Health Technology Assessment 1998; Vol. 2: No. 19

Review

Systematic reviews of trials and other studies

AJ Sutton
KR Abrams
DR Jones
TA Sheldon
F Song

Health Technology Assessment
NHS R&D HTA Programme

Standing Group on Health Technology

Chair: Professor Sir Miles Irving,
Professor of Surgery, University of Manchester, Hope Hospital, Salford †

HTA Commissioning Board

Chair: Professor Charles Florey, Department of Epidemiology & Public Health,
Ninewells Hospital & Medical School, University of Dundee †

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Systematic reviews of trials and other studies

AJ Sutton[1]
KR Abrams[1]
DR Jones[1]
TA Sheldon[2]
F Song[3]

[1] Department of Epidemiology and Public Health, University of Leicester, Leicester, UK
[2] York Health Policy Group, University of York, York, UK
[3] NHS Centre for Reviews and Dissemination, University of York, York, UK

Published March 1999

This report should be referenced as follows:

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *Health Technol Assess* 1998;**2**(19).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE. Copies of the Executive Summaries are available from the NCCHTA web site (see overleaf).

NHS R&D HTA Programme

The overall aim of the NHS R&D Health Technology Assessment (HTA) programme is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and work in the NHS. Research is undertaken in those areas where the evidence will lead to the greatest benefits to patients, either through improved patient outcomes or the most efficient use of NHS resources.

The Standing Group on Health Technology advises on national priorities for health technology assessment. Six advisory panels assist the Standing Group in identifying and prioritising projects. These priorities are then considered by the HTA Commissioning Board supported by the National Coordinating Centre for HTA (NCCHTA).

This report is one of a series covering acute care, diagnostics and imaging, methodology, pharmaceuticals, population screening, and primary and community care. It was identified as a priority by the Methodology Panel and funded as project number 93/52/03.

The views expressed in this publication are those of the authors and not necessarily those of the Standing Group, the Commissioning Board, the Panel members or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for the recommendations for policy contained herein. In particular, policy options in the area of screening will be considered by the National Screening Committee. This Committee, chaired by the Chief Medical Officer, will take into account the views expressed here, further available evidence and other relevant considerations.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

Series Editors: Andrew Stevens, Ruairidh Milne and Ken Stein
Editorial Assistant: Melanie Corris

The editors have tried to ensure the accuracy of this report but cannot accept responsibility for any errors or omissions. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

Contents

Part A: Methods

Part B: Results I – pre-synthesis methods

Part C: Results II – the critical appraisal of research evidence

Part D: Results III – statistical analyses (classical and Bayesian)

Part E: Results IV – other important issues when combining estimates

Part F: Results V – applications of meta-analysis in other contexts and using other data types

Part G: Results VI – extensions of meta-analytic methods

Part H: Results VII – summary, recommendations and further research

List of abbreviations

AMI	acute myocardial infarction
ARE	asymptotic relative efficiency
ANOVA	analysis of variance
BFs	Bayes factors
BIDS	Bath Information and Data Services
BLDSC	British Library Document Supply Centre
BMT	bone marrow transplantation
CCT	controlled clinical trial
CHD	coronary heart disease
CI	confidence interval
CMWG	Cochrane Methods Working Group
CONSORT	Consolidated Standards of Reporting Trials
CRD	Centre for Reviews and Dissemination
CWG	Cochrane Working Group
df	degree of freedom
EB	empirical Bayes
EBCTCG	Early Breast Cancer Trialists' Collaborative Group
EM	expectation maximisation
ETS	environmental tobacco smoke
FDA	Food and Drug Administration
FNR	false negative rate
FPR	false positive rate
GAO	US General Accounting Office
GEE	generalised estimating equations
GLS	Generalised Least Squares
HIV	human immunodeficiency virus
HTA	Health Technology Assessment
ID	identification
IPD	individual patient data
ISI	Institute of Scientific Information
IVSK	intravenous streptokinase
LRR	log(relative risk)
LRT	likelihood ratio test
MAL	meta-analysis of literature
MAP	meta-analysis of individual patient data
MCMC	Markov Chain Monte Carlo
MeSH	medical subject headings
ML	maximum likelihood
MLE	maximum likelihood estimate
NNT	number needed to treat
NRR	National Research Register
NTIS	National Technical Information Service
O–E	observed deaths minus expected deaths
OR	odds ratio
RCT	randomised clinical trial
REML	restricted maximum likelihood
ROC	receiver operating characteristic
RR	relative risk
RRR	relative risk reduction
SAS	Statistical Analysis System
SE	standard error
SIGLE	System for Information on Grey Literature
SMR	standardised mortality ratio
SROC	summary receiver operating characteristic
TNR	true negative rate
TPR	true positive rate

Executive summary

Objectives

Systematic review and meta-analytical methods are already common approaches to the assessment of health technology and related areas, and increasing adoption of such approaches may be foreseen, in part in response to increasingly wide emphasis on evidence-based approaches to medicine and health care. This report is intended:

- to identify applications of systematic review and meta-analytical methods in Health Technology Assessment (HTA)
- to promote further, appropriate use of such approaches in these areas of application
- to begin to identify priorities for further methodological developments in this field.

How the review was conducted

Systematic literature searches using MEDLINE, EMBASE, and Institute of Scientific Information (ISI) Science/Social Science electronic databases and the Cochrane methods database were carried out to find relevant articles. Relevant reference collections of the study team were pooled. Grey literature and unpublished articles were obtained by writing to prominent researchers, and through the Internet; further papers were identified by inspecting the reference lists of all previously obtained articles.

Review findings

A large number of papers concerning methodology relevant to different aspects of systematic reviews were identified. While the ordering of the report follows the stages involved in carrying out a systematic review, it is highly structured in a way which enables readers with specific interests to locate particularly relevant sections easily. The main features of the report are now summarised briefly in turn.

A brief overview of the important issues to be considered prior to the appraisal and synthesis of studies, including a critical appraisal of search methods, is presented.

Methodology for critical appraisal of the research evidence, including ways of assessing the quality of the primary studies, and its incorporation into a review, is explored. No consensus has been developed as to which method is most appropriate for doing this.

An important consideration is the possibility of heterogeneity between study outcome estimates. Many assessments and formal tests for detecting heterogeneity are described. Methods for accounting/adjusting for heterogeneity are identified and assessed. No consensus has been reached concerning the best strategy for dealing with heterogeneity; currently a large degree of subjectivity is required on the part of the reviewer.

Both classical and Bayesian statistical approaches have been developed to combine study estimates. These encompass the relatively simple fixed effect approaches, through random effects models, to more sophisticated hierarchical modelling. The more complex methods were largely devised to deal with heterogeneous outcomes, systematic variation between studies, and the need to incorporate a fuller set of components of variability into the model. Several of these methods have come under criticism; it is concluded that neither fixed nor random effect analyses can be considered ideal.

In addition to these general methods, approaches specific to particular outcome scales/measures, and data types are identified. These include methods for combining ordinal, binary, and continuous outcomes; survival data; diagnostic test data; correlated outcomes; individual patient data; single arm studies; crossover trials; and finally, studies of differing designs. While some of these methods have become standard, others are less commonly used and so are at early stages of development.

Problems encountered by meta-analysts were identified. Two potentially serious ones are publication bias and missing data. Methods for detecting/adjusting for publication bias exist, and others are currently being developed. The validity of most is largely undetermined. Additionally, long-term policy measures such as registries for all trials have been suggested. Dealing with missing data within a meta-analysis has not been considered to the same extent. General methods do exist (in other literatures), but many of them are untested in a meta-analytical setting.

Further issues identified include methods used to report the results of systematic reviews; use of sensitivity analyses; prospective meta-analysis; and alternatives to traditional meta-analysis.

Several of the key methods are illustrated using a dataset comprising cholesterol lowering studies.

Recommendations

Recommendations for good practice for the most part follow standard and widely agreed approaches. Greater latitude in the nature of studies potentially eligible for review, including non-randomised studies and the results of audit exercises, for example, may, however, be appropriate. The key stages are (with extensions and/or less widely agreed aspects in parentheses):

1. Specification in a protocol of the objectives, hypotheses (in both biological and health care terms), scope, and methods of the systematic review, before the study is undertaken.
2. Compilation of as comprehensive a set of reports as possible of relevant primary studies, having searched for all potentially relevant data, clearly documenting all search methods and sources.
3. Assessment of the methodological quality of the set of studies (the method being based on the extent to which susceptibility to bias is minimised, and the specific system used reported). Any selection of studies on quality or other criteria should be based on clearly stated *a priori* specifications. The reproducibility of the procedures in 2 and 3 should also be assessed.
4. Identification of a common set of definitions of outcome, explanatory and confounding variables, which are, as far as possible, compatible with those in each of the primary studies.
5. Extraction of estimates of outcome measures and of study and subject characteristics in a standardised way from primary study documentation, with due checks on extractor bias. Procedures should be explicit, unbiased and reproducible.
6. Perform, where warranted by the scope and characteristics of the data compiled, quantitative synthesis of primary study results (meta-analysis) using appropriate methods and models (clearly stated), in order to explore and allow for all important sources of variation (e.g. differences in study quality, participants, in the dose, duration, or nature of the intervention, or in the definitions and measurement of outcomes). This will often involve the use of mixed/hierarchical models, including fixed covariates to explain some elements of between-study variation, in combination with random effects terms.
7. Performance of a narrative or qualitative summary, where data are too sparse, or of too low quality, or too heterogeneous to proceed with a statistical aggregation (meta-analysis). In such cases the process of conduct and reporting should still be rigorous and explicit.
8. Exploration of the robustness of the results of the systematic review to the choices and assumptions made in all of the above stages. In particular, the following should be explained or explored:
   a) the impact of study quality/inclusion criteria
   b) the likelihood and possible impact of publication bias
   c) the implications of the effect of different model selection strategies, and exploration of a reasonable range of values for missing data from studies with uncertain results.
9. Clear presentation of key aspects of all of the above stages in the study report, in order to enable critical appraisal and replication of the systematic review. These should include a table of key elements of each primary study. Graphical displays can also assist interpretation, and should be included where appropriate. Confidence intervals around pooled point estimates should be reported.
10. Appraisal of methodological limitations of both the primary studies and the systematic review. Any clinical or policy recommendations should be practical and explicit, and make clear the research evidence on which they are based. Proposal of a future research agenda should include clinical and methodological requirements as appropriate.

Further areas of research related to the methods used for systematic reviews

Two priority areas are indicated below. Additionally, other areas needing further research are highlighted.

Priority topics

- Sensitivity analysis of the impact of many aspects of the design and analysis of the systematic

review, and in particular of the meta-analysis, has been advocated. The result is a complex set of inter-related sensitivity analyses. Research into optimum, or at least efficient, strategies of multi-dimensional sensitivity analysis in these contexts would thus be useful.
- Evaluation of the role in HTA of meta-analysis of observational studies, and cross-design synthesis (which often features the inclusion of non-randomised evidence), possibly through systematic research and workshops of researchers active in the field.

Other areas needing further research

Study quality

- Investigation into the relevant dimensions of methodological quality and empirical research which establishes the relative importance of these dimensions in different contexts. This should eventually lead to the development of rigorous, validated, and parsimonious scales which can be adapted to a wide range of studies.
- Exploration of study quality as an explanation of heterogeneity.
- Empirical investigation into the basis for choice of cut-off values for exclusion of studies on grounds of quality.
- Systematic approaches to quality assessments of non-randomised controlled trials.

Heterogeneity

- Further investigation of its relationship with publication bias.
- Development of guidelines/recommendations for identifying and exploring heterogeneity.
- Investigation of degree of heterogeneity (both quantitative and qualitative) beyond which combining of all the studies should not be considered.
- Investigation into the effects of choice of measurement scale from both: a) a statistical perspective, and b) a clinical perspective.

Publication bias (HTA has commissioned a separate review in this area)

- Assessing the impact of the pipeline problem.
- Empirical study of degree and mechanisms of publication bias in meta-analysis of epidemiological and other non-randomised studies.
- Investigation into the extent to which the use of a prospective register for trials minimises publication bias.
- Further investigation into proposed statistical methods, including their power to detect publication bias, and their sensitivity towards its detection.

Approaches to modelling and analysis

- Investigation of the relative merits of the different approaches to combining studies in which some arms report no events (zeros in 2 × 2 tables)
- Comparison of new methods for random effects modelling which fully incorporate parameter uncertainty.
- Investigation of robustness of random effects models to departures from normality.
- Empirical investigation of model attributable weights with particular reference to over-weighting of large samples, in some models.
- Investigation of the impact of missing data at both the study level and patient level.
- Development of experience with practical applications of mixed models.
- Development of methodology for combining individual patient data with study level data.
- Investigation of the role of cumulative/sequential application of meta-analysis as a research methodology.
- Further development of methods for integration of qualitative assessments of studies with quantitative estimates of the results.
- Development of random/mixed effects models for meta-analysis of survival data.
- Use and implications of exact statistical methods for combining small studies.
- More extensive but critical use of Bayesian methods, including:
  a) encouragement of **expository papers** in the applied literature on the application of Bayesian methods
  b) more research on obtaining and using **elicited prior beliefs**.
- More research into the use of meta-analytic techniques in conjunction with decision analysis methods.
- General investigation of the impact of missing values, and extension of currently available methods to a wider range of circumstances with missing data, including the use of Bayesian methods.
- Development of the use of simulation of results of new studies before they are published or of hypothetical studies to allow their impact on meta-analysis to be assessed.

Miscellaneous

- More research into extrapolating the results of a meta-analysis to clinical practice.
- Further development of detailed publication guidelines to encourage uniform reporting of the results of studies, particularly of types other than randomised clinical trials.

Preface

Use of systematic review and meta-analytical methods in HTA and related areas is now common. This report is intended to promote appropriate application of such methods, and to begin to identify further appropriate methodological developments.

It is not intended as a text book of these methods but as a structured survey of practice and problems in the area. We hope that readers will be rapidly able to find and understand a review of the use of these methods in contexts relevant to their particular interests. The strongly subdivided but cross-referenced text, selected worked examples of key methods, and relatively heavy use of quotation from original sources are all intended to aid the reader in so doing. Similarly, the explicit documentation of the search strategy used should allow readers to update the review in areas of particular interest to them. The selection of the quotes included in the report was somewhat arbitrary. They were largely included where particular issues were expressed eloquently and precisely in papers, and it was felt that rewording might detract from the view expressed, or to reflect the range of opinions expressed by experts in the field.

We are pleased to be able to acknowledge the assistance we have received from several people and groups in the course of compiling this review. In particular we wish to thank: Dr Martin Hellmich, a Visiting Fellow in the Department of Epidemiology and Public Health at Leicester, for contributing the chapter on meta-analysis of diagnostic test accuracy; Sage publications for supplying several books without charge; Julie Glanville and Janette Boynton at the Centre for Reviews and Dissemination, York, for their advice on searching for the literature; the Cochrane Collaboration for supplying their database of literature, and dealing with several queries; Teresa Smith from the Biostatistics Department at Cambridge for her code to plot some of the meta-analyses diagrams found in this report; the authors who supplied copies of their work prior to its submission/publication in journal format for inclusion in this review; Cindy Billingham, also doing a NHS HTA review (93/50/04 Quality of life assessment and survival data), for her advice throughout the project.

Part A:

Methods

Chapter 1

Introduction

Systematic review and meta-analytical methods are already common approaches to the assessment of health technology and related areas, and increasing adoption of such approaches may be foreseen, in part in response to increasingly wide emphasis on evidence-based approaches to medicine and health care. The potentially relevant methodological literature is already substantial. This review, for example, draws on a database of about 1000 potentially relevant references, and cites about 600 of them. This report is the outcome of the NHS Research and Development Health Technology Assessment Programme methodology project number 93/52/03. It is intended:

- to identify applications of systematic review and meta-analytical methods in Health Technology Assessment (HTA)
- to promote further, appropriate use of such approaches in these areas of application
- to begin to identify priorities for further methodological developments in this field.

The review and this report focus primarily on the use of quantitative methods to obtain overall estimates of effectiveness of interventions by means of the statistical pooling of the results of studies or methods of exploring variations in their results. In many health technologies; however, the evaluations are too dissimilar, or the outcomes too varied to permit the use of statistical analysis of the studies as a single set. In these situations researchers will not be able to use formal statistical techniques to derive estimates of the effectiveness of interventions. When this is the case, it is important that the systematic review still adopts the comprehensive, rigorous and explicit approach used when more quantitative methods can be applied.

Qualitative approaches to study synthesis will still need to appraise studies critically to assess their validity. However, this will not be applicable in a quantitative manner to obtain overall estimates. Qualitative analysis should examine variation in outcomes and attempt to explore this in terms of study design characteristics, the participants and nature of the interventions/exposures. The result of a qualitative analysis is likely to be a range of plausible effect sizes and a judgement of the direction of likely benefit. These, however, should be justified explicitly on the basis of the study results, and any implicit weightings made clear.

In many ways, the quantitative analyses considered in this report represent special cases of the qualitative analysis. Although the emphasis in this report is on the more technical aspects of analysis, quantitative studies should of course not neglect the simpler aspects of analysis and presentation which they share with all qualitative studies, including adequate description of the primary studies on which they are based. The conduct of a quantitative analysis, however, should not be used as an excuse for inadequate description of the studies included.

This report is divided into eight parts. Part A describes the methods adopted for the project. Part B outlines the methods for the pre-synthesis stage of a review. Part C discusses methods for the critical appraisal of the research evidence. Part D describes the statistical methods used to combine study results. Part E discusses other issues which are important when synthesising evidence. Part F describes further methods specific to certain contexts. Part G presents extensions to the meta-analytic methods described in previous sections. Part H summarises the recommendations and topics needing further research.

Appendix 1 summarises the literature search strategies used in compiling the database of literature on which the report is based, to help identify the coverage explicitly, and to facilitate updating of the database, and hence the review, in future. Appendix 2 lists papers identified shortly before the completion of this report, which could not be included in the main text because of time constraints. The report concludes with a Bibliography of all relevant papers identified for this review (whether they were actually cited in the text or not).

Chapter 2

Methods adopted for this project

The literature search

Objectives

The primary objective of the literature search was to locate all (or as much as feasibly possible) of the literature concerned with the methodology used in the systematic review of evidence. This meant, although our interest was methods for evidence based care, a search for synthesis methodology was done irrespective of discipline.

The search strategy

Several approaches were taken to search for the relevant literature, with the intention of achieving the highest retrieval rate possible:

Electronic databases: The databases Institute of Scientific Information (ISI) Science, ISI Social Science, and EMBASE were all searched via the Bath Information and Data Services (BIDS) computer service. In addition, MEDLINE was searched on CD-ROM using the OVID search engine. All these databases were searched from the beginning of 1991, through August 1996, with the exception of MEDLINE which was searched from the beginning of 1992.[1] Simplified updates of these searches were carried out into the first quarter of 1997. Appendix 1 gives a detailed description of the search strategies used.
Cochrane Database: Papers concerned with methodology used for meta-analysis had previously been compiled by Oxman. This list (which has since been updated) was available electronically via the Cochrane library (1).[2] The vast majority of the references were directly relevant to this project.
Private collections: Two of the members of the study team (DRJ, KA) had worked in the area of methodology for systematic reviews prior to this project, and hence had private literature databases to draw on.
Other methods: The reference lists of each relevant paper obtained were examined to identify papers not found previously by the above methods. Known researchers in the field were contacted for work completed, but as yet unpublished. Unpublished papers and technical reports were retrieved from researchers home pages via the Internet.

Searching methods/results

The electronic database searches were carried out first. The results of these searches were compared with the Cochrane database and the private collections, to assess how successful they were, by calculating the proportion of the known articles the search strategies retrieved. Because the different sources used varying time windows (the Cochrane database had no articles post-1994 and the electronic database searches covered 1991–1996), an exact evaluation was not made; however, it was clear that utilising all three sources was beneficial, as each highlighted substantial numbers of references the other two methods had not found. Searching the reference lists of the papers found by these methods again brought to light a substantial number of new references.

A database of these was created using the Reference Manager (version 7) (2) software package. Each reference was keyworded by one of the researchers (AS) using a unique and personal keyword system to help order and categorise the large body of literature.

Using this system, it was difficult to ascertain all the original sources of the references when looking retrospectively, since if more than one source had retrieved a reference, then the duplicates would not be included in the database. This means that an assessment of the performance of each database was not possible.

The search and retrieval of literature continued throughout the duration of the project, this included obtaining literature published after the initial searches. By the completion of the project, 1005 potentially relevant methodology references had been identified. Thirty-four of these were identified too late to include in the review (see appendix 2 for a listing). Of the remaining 971 references, 781 were obtained and inspected/read. The remaining 190 were not obtained, in the vast majority of instances due to one of the following reasons:

[1] This was due to the available CD-ROM version only covering 1992 onwards.

[2] The pre-update list, which is the one that was available when this search was carried out, is also available in printed form (3).

- Although the paper was in some way related to carrying out research synthesis, a consensus decision, after reading abstracts or other information, deemed the paper to contain no new developments in methodology. In such instances the article was often an introductory or tutorial paper. Additionally, a considerable number of papers appear to discuss/review the issues involved in a meta-analysis but do not contribute to new methodology (the majority of the papers not obtained were due to these reasons).
- The literature could not be obtained. It may have been badly referenced, or extremely diverse, and the National Library could not trace it/obtain it. Alternatively, it could have been referenced as an unpublished work with little or no indication how to obtain it.
- It may have been an 'old' reference (pre-1970), the relevant contents of which had been summarised in easier to obtain formats.

It is important to note that just because a paper was obtained and read did not mean it was automatically included in this report. A considerable number of papers read did not present new methodology, or any methodology content had been described and written about from other sources.

To make it absolutely clear what sources were considered in the various sections of the report, several reference lists have been compiled. There is a reference list at the end of every chapter which includes all references cited therein. Not all these references may be directly relevant to meta-analysis; for instance they may have been included to provide background reading on a particular topic. The main bibliography at the end of the report provides a list of all the references concerning meta-methodology (in some way), found during the project, whether they were actually cited in the main text or not. This list **excludes** the non-meta-analysis citations found in the text. Additionally, appendix 2 includes the 34 references known, or suspected of having new methodology in, that came to our attention too late to include in the review.

Discussion

Searching databases for the methodology references on a particular subject is notoriously difficult.[3] Their seems to be no simple strategy for effectively retrieving the relevant information; moreover, the suggestion given to us[4] that there is no substitute for time invested in simply scanning through the huge numbers generated by the first level search. The fact that approximately 300–400 new references were identified by scanning lists of references already retrieved (missed by the database searches) would indicate that the electronic database search strategies were not sufficient in themselves. Indeed, it would appear that scanning reference lists is a very time effective way of locating the literature. It is also interesting to note that relevant papers were picked up by doing this that would never have been picked up via a database search. An example of this is the paper below:

Emerson JD. Combining estimates of the odds ratio: the state of the art. *Stat Methods Med Res* 1994;**3**:157–78.

This paper has much valuable advice on combining odds ratios without ever explicitly mentioning meta-analysis, synthesis or combining studies, and hence was not retrieved in the first stage search (see appendix 1). This raises interesting questions for people carrying out methodological reviews, and highlights the benefit of using supplementary searching methods such as scanning reference lists and handsearching core relevant journals.

For completeness, below is set out what we believed would be the 'ideal' search strategy, devised several months into the searching. As one can see several of the points were not carried out. This was simply due to time constraints.

Overall search strategy

- Identify key existing collections.
- Search BIDS EMBASE 1991–1996, using standard search.
- Search ISI Science 1991–1996, using standard search.
- Search OVID MEDLINE 1992–1996, using adapted standard search and selected parts on CRD search.
- Search PsycLIT (Psychology database).
- Search Education database (ERIC–ERIC international or ISI Soc Science).
- Search samples of large sets not already imported from above databases – if many are found, continue.

[3] Personal communication with Julie Glanville (Information officer at the University of York).

[4] Personal communication with Julie Glanville (Information officer at the University of York).

- Select approximately 1% of non-intersect papers from the meta-analysis pools (year by year) to see what is missing.
- Search back to 1980 in at least EMBASE, ISI Science and MEDLINE.
- As a check, check through last 5 years (or perhaps from 1980) of statistics journals.
- As a check, download citation searches on (six) key papers.

For the researcher wishing to keep this review up to date, the authors offer the following advice on the searching methods. To carry out an updated search using all the databases and strategies of appendix 1 would be a time-consuming procedure, not least due to its relatively low hit rate. Without further investigation, it would be difficult to suggest which parts were least effective/only duplicated other parts of the search however. It would appear that different strategies could be most effective for different topic areas. For instance, if one is only looking for papers presenting statistical methods, a different approach should be taken from finding literature on say searching methods. However, whatever database searching strategy is used primarily, our advice would be always to inspect the reference lists of all relevant papers obtained.

Implications for updating the review

It is difficult to ascertain how fast the field is currently moving, or how fast it will move in the future. The number of papers published each year (on meta-analysis/systematic review methods) gives some indication that this is very much a growth area and one that has grown at an accelerating rate over the past 10 years. Indeed, it has been reported that the number of papers which report applications of meta-analysis is increasing exponentially. It would seem realistic to expect that the methodological developments will increase as the application of methods increases, and as the areas of application broaden, new methods will be required.

Standard methods (i.e. fixed, random, and mixed modelling) seem pretty much in place now but experience of using and choosing between models needs to be developed. The Bayesian alternative is now a real alternative due to increased computational power; recently extensions from a Bayesian perspective have appeared, and we suspect they will continue to do so. Due to this, there may will be a shift to using more sophisticated and realistically complex modelling techniques.

Developments for specific situations continue to be presented, including methods to deal with missing data, and combining information from disparate sources. The area of fastest growth, however, is methods to assess and adjust analyses for publication bias. More research is likely to be produced on this topic in the near future.

A list of potentially important papers published too late for inclusion in this review is included in appendix 2.

References

1. The Cochrane Library. Issue 1 edn. BMJ Publishing Group, 1996.
2. Reference Manager (computer program). 7. 2355 Camino Vida Roble, Carlsbad CA 92009-1572, USA: Research Information Systems; 1995; Windows.
3. Altman D, Chalmers I, editors. Systematic reviews. London: BMJ Publishing Group, 1995.

Part B:

Results I – pre-synthesis methods

Chapter 3

Procedural methodology (for meta-analysis)

Introduction

It is the aim of this section to outline the rigorous procedural methodology that has been advocated when conducting any meta-analysis. At least two sets of guidelines have been published laying out the procedural path to be followed when conducting a systematic review (1) (referred to as the Cochrane Handbook in the text), and (2) (referred to as CRD4). The Cochrane Collaboration began its work in 1992, with the aim to prepare, maintain and disseminate systematic, up-to-date reviews of randomised clinical trials (RCTs) of health care, and, when RCTs are not available, review the most reliable evidence from other sources. In 1997 the second edition of their handbook was produced, which lays out the procedural methodology to be followed when conducting a review within the collaboration (1). Similarly, the NHS Centre for Reviews and Dissemination (CRD) at York have produced a similar document outlining guidelines for those carrying out or commissioning reviews for themselves or other research bodies (2). These two sets of guidelines are similar in both structure and content. The sections below outline the stages involved in carrying out a systematic review, as identified in these documents, and aim to give the reader an idea of the suggested procedures to follow when carrying out a systematic review. If the reader is carrying out a review for a specific body (such as the Cochrane Collaboration), then clearly it is necessary to follow their guidelines strictly. If one is carrying out a systematic review independently, then the rigorous methods put forward by these two organisations will stand the researcher in good stead for carrying out a worthy review of their own. Both sets of guidelines will give more detailed accounts of the procedures outlined here and are recommended reading. Both guidelines also discuss the logistics of doing a review – a subject not covered in this report.

Identification of the need for the review[1]

Even before a review is undertaken it is important to establish the need for such a review, as CRD4 states (2):

> ‘It is important to be clear about the aim and requirements of each systematic review before it is started, and to be aware of other reviews in the field of interest that have previously been published or are currently in progress.’

One should check for other reviews (published or in preparation) using the Cochrane Database of Systematic Reviews (3–5), the CRD Database of Reviews of Effectiveness (DARE) (6) and the NHS National Research Register (NRR) (7). Also, the more common electronic databases (such as MEDLINE, EMBASE)[2] should also be searched. Key research groups within the field could also be contacted.[3] A further issue that needs consideration is:

> ‘Background information describing the epidemiology of the health care problem and the patterns of use of a health technology and its alternatives should be briefly reviewed. An outline should also be given of the present options and arrangements for health care provision in the review area, together with routine statistical data describing their use. It may be of value to include information on the historical, social, economic and biological perspectives to the review problem.’ (2)

In addition (1), presented below are general points regarding issues that need taking into account when considering and undertaking a systematic review:

- Questions should address the choices (practical options) people face when deciding about health care.

[1] These guidelines mainly assume that one is assessing the effectiveness of a treatment. However, if the review is about some other topic such as a diagnostic test (see chapter 21) or a risk factor then these guidelines may need modifying.

[2] A discussion of these electronic databases is given in chapter 4. Also CRD4 gives search strategies for locating review articles in MEDLINE (2).

[3] Not all review articles are systematic; thus they need to be critically appraised, this can be done via checklists given in (1).

- Reviews should address outcomes that are meaningful to people making decisions about health care.
- The methods used in a review should be selected to optimise the likelihood that the results will provide the best current evidence upon which to base decisions.
- It is important to let people know when there is no reliable evidence, or no evidence about particular outcomes that are likely to be important to decision makers.
- It is not helpful to include in a review evidence where the risk of bias is high, even if there is no better evidence.
- Similarly, it is not helpful to focus on trivial outcomes simply because those are what researchers have chosen to measure.
- So far as is possible, it is important to take an international perspective. The evidence collected should not be restricted by nationality or language without good reason.
- Results should be presented in such a way that their applicability in different contexts can be assessed by decision makers.
- Reviewers should bear in mind that different people might make different decisions based on the same evidence (for good reasons). The primary aim of a (Cochrane) review should be to summarise and help people to understand the evidence. Reviewers must be careful not to impose their own values and preferences on others when answering the questions they pose.

In addition to the above, it is also important to establish that the results of any proposed review are not invalidated by the publication of a current RCT/study. This could for example be avoided by checking trial registers and contacting experts in the area. In this way, pipeline bias could be avoided (see chapter 16).

CRD4 also makes the suggestion that the target audience for the review (i.e. people who will use the results) should be identified early on.

Background research

Having decided on the appropriateness of a review, the next stage is to explore the existing information on the topic further. It is necessary to determine the scope of the review and the specific questions that the review will address (8–10).

A preliminary assessment of the primary research that is available should be made, it should be done considering the following points (1):

- Assessing the volume of literature in the field – can be done using electronic databases.
- Assessing study designs used in the primary research – decisions have to be made on which designs are to be included in the review.[4]
- Assessing effectiveness using causal pathways – the effectiveness of a treatment policy may involve a sequence of interventions that cannot be evaluated in a single study. 'If the literature search reveals that there are no complete evaluations of the effectiveness of the intervention policy than an analysis of the components of the components of the policy should be considered. Where possible these should be mapped out by a causal pathway (11–13)'.
- Identification of questions to be addressed in the review – the most important decision (1).
- Identification of outcomes.
- Identification of effect modifiers: 'There may be factors, such as the characteristics of the patients and settings, choice and measurement of outcomes, or differences in the nature or delivery of interventions, which influence the estimates of effectiveness of the intervention under investigation. It is important that these 'effect modifiers' are identified as they may explain apparent differences in the findings of the primary studies.' (1)
- Identification of particular issues related to validity – checking the primary studies are methodologically sound; issues include randomisation, unsuitable comparison groups, a lack of blind outcome assessments, inadequate follow-up times, a lack of suitable gold standard diagnostic tests, inability to define and assess relevant outcomes, unreliable measurement techniques, or inappropriate statistical analysis.
- Identification of issues related to generalisability – it should be noted whether the design, setting and participants of the primary studies will reduce the generalisability of the review's findings.

Whether the existing literature helps to focus a review or whether ignoring given background information yields an impossible review is a fine balance, that is very dependent upon the topic area to be reviewed. Ultimately, the benchmark

[4] (2) Gives a search strategy for identifying RCTs using MEDLINE.

by which to judge a review is whether it will help to inform healthcare/policy decisions.

When carrying out a review, one also needs to balance scientific validity and work load, it should be kept in mind that: 'There is little value in doing a review which will produce an unreliable answer.' (2)

The review protocol

(2) The protocol specifies the pre-determined plan (14, 15) which the research exercise will follow. It is very important to establish methods before the review is started, to avoid biases. The Cochrane Handbook warns that (1):

> '*Post hoc* decisions (such as excluding selected studies) that are made when the impact of the results of the review is known are highly susceptible to bias and should be avoided. As a rule, changes in the protocol should be documented and reported, and "sensitivity analyses" (see chapter 27) of the impact of such decisions on the results of the review should be made when possible.'

The methods described should be rigorous and clearly defined, and should have repeatability; that is, someone else should be able to replicate the methods/results.

The following sections outlined below should all be detailed in the protocol. Most of these subjects are dealt with in detail later in this report: links are given where appropriate.

Problem specification

The protocol should state in detail the main questions or hypotheses which will be investigated in the review.

The Cochrane Handbook states:

> 'There are several key components to a well-formulated question. A clearly defined question should specify the types of people (participants), types of interventions or exposures, and the types of outcomes that are of interest. In addition, the types of studies[5] that are relevant to answering the question should be specified. In general the more precise one is defining components, the more focused the review.' (1)

When doing this, it is worth keeping in mind the below comment from CRD4:

> 'While questions should be posed before initiating the actual review, it is important that these questions do not become a straight-jacket which prevents exploration of unexpected issues' (2) If changes are made at a later date, then these amendments should be stated in the protocol (14).

Searching for studies

The proposed search strategy should be described naming databases and other sources of information. Any restrictions, such as limiting the language of reports, should also be stated (2). The methods used for carrying out this stage of a research synthesis are documented in chapter 4.

Deciding on study inclusion criteria

The health intervention/technology of interest, the setting and relevant patients or client groups, and the outcome measures used to assess effectiveness should all be clearly defined (2).

The types of study design to be included should also be specified: 'Note that even though RCTs may be the preferred design there are several areas of health care which have not been evaluated using RCTs.' (2) (see chapter 6).

On deciding how broad or narrow to make the inclusion criteria, there is a trade off between reducing generalisability of the results and obtaining information which is hard to compare and synthesise (16,17). If inclusion criteria are quite liberal, it may be possible to investigate theories concerning the effects of differences in the study characteristics and other effect-modifiers using 'meta-regression' or other statistical methods (see chapter 11).

Study validity

The basic checklists which will be used to assess the validity of the primary studies should be included in the protocol (2). Scales and checklists and their use are covered in chapter 6.

Data extraction

A data-extraction sheet could be included to assist the evidence extracting process (2). This topic is covered on page 17.

Study synthesis

Although it may not be possible to state explicitly which statistical, or other methods, will be used until after the studies have been assessed, the general modelling approaches that are likely

[5] This is discussed on pages 16–17 and 23–4.

to be used should be specified (2). Also, any hypothesis-testing and subgroup analyses (see page 209) should be specified here *a priori* (14). This is done to prevent many analyses being carried out post hoc, which potentially may lead to spurious associations being found. For the same reason, it may be important to have a limit on the number of such hypotheses in the protocol. The statistical methodology used to combine results from different studies forms a large proportion of this report. Sections D, E, F, and G discuss the statistical methods that have been used for research synthesis.

Summary

This section is not intended to be anything more than a brief overview of the issues that are important when one is considering carrying out research synthesis. It may help the researcher who is new to the subject to get a feel for the discipline, and serve as a springboard into later sections of this report, as many of the issues touched on here are expanded in later sections.

References

1. Oxman AD, editor. The Cochrane Collaboration handbook: preparing and maintaining systematic reviews. Second edn. Oxford: Cochrane Collaboration, 1996.
2. Deeks J, Glanville J, Sheldon T. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. Centre for Reviews and Dissemination, York: York Publishing Services Ltd, #4, 1996.
3. Chalmers I, Sandercock P, Wennberg J. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann N Y Acad Sci* 1993;**703**:156–65.
4. Chalmers I, Haynes B. Reporting, updating, and correcting systematic reviews of the effects of health care. *BMJ* 1994;**309**:862–5.
5. The Cochrane Collaboration. The Cochrane Database of Systematic Reviews (database on disk and CD ROM). London: BMJ Publishing Group, 1997.
6. The NHS Centre for Reviews and Dissemination. The Database of Abstracts of Reviews of Effectiveness (database on line). York: University of York, 1997.
7. NHS Executive. The National Research Register (database available on disk). Welwyn Garden City: VEGA Group Plc, 1995.
8. Hall JA, Tickle-Degnen L, Rosenthal R, *et al.* In: Cooper H, Hedges LV, editors. 2, Hypotheses and problems in research synthesis. The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 17–28.
9. Goodman C. Step 1: specify the assessment problem. In: Literature searching and evidence interpretation for assessing health care practices. Stockholm, SBU: The Swedish Council on Technology Assessment in Health Care, 1993.
10. Light RJ, Pillemar DB. Summing up: the science of reviewing research. Cambridge, Mass: Harvard University Press, 1984.
11. Canadian Task Force for the Periodic Health Examination. The periodic health association. *Can Med Assoc J* 1979;**121**:1193–7.
12. Effective Health Care. Screening for osteoporosis to prevent fractures. Leeds: University of Leeds, 1992.
13. Woolf SH, Battista RN, Anderson GM, Logan AG, Wang E. Canadian Task Force on the Periodic Health Examination. Assessing the clinical effectiveness of preventive manoeuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol* 1990;**43**:891–905.
14. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis (review). *J Clin Epidemiol* 1995;**48**:167–71.
15. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analysis of randomized controlled trials. *N Engl J Med* 1987;**316**:450–5.
16. Horwitz RI. Large-scale randomized evidence: large, simple trials and overviews of trials: discussion. A clinician's perspective on meta-analyses (comment). *J Clin Epidemiol* 1995;**48**:41–4.
17. Eysenck HJ. Systematic reviews – meta-analysis and its problems. *BMJ* 1994;**309**:789–92.

Chapter 4

Searching the literature and identifying primary studies

The importance of the literature search

As has been identified in CRD4:

> 'The aim of the search is to provide as comprehensive a list as possible of primary studies, both published and unpublished, which may fit the inclusion criteria and hence be suitable for inclusion in the review.' (1)

It is worth remembering that, the 'precision of the estimate of effectiveness depends on the volume of information obtained.' (1) Also, it is 'important to ensure that the process of identifying studies is not biased, minimising the possibility of the review's conclusions being weakened through publication bias' (1) (see chapter 16). Unfortunately, this is only possible if a prospective comprehensive research register is maintained for the topic, as only then is the ability to identify a study not influenced by its findings (2,3) (see pages 132–3). Indeed a comprehensive, unbiased search is one of the key differences between a systematic review and a traditional review (4).

The methodology of searching and collecting studies for a meta-analysis has been somewhat overshadowed by the research related to statistical methods to combine studies (5). It is clear, however, that the validity of the results of statistical analyses depends on the validity of the underlying data (6), and every effort should be made to locate the primary studies.

In many topic areas the potentially useful literature may be very large. This has led some meta-analysts to doubt whether comprehensive searches are worth the effort (7). But as White observes (8), even they seem to have more rigorous standards for uncovering studies than librarians and information specialists typically encounter.

The guidelines of the CRD and the Cochrane Collaboration (1,4) seem to imply that the search should be exhaustive, that is to say, trying to find every study on a given topic, however some consider this is unrealistic and White has made the comment (8):

> 'The point is not to track down every paper that is somehow related to the topic. Research synthesists who reject this idea are quite sensible. The point is to avoid missing a useful paper that lies outside one's regular purview, thereby ensuring that one's habitual channels of communication will not bias the results of studies obtained by the search.'

It should be noted that little evidence is available which compares the results obtained using exhaustive and non-exhaustive approaches.

However, given the increasing availability of topic specific and general databases of known published studies and study registers, the amount of effort required in conducting an exhaustive search is usually not prohibitive.

Methods available for searching

Cook *et al.* (9) present a list of possible sources of literature that could be included in a systematic review *(Box 1).*

The most commonly used of these will be considered below, discussing their relative merits.

BOX 1 Possible sources of primary studies for inclusion in a systematic review

- Trial (research) registries
- Computerised bibliographic databases of published and unpublished research
- Review articles
- Published and unpublished research
- Abstracts
- Conference/symposia proceedings
- Dissertations
- Books
- Expert informants
- Granting agencies
- Industry
- Journal handsearching

Research registers

These can be defined as a 'database of research studies, either planned, active, or completed (or any combination of these), usually oriented around a common feature of the studies such as subject matter, funding source, or design.' (10)

In a Cochrane Review, the primary source of studies is a review group's trials register. Where they exist, these are the most valuable source of studies. If all trials, ever carried out, had registered, at onset, then there would be little or no need for other forms of searching. However although registers are on the increase, certainly not all, and especially many older trials, will be found in these.

It is worth noting that the journal *Controlled Clinical Trials* instituted a column that focuses on registers and maintains a 'register of registers' (11). Dickersin (10) also presents a list of research registers.

For an example of the use of research registers, the Oxford Database of Perinatal Trials clearly demonstrates their benefits. This was one of the earliest registers to be set up, and up to 1993, Dickersin (10) reports that over 400 meta-analyses have come out of it alone.

For a further discussion of research registers see pages 132–4, which are dedicated to the subject.

Electronic databases

Another powerful tool for identifying primary studies are electronic databases. They are now available in several formats, including on-line access (via the Internet) and CD-ROM. Although these databases allow access to hundreds of thousands of references, they do have several potential drawbacks. These are discussed below.

Firstly, one should be aware that not all studies are included in even the best databases. For instance, using MEDLINE,[1] only 30–80% of all known published randomised controlled trials are identifiable, depending on the area or specific question (6). Non-English-language references are under-represented in MEDLINE and only published articles are included (4), so there is the potential for publication bias (see chapter 16) (6,12,13) and language bias (4). Depending on the country of origin, there is also potential for geographical biases (1).

Another problem with databases is that even though many of the studies may be in a database such as MEDLINE, it may not be easy to identify all those which are relevant (10). A study investigating this problem (14) reported MEDLINE failed to find 44% of known trials of intraventricular haemorrhage, and 71% of known trials of neonatal hyperbilirubinaemia using 'sensible' search strategies (where the vast majority of RCTs were known to be included in MEDLINE). Possible reasons for poor retrieval are: 1) the search used was too narrow; 2) the indexing of studies in MEDLINE is inadequate ['the precision with which subject terms are applied to references should be viewed with healthy scepticism' (4)]; 3) the original reports may have been too vague, hampering indexing. Clearly only the first point can be easily rectified – a discussion on searching strategies is given below. The possible existence of points two and three highlights the need to use multiple sources to identify studies (4).

A further problem has been reported concerning the search-engine front end of databases. In 1992 Adams *et al.* (15) used SilverPlatter MEDLINE to identify RCTs on a particular topic. They found random deterioration in its ability to cope with an extended, but logical search sequence. The same thing happened with updated software (Silver Platter version 3.1) The authors warn to re-do searches to test their consistency.[2]

In addition to this, other electronic resources and special collections are becoming available on the Internet which may assist in the identification of primary studies. With the number of sources of information increasing it may be worth seeking a librarian's advice on which databases to use (16).

Locating databases to search

For systematic reviews of most (if not all) clinical topics the initial databases to search will be MEDLINE and EMBASE and SCISEARCH. The paragraph below describes the relationship between these databases (1):

> 'Medline provides wide coverage of many English language journals, EMBASE can be used to increase coverage of articles in the European languages. SCISEARCH (the Science Citation Index) can also be used to trace citations of important papers through time, which may yield further useful references.'

[1] One of the largest and most popular medical electronic databases (electronic version of Index Medicus).

[2] It is unknown to the authors whether this fault has been rectified.

The overlap between MEDLINE and EMBASE is approximately 34% (17), although it can vary between 10 and 75% for specific topics (4). It is for this reason that one cannot rely on searching a single database.

Many specialist databases also exist. A directory of computer databases (includes over 6000) exists: A Directory and Data Sourcebook (Detroit: Gale Research, 1991), as well as two other guides; Online Medical Databases[3] and Online Databases in the Medical and Life Sciences[4] which may be useful to check all relevant databases have been identified (1). In addition, many resources and special collections are now being made available via the Internet.

Designing electronic search strategies

It is critical to plan and execute a logical search strategy. Failure to do this may result in wasted time, excessive costs, and irrelevant or missed citations (16).

Indeed this point is made in CRD4:

> 'A balance must be struck between high recall rates and high precision to ensure that whilst a search is relatively comprehensive it does not result in an unmanageable volume of inappropriate references.' (1)

Strategies for retrieving studies

The first step is to identify critical terms, descriptive of the topic under investigation (16). This can be achieved by consulting index manuals, or by identifying appropriate articles and noting the manner in which they have been indexed (1).

The Cochrane Handbook comments: 'Developing a search strategy is an iterative process in which the terms that are used are modified based on what has already been retrieved.' (4)

Many databases use special indexing terms; in MEDLINE they are called medical subject headings (MeSH). The reference lists of these headings should be searched for the ones relevant to the topic of interest. Additional keyword and free-text words (words appearing anywhere in the database entry) will usually be required to supplement index terms.

Most databases structure searches by combining search terms using Boolean relationships (AND, OR and NOT). A broader search can be made using the OR command, and similarly narrower using the AND operator.

It is worth noting that different databases use different indexing and search engines. Therefore, it is necessary to be aware that search strategies developed may need modifying to use on other databases.

Appendix 5c of the Cochrane Handbook (4) gives a search strategy for locating RCTs. In addition, appendix 1 of CRD4 (1) gives examples of search strategies for using MEDLINE to retrieve review articles [one of the most convenient sources of trial references (4)]. Another way of facilitating the searching process is to seek the advice of, or work with, specialist librarians (4). Indeed, the NHS CRD advise that a librarian, preferably with some experience in carrying out systematic reviews, is part of the study team. Once a strategy has been devised, it can be tested by seeing if it picks up key references already known (16).

The indexing used in electronic databases

Dickersin *et al.* (6) comment that the National Library of Medicine introduced the publication type[5] RANDOMIZED CONTROLLED TRIAL (indexed in MEDLINE) in 1991, and from January 1995 introduced CONTROLLED CLINICAL TRIAL (CCT) (defined by Cochrane Collaboration's criteria, and was used to index trials not contained in RANDOMIZED CONTROLLED TRIAL). As handsearching is done these terms will be applied retrospectively to previously unindexed trials.

Counsell and Fraser (18) along with Dickersin *et al.* (6) call for better, more consistent and more specific indexing of papers (trials) in MEDLINE.

[3] Lyon E (1991) Online medical databases. London: Aslib.

[4] Online Database in the Medical and Life Sciences (1987) New York: Cuadra/Elsevier.

[5] This is another field used to index in MEDLINE. One can restrict the search to certain types of articles using the publication type field. However, since RCT was only introduced in the 1990s, at present it is necessary to use the previously mentioned strategies to identify RCTs reported earlier than this and in other databases. However, with retrospective indexing this feature should greatly aid the retrieval of RCTs and CCTs in the future.

The above points in the main relate to RCTs, similar issues are also important when searching for observational studies.

The book chapter by Reed and Baxter (16) is dedicated to electronic database searching. This includes a section explaining many different databases, and provides a list of the most common ones. It is recommended reading for a more detailed description of the contents of this section.

Citation searches

In carrying out a citation search, the searcher begins by identifying several important references on the topic of interest. A citation index will identify, for a specified time period, all articles, reports, or other materials in the body of literature it covers that have cited the important references identified (16).

The advantages of this method are that it: '... allows the researcher to avoid reliance on the subjectivity of the indexers. It also avoids the inherent currency lag and biases of controlled vocabularies.', and allows cross discipline referencing (16).

The ISI databases as well as SCIEARCH allow citation searching; however, the ISI databases are restricted to journal articles.

Citation searches are not frequently carried out. Cooper (19) reported in his survey that only 14 and 9% of reviews do citation searchers (manually and computerised, respectively) and he considers this to be 'disturbingly low'. However, although citation searching tends to produce different 'hits' from searches using natural language and controlled vocabulary, the Cochrane Handbook (4) states that 'insufficient evidence is currently available to suggest that routine use of citation searches is warranted, given the costs involved.'

Extensions to citation searching

Reed and Baxter (16) suggest an extension to the searching process. This is to find papers which cite two specified papers. A CD-ROM ISI innovation makes it possible to retrieve articles that are 'bibliographically coupled', i.e. cite identical references. ISI also makes 'research fronts' retrievable on-line by entering a number code. These research fronts identify 'clusters' of papers which have cited similar references. For example, a research front for the subjects 'meta-analysis of clinical trials; test validation; validity generalization' exists and all papers in this set can be retrieved simply using code RF number 9324_94 (ISI – through BIDS).

Other search strategies

Scanning reference lists (footnote chasing)

Scanning the reference lists of articles found through database searches may identify further studies for consideration (1). The Cochrane Handbook advises:

> 'Reviewers should check the reference of all relevant articles that are obtained. Additional, potentially relevant, articles that are identified should be retrieved and assessed for possible inclusion in the review. The potential for reference bias (a tendency to preferentially cite studies supporting one's own views) when doing this should be kept in mind.' (4) This should be guarded against by using (several) other strategies.

The idea of reference bias was originally suggested by Sackett (20). His study (20) found evidence of reference bias and also commented on many multiple publications of (the same) trials, another potential source of bias to be aware of when carrying out a review (see chapter 16).

Handsearching

Key journals can be handsearched to check if the searcher has missed anything using the alternative methods, e.g. due to problems such as things badly indexed in electronic databases. By handsearching carefully selected journals, a small amount of work can reveal a high percentage of relevant studies.

It is worth noting that currently, the Cochrane Collaboration is organising handsearches of entire series of journals with all studies found being indexed. This is being coordinated to avoid duplication of work with the intention of developing an International Register of Clinical Trials and hence eliminating the need for individual research groups to carry out retrospective handsearching.[6]

Identifying grey material

Results may have been published in reports, booklets, conference proceedings, technical reports, discussion papers or other formats which are not indexed on the main databases (21,22). All these sources can be called 'grey literature'. Identifying such literature is not easy, however,

[6] Note: It is a good idea to inform the Cochrane Centre of any unusual/poorly indexed trials you locate for inclusion on their register.

databases do exist, such as SIGLE (System for Information on Grey Literature), NTIS (National Technical Information Service), DHSS-Data, and the British Reports, Translations and Theses, which is received by the BLDSC (British Library Document Supply Centre). One should be aware that even if you identify material such as conference reports, obtaining them may be a problem (16).

In addition, results may exist in interim reports, unsubmitted papers and manuscripts, presented papers (not published), rejected papers, and partly completed reports (23), most of which will not be included in the above databases. Clearly identifying and getting hold of this information can be extremely difficult. Possibly the best chance one has is through personal communication with the relevant researchers, either formally or informally depending on appropriateness (4). Other approaches are to use electronic networks/lists, contact with public policy organisations and advertising (23). It is important to point out that the inclusion on grey literature, such as unpublished studies is somewhat controversial. It has been argued that since it has not been peer reviewed, it may be of dubious quality (see chapter 6 for more on this topic).

The book chapter (23) deals exclusively with this subject and is recommended reading for researchers wanting more detailed information on the methods available.

A few comments specific to different forms of grey literature are given below.

Conference proceedings. These are a good source for information on research in progress as well as completed work. A note of caution is that data from the abstracts of conference proceedings is notoriously unreliable. For this reason, an attempt to make contact with the authors and obtain any other relevant information/reports should be made (1). Conference proceedings are recorded in several databases in including the Index of Scientific and Technical Proceedings (available via the BIDS), the Conference Papers Index (available via Dialog) and in printed forms such as the Index of Conference Proceedings received by the BLDSC.

Consultation (with leading researchers and practitioners). The Cochrane Handbook states: ‘Experts in the topic of the research synthesis can be an important source of information on recent trials that have not yet been published, or on older trials which were never published.’ (4) It also suggests making a list of relevant articles and sending it with a letter asking whether they know of any other relevant trials (published or not) in the field.

White (8) strongly encourages this method saying can be very fruitful: ‘The only danger lies in reliance on a personal network to the exclusion of alternate sources.’ Its strength is at finding unpublished studies. One has to be aware of selection bias when doing this.

Consultation with the pharmaceutical industry. A similar approach to above can be taken to contacting pharmaceutical companies. They may be willing to release results that have not already been published (1).

Problems and issues with searching

As hinted above one needs to be aware that searching more than one database is necessary, due to differential coverage (16). In addition most publications pre-mid-1960s are not in electronic form. In addition, mainstream sources such as book chapters are usually not referenced in databases, which is a problem.[7]

Reporting searching

The reporting of the search strategy (even if it has been carried out well) is also often neglected.

The failure of almost all integrative review articles to give information indicating the thoroughness of the search for appropriate primary sources does suggest that neither the reviewers nor their editors attach a great deal of importance to such thoroughness (8).

However, the Cochrane Handbook states that: ‘The search strategy should be described in sufficient detail that the process could be replicated.’ (4)

Indeed, a format for the necessary details of the search process that should be included in the final report, has been advocated (8).

[7] PsycLIT did include book chapters but only for the years 1987–1990.

Selecting studies

Judging study relevance

The list generated by the search strategy should firstly be inspected. 'If, given the information available, it can be determined that an article definitely does not meet inclusion criteria, it can be excluded. If the title or abstract leave room for doubt in the reviewer's mind that the article cannot be definitely be excluded, the full text of the article should be retrieved (4).' In deciding which articles to include the researcher should initially err on the side of caution.

As White points out (8), this should not be too difficult as one should have knowledgeable and motivated people reading the articles.

The selection process

The articles selected through the search process must be assessed to see whether the inclusion criteria for the review have been met.

The Cochrane Handbook (4) lists the following issues that must be decided upon:

- whether more than one reviewer will assess the relevance of each article
- whether the decisions concerning relevance will be made by content area experts, non-experts, or both
- whether the people assessing the relevance of studies will know the names of the authors, institutions, journal of publication and results when they apply the inclusion criteria
- how disagreements will be handled if more than one reviewer applies the criteria to each article.

A suggestion made (4) is to have two reviewers, one an expert in the field and one who is not to safeguard against pre-formed opinions. However, we note that much of this will depend on the time available and how difficult (subjective) are the opinions needed.

It has been suggested that reviewers should be blinded from information such as source, authors, institution and magnitude and direction of the results by editing articles, with the intention of removing reviewer prejudices. This, however, takes much time and there is no empirical evidence suggesting benefits from doing so (4).

Any disputes about inclusion/exclusion can usually be cleared up by discussion between reviewers. If this is not the case additional information should be sought (4). Deeks *et al.* (1) also comment that any disagreement on inclusion can be explored using a sensitivity analysis (see pages 209–10).

Note that it is recommend to pilot test the inclusion criteria so it can be refined and clarified (4).

A final word of warning is to be aware of the potential of language bias (24); this occurs when inclusion criteria are limited by the language of the study report and there is an association between effect size and language of publication. For example, one could argue that highly significant studies would be published in high profile English language journals, while researchers in non-English speaking countries may be more likely to publish non-significant results in native journals. In this case, only considering English language journals would produce a biased set of trials. A related issue is multiple publication bias. One should be aware that the same trials results can be published in more than one place, it is necessary to check whether each study found reports data which is exclusive from the other studies, otherwise results may be included twice into a meta-analysis (see chapter 16 for more information).

Documenting the selection process

This selection process should be documented; a Cohen's Kappa (a measure of chance corrected agreement) statistic can be used to describe the reproducibility of the decisions of the assessors (1).

The final report should contain tables detailing studies included and excluded from the synthesis, with reasons given for each exclusion (1).

Assessing study validity

Deeks *et al.* comment: 'The assessment of validity aims to grade studies according to the reliability of their results, so that they can be given appropriate weight in the synthesis, and when drawing conclusions'. (1) The aim of this is to reduce bias, with high quality studies likely to be least biased. Primary studies can be graded into a hierarchy according to their design. Ideally, a review will concentrate on studies which provide the strongest evidence, but where only a few good studies are available weaker designs may have to be considered (1).

Table 1 is reproduced from (1), giving a suggested hierarchy to the various study designs commonly used.

Table 1 should be used with caution, however, since the study validity not only depends on the type of study but how well it was designed, carried out and analysed. 'A poor RCT may be less reliable than a

***TABLE 1** [adapted from (1)]. An example of a hierarchy of evidence*

I	Well-designed randomised controlled trials
	Other types of trial:
II–1a	Well-designed controlled trial with pseudo-randomisation
II–1b	Well-designed controlled trials with no randomisation
	Cohort studies:
II–2a	Well-designed cohort (prospective study) with concurrent controls
II–2b	Well-designed cohort (prospective study) with historical controls
II–2c	Well-designed cohort (retrospective study) with concurrent controls
II–3	Well-designed case–control (retrospective) study
III	Large differences from comparisons between times and/or places with and without intervention (in some circumstances these may be equivalent to level II or I)
IV	Opinions of respected authorities based on clinical experience; descriptive studies and reports of expert committees

well conducted observational study.' (1) Chapter 6 outlines the various methods of assessing the quality of primary research, that can be used for meta-analysis.

Fleiss and Gross have made a comment on including studies of mixed validity (25):

> 'There is considerable doubt about the validity of statistically combining the results of studies with different designs or synthesising results of observational or uncontrolled studies, on the grounds that this might pool biases.'

Recently, however, attempts to do just that have been made. Chapter 26 outlines methods for the generalised synthesis of evidence, proposed for this purpose.

Methods of data extraction

The protocol should contain an example of a data extraction form which lists the data items to be extracted from each of the primary studies (26). Data extraction is best done using special forms, examples are given in appendix 3 of (1), although these may have to be modified for a particular meta-analysis.

Deeks *et al.* (1) recommend, due to the risk of errors, data extraction should be done independently by at least two people and the level of agreement ascertained. However, time and research constraints make this difficult. Any disagreements that cannot be resolved should be investigated in a sensitivity analysis (see pages 209–10).

Missing data may be a problem, if this is the case the authors of the original studies should be contacted – if this proves unfruitful, then statistical methods do exist for dealing with missing data (see chapter 17).

There is always the possibility of contacting the original researchers for every study located and requesting individual patient data (IPD). If all the data is received then the analysis can be based at the patient level (as opposed to the study level) (see chapter 24 which describes the relevant methods). Even if the intention is not to carry out an analysis at the individual patient level, Cook *et al.* (9) recommend obtaining individual patient level data when the published data do not answer questions about: intention to treat analyses, time-to-event analyses, subgroups, dose–response relationships.

If the primary reports do not present data in the way desired for synthesis, then it may be possible to transform or estimate the desired values. Techniques specific to epidemiological studies have been developed (27) and are outlined on pages 148–52. Also, it may be possible to contact the original authors for missing data/or new analyses.

Comparative investigations of searching

The following have studied the effects of different search methods on research synthesis.

Dickersin *et al.* (6) compared state of the art (hand and MEDLINE) with only MEDLINE searches of different types. They concluded that using MEDLINE only omitted half of the relevant studies. Additionally, Clarke (28) gives an example of searching, performed for a meta-analysis, and highlights how MEDLINE alone was not sufficient. Also, Adams *et al.* (29) summarise further investigations into searching using MEDLINE, and conclude that between 20 and 60% of RCTs are missed by skilled MEDLINE searches when compared to handsearching or using trial registers. Spoor *et al.* (30) used capture–recapture techniques to compare searching an electronic database with handsearching. They found that MEDLINE missed 35 relevant articles, handsearching (human error) missed eight, with an estimated two articles [95% confidence interval (CI) 0–6] were missed by both techniques. Dickersin *et al.* (14) compared MEDLINE with a Perinatal Trials Database. Two MEDLINE searches were carried out; one by an expert, and the second a 'quick and dirty' one (the original paper gives both search strategies). The authors note that no abstracts are held in MEDLINE pre-1975, so text searching is less effective before this date. They concluded that most of the trials are in MEDLINE, but a search has to be very broad to retrieve them all.

Jadad and McQuay (31) investigated; 1) the time involved in identifying pain research reports published in specialist journals in 1970, 1980, and 1990 using a refined search strategy for MEDLINE and hand searching; 2) the levels of precision and sensitivity of the MEDLINE search strategy over a 20-year period and to determine the causes of failed identification; 3) methods to determine efficient combinations of MEDLINE and selective hand search to achieve high sensitivity and minimal cost. Among their finding was the result that MEDLINE was most time efficient; it identified 87% of known trials with 52% precision, and the search took one-tenth of the time of that of hand searching. The same authors (Jadad and McQuay) come to the defence of MEDLINE (32) commenting that when used correctly:

> '(A) restricted "pilot" hand search to refine the strategy, followed by a high yield Medline search and hand search of non-indexed journals, may be a cost effective way of meeting the fundamental challenge.' (32)

Kleinjen and Knipschild (33) investigated to see if computer database searches alone were sufficient for locating studies. They used MEDLINE and EMBASE and explored three subject areas. They concluded that number of articles found with computer searches depends very much on the subject at hand, and that the better methodological studies were found (on the whole) in the electronic databases. Gotzsche and Lange (34) compare different search strategies for recalling double-blind trials from MEDLINE. They conclude that using 'comparative study' as a MeSH term is better than using, 'double-blind method' (even when it is used as a text word also), However, the success of both terms was > 90%, which they comment was surprisingly high, much higher than previous studies.

Further research

There is a lack of helpful research which allows both the quality of research and also design to be put on one validity scale.

Summary

This section has concentrated on searching the literature and identifying primary studies that might potentially be included in a systematic review or meta-analysis. The main point identified is that there is no one single search strategy that would provide adequate results, and that in performing reviews researchers should maintain a healthy degree of scepticism about any or all their searches. However, a second key point is that all searches/ methods that are used should be sufficiently well documented so that they may be replicated by other researchers. This latter point is equally important as regards study inclusion/exclusion.

Finally, changes are happening rapidly in terms of electronic publishing and databases. Such changes will undoubtedly have profound implications for conducting systematic reviews in the future.

References

1. Deeks J, Glanville J, Sheldon T. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. Centre for Reviews and Dissemination, York: York Publishing Services Ltd, Report #4, 1996.

2. Simes RJ. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* 1986;4:1529–41.

3. Simes RJ. Confronting publication bias: a cohort design for meta-analysis. *Stat Med* 1987;**6**:11–29.

4. Oxman AD, editor. The Cochrane Collaboration handbook: preparing and maintaining systematic reviews. Second edn. Oxford: Cochrane Collaboration, 1992.

5. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science (review). *Epidemiol Rev* 1992;**14**:154–76.

6. Dickersin K, Scherer R, Lefebvre C. Systematic reviews – identifying relevant studies for systematic reviews. *BMJ* 1994;**309**:1286–91.

7. Laird NM, Wachter KW, Straf ML, editors. A discussion of the Aphasia Study. In: The future of meta-analysis. New York: Russell Sage Foundation, 1992, p. 47–54.

8. White HD, Cooper H, Hedges LV, editors. Scientific communication and literature retrieval. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 41–56.

9. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis (review). *J Clin Epidemiol* 1995;**48**:167–71.

10. Dickersin K, Cooper H, Hedges LV, editors. Research registers. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 71–84.

11. Dickersin K. Keeping posted. Why register clinical trials? – revisited. *Controlled Clin Trials* 1992;**13**:170–7.

12. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data (with discussion). *J R Statist Soc A* 1988;**151**:419–63.

13. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith HJ. Publication bias and clinical trials. *Controlled Clin Trials* 1987;**8**:343–53.

14. Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Pursuing the literature: comparison of MEDLINE searching with a perinatal trials database. *Controlled Clin Trials* 1985;**6**:306–17.

15. Adams CE, Lefebvre C, Chalmers I. Difficulty with MEDLINE searches for randomised controlled trials. *Lancet* 1992;**340**:915–16.

16. Reed JG, Baxter PM, Cooper H, Hedges LV, editors. Using reference databases. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 57–70.

17. Smith BJ, Darzins PJ, Quinn M, Heller RF. Modern methods of searching the medical literature. *Med J Aust* 1992;**157**:603–11.

18. Counsell C, Fraser H. Identifying relevant studies for systematic reviews. *BMJ* 1995;**310**:126.

19. Cooper HM. Literature searching strategies of integrative research reviewers. *Am Psychol* 1985;**40**:1267–9.

20. Sackett DL. Bias in analytic research. *J Chron Dis* 1979;**32**:51–63.

21. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, *et al.* Should unpublished data be included in metaanalyses – current convictions and controversies. *JAMA* 1993;**269**:2749–53.

22. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I: control of bias and comparison with large co-operative trials. *Stat Med* 1987;**6**:315–25.

23. Rosenthal MC, Cooper H, Hedges LV, editors. The fugitive literature. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 85–96.

24. Egger E, ZellwegerZahner T, Schneider M, Junker C, Lengeler C. Language bias in randomised controlled trials published in English and German. *Lancet* 1997;**350**:326–9.

25. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127–39.

26. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Int Med* 1987;**107**:224–33.

27. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.

28. Clarke M. Searching Medline for randomised trials. *BMJ* 1993;**307**:565.

29. Adams CE, Power A, Frederick K, Lefebvre C. An investigation of the adequacy of Medline searches for randomized controlled trials (RCTs) of the effects of mental-health-care. *Psychol Med* 1994;**24**:741–8.

30. Spoor P, Airey M, Bennet C, Greensill J, Williams R. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. *BMJ* 1996;**313**:342–3.

31. Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. *Online J Curr Clin Trials* 1993;Doc. No. 33.

32. Jadad AR, McQuay HJ. Searching the literature: be systematic in your searching. *BMJ* 1993;**307**:66.

33. Kleinjen J, Knipschild P. The comprehensiveness of Medline and Embase computer searches. *Pharm Weekbl [Sci]* 1992;**14**:316–20.

34. Gotzsche PC, Lange B. Comparison of search strategies for recalling double blind trials from MEDLINE. *Dan Med Bull* 1991;**38**:47–68.

Part C:

Results II – the critical appraisal of research evidence

Chapter 5

Example: the cholesterol lowering trials

Introduction

To clarify some of the methods discussed in the remainder of this report, a practical example has been included. A dataset consisting of trials investigating the effect of lowering serum cholesterol levels is used to illustrate methods for binary outcomes; this is described below. Additionally a further example (the effect of mental health treatment on medical utilisation) is described, and used, in chapter 9, to illustrate the analysis of continuous outcomes.

Effect on mortality of lowering serum cholesterol levels

Since 1962 several studies have investigated the effect of lowering cholesterol levels on the risk of death, primarily from coronary heart disease but also from all other sources. This dataset consists of 35 RCTs, originally compiled by Smith, Song and Sheldon for a meta-analysis (1) (see this paper for a listing of references to these trials). Only a subset of these 35 RCTs will be used, primarily to reduce the amount of computation required for the purposes of illustration. The subset of trials chosen comprises of those trials in which patients were largely without pre-existing cardiovascular disease. In the original report the trials were numbered 1–35. The subset of trials considered here (initially in chapter 9), in order to be consistent with this numbering, are labelled 16, 20, 24, 28, 29, 30, 31. This will enable the interested reader to cross refer back to (1). The subset consists of the trials that used cholesterol lowering as a primary intervention.

It should be noted that since this list was compiled, further studies have been carried out. It therefore should be stressed that the analyses presented are to **illustrate** the various methods discussed in the report, and are in no way meant to indicate a definitive analysis.

Reference

1. Smith GD, Song F, Sheldon TA, Song FJ. Cholesterol lowering and mortality: the importance of considering initial level of risk. *BMJ* 1993;**306**:1367–73.

Chapter 6

Study quality

Introduction

It has been noted (1) that the subject of judging research quality in synthesis dates back to Glass in 1976 (2). The primary concern is that combining study results of poor quality may lead to biased, and therefore misleading, pooled estimates being produced. Detsky *et al.* put this more precisely in statistical terms:

> 'The effects of quality on the study's estimate of effectiveness can be expected to have two components, bias effects and precision (added variability) effects. ..... Meta-analyses that combine studies of varying quality could suffer from bias resulting in a Type 1 error or a Type 2 error. Meta-analyses that combine studies of varying quality could also suffer from a lack of precision resulting in a Type 2 error.' (3)

Thus, there is a real danger of producing misleading results if the quality of the primary studies is dubious. This has led to a warning from Thacker to those who do not consider the quality of their data '....sophisticated statistics will not improve poor data, but could lead to an unwarranted comfort with one's conclusions'. (4)

The importance of considering the quality of the primary studies was again highlighted by Naylor:

> '... in some respects, the quantitative methods used to pool the results from several studies in a meta-analysis are arguably of less importance than the qualitative methods used to determine which studies should be aggregated.' (5)

However, assessment of quality is not without its controversies. Greenland (6) has indicated that quality assessment is the most insidious form of bias in the conduct of meta-analysis.

Detsky *et al.* (3) identify three basic issues that need addressing when considering study quality in research synthesis:

- How much does quality matter?
- How best can we measure quality?
- Incorporating measure of quality in a meta-analysis.

They state the answer to the first question is unknown and that there is no 'gold standard' method for part two either, adding that different methods seem to produce reasonably congruent results. Again there is no one 'right' approach to point three.

In addition, it has been pointed out that one of the roles of meta-analysis should be to clarify, or even quantify, weaknesses in the existing data on a scientific question and to encourage better quality in future studies (7).

The section below outlines ways in which the quality of studies may vary. This is then followed with a description of the various approaches to assessing and dealing with study quality that meta-analysists have taken. As another review group (8) is covering this topic, its treatment here is quite brief. In addition, for an excellent review of scoring systems see Moher (9).

Methodological factors that may affect the quality of studies

Table 1 suggests a possible hierarchy to the sources of best evidence. The reasoning behind this was that different study designs are susceptible to biases in varying degrees, and thus vary in the reliability of the results. It has become accepted that RCTs are the 'gold standard' source of evidence, giving unbiased estimates of intervention effects. However, no empirical measure of the amount of bias, on average, that other study designs are susceptible to is available.[1] Despite this, due to specific features which are known to increase/reduce bias, such as matching, collecting the data retrospectively, and using a historical comparison group, an argument can be made for the superiority of evidence from one study design over another, although it is not possible to quantify how much more superior it is. As one can only determine the methodological

[1] Although Macarthur *et al.* (46) note three broad categories of bias are generally recognised, namely, sample distortion bias, information bias, and confounding bias.

quality of a study to the extent that study design and analytic methods are reported (10), we restrict ourselves to such factors in the remainder of this chapter.

CRD4 (11) states that the first division that can be made is between experimental and observational studies. So we shall focus initially, on clinical trials.

Experimental studies

Moher *et al.* (9) suggest the design features of trials, which effect the trials quality, and can be assessed, can be split into four areas, namely assignment, masking, patient follow-up, and statistical analysis.

Assignment

This could well be the single most important design feature of a study. As randomised controlled trials provide the most valid basis for the comparison of interventions in health care (12), it is clearly a desirable feature and thus RCTs are considered the most reliable method on which to assess the efficacy of treatments (13).

Despite this, the details of randomisation are not often reported (12). Another disturbing problem stemming from the 'unnatural' balance of numbers in the arms of many trials, is that, there is evidence that unadulterated randomisation has not gone on, i.e. groups are 'too equal' (this is sometimes called random manipulation) (12). The motivation for this 'fudging' is that researchers believe equal groups increases the credibility of the results. This therefore raises the question, even if a study is described as randomised, can you believe it? This is also important if the outcome is affected by baseline value, e.g. size of wound when the outcome is percentage of wound healed or absolute reduction in size.

Masking/blinding

Blinding is generally desirable in trials to minimise biases. Patients are said to be blinded if they do not know which intervention arm of the study they are in. Similarly the health professional administering treatment is blinded if they do not know which treatment the patient is getting. Finally, the person assessing the effect of the intervention may also be blind. This will be particularly important when the outcome measurement of interest involves some subjective/human judgement. Obviously, by the nature of some interventions, blinding of one or more of the above groups of people may not be feasible. Allocarion concealment relates to patient assignment in which the masking may not be kept after patients are allocated. Bias has been detected in trials not reporting adequate allocation concealment (14); however, this is thought not to be as important as the generating of assignments *per se* (12,15).

Patient follow-up

In trials, patients drop out for several reasons. Patients may also switch to other arms, in instances such as when the patient was allergic to the original treatment. How these events are documented and subsequently dealt with in the analysis can effect the overall treatment estimate. This is also a huge potential source of bias.

Statistical analysis

Obviously if an inappropriate statistical analysis was carried out, or a correct type of analysis, but with mistakes was produced, this could lead to misleading results.

Other

In addition, if crossover designs have been used, if they are used inappropriately (such as in fertility treatment where they are quite often misused), this will produce strongly biased results (16).

Observational studies

Since treatment allocation is left to a haphazard mixture in observational studies (11), this is one reason why they have a greater susceptibility to bias than clinical trials. Similarly, ascertaining that differences observed between groups of patients in observational studies are the effect of the interventions is a far harder exercise than it is in experimental studies (11). Cohort studies, in which groups receiving the different interventions being compared are evaluated concurrently, are regarded as more valid than studies which make comparisons with 'historical' controls (11). Similarly, studies which are planned prospectively are also less likely to be biased than studies which are undertaken retrospectively (11). Case–control studies are prone to many extra biases, and therefore fall below cohort studies in the hierarchy (11). If they are included in a meta-analysis it may be possible to grade them according to the suitability of choice of the control group. It is also worth being aware that treatment effects could be underestimated due to over-matching on factors which are related to allocation of the intervention (11).

However, clearly the study design is not the only factor which effects the quality of a study. How well it was designed, carried out and analysed all contribute to its quality. In this way a poor RCT may be less reliable than a well conducted observational study (11). Assessment of such factors can be made more systematic with the use of checklists (11); these are discussed on pages 25–6.

Evidence of bias and study quality

Several studies have been carried out to investigate the effect of study quality on the magnitude and direction of the results (14, 17–21). The findings from these studies are variable and not totally consistent. Several of these, contrasting in their findings, are outlined below.

Emerson *et al.* (17) investigate whether a relationship exists between treatment difference magnitude and a given quality score for a selected groups of studies. They found no evidence for this relationship. A possible explanation, put forward by the authors, for this, is that the studies assessed came form previous meta-analyses and may have been of greater than average quality. This led the authors to comment that the result leans to recommending the inclusion of all RCTs in a meta-analysis and not adjusting weights for quality either.[2]

Colditz *et al.* (19) investigate the association of study type with result and concludes: ‘We observed that several features of study design influence the likelihood of a report that patients perform better on the innovation than on standard therapy. These features included randomisation, blinding, the use of placebo and the inclusion of patients refractory to standard therapies.’ The authors go on to suggest that one may wish to adjust for the average level of bias associated with a given design when pooling studies in a meta-analysis and suggest values for each design feature; however, the authors of this report are not aware of any instances when they have been used.

Studies have been undertaken to investigate the difference randomisation makes in a study. As Wortman reports (1), designs that used non-random allocation overestimated the effect by at least one-third. However he also notes that systematic biases in quasi-experiments can also underestimate effects. Also, Sowden *et al.* (22), in a review of observational studies, found that the effect size varied according to the quality of adjustment for case mix.

Schultz *et al.* (21) set up an investigation to determine whether inadequate approaches to randomised controlled trial design and execution are associated with evidence of bias in estimating treatment effects. They investigated the effect of inadequate allocation concealment, exclusions after randomisation, and lack of double-blinding. They found that larger treatment effects were reported when concealment was either inadequate or unclear, trials that were not double-blinded yielded larger estimates of effect size, and there was no association with effect size for trials which excluded patients after randomisation.

Assessing the quality of studies

As Wortman states (1), the literature contains two approaches for coding research quality. The first system (23) applies the validity framework developed by Campbell *et al.* (24). This approach provides a matrix of designs and their features or ‘threats to validity’. Its focus is on non-randomised studies found in the social science literature. [For this reason it will not be pursued further here, and the interested reader is referred to the above cited papers and (1).]

The second system was developed by Chalmers *et al.* in 1981 (25). This was later extended by them to a framework for the ‘quality assessment’ of meta-analyses (26). It concentrates on the randomised control trial study design, and has the objective of providing an overall index of quality rather than the estimation of bias.[3,4]

Many different checklists and scales have appeared in the literature, initially for trials but now scales are available for assessing observational studies also. Another project in this series has been commissioned to look at the quality of randomised controlled trials exclusively (8).

For more information on assessing the quality of quasi- and uncontrolled experiments, see Wortman (1).

[2] These are two possible methods of incorporating study quality into the statistical analysis. These are discussed on pages 26–9.

[3] It could be argued that using individual markers could be considered as a third (27); however, they are not considered further here.

[4] (1) Includes a detailed comparison of the two methods. The largest differences are that Campbell’s encompasses a larger variety of designs (randomised and non-randomised), while Chalmers’ is more in depth for just randomised trials, and includes a scoring system.

Checklists and scales for trials

The first checklist for trials was published in 1961 (9) and eight more had been published by 1993 (9).The first scale for assessing the quality of trials was published in 1981 (25). By 1993, an additional 24 scales had been developed (9).The interested reader is referred to these references through the excellent review article (9) for more specific detailed information.

Although many of these scales and checklists for trials are similar (though many are designed only for RCTs[5]), they emphasise different dimensions of quality. Moher *et al.* (27) assessed the variability of using different scales and found it was considerable. For this reason the content of the checklist should be stated in the protocol (11).

These scales have their critics. Jadad *et al.* comment: '... there is a dearth of evidence to support the inclusion or exclusion of items and to support the numerical scores attached to each of those items.' (28)

Another problem is the effect that the level of reporting has on the quality score. Jadad *et al.* comment:

> 'Given space constraints in most journals, editorial decisions may end up having a major effect on the quality score achieved by a given study.' (3)

They go on to state that, incomplete reporting may be avoided in the long term if journals adopted more uniform reporting standards for trials and authors routinely made additional protocol details available on request.

Schulz *et al.* (12) state, as a very minimum, reports of RCTs should include: 1) the type of randomisation, 2) the method of sequence generation, 3) the method of allocation concealment, 4) the persons generating and executing the scheme and 5) the comparative baseline characteristics.

With the arrival of the Consolidated Standards of Reporting Trials (CONSORT) statement (29) (a list of 21 items that should be included in a report as well as a flow chart describing patient progress through the trial), hopefully this issue should no longer be a problem.

Moher *et al.* (9) concluded, from their investigation of 25 scales, that all but one of them have major weaknesses, not least that they have evolved with little or no standard scale development techniques. Other than the exceptional scale (28), they chose items from 'accepted criteria' from standard clinical trial textbooks. Moher *et al.* commented that:

> 'Although these criteria may be useful, some of them are based on conviction whereas others are based on empirical evidence.' (9)

The illustrative example of informed consent, which is included in some checklists, is given, and the authors question how this affects the quality of the study.

Checklists for observational studies

Although much smaller in number, checklists do exist for epidemiological studies which assess potential links between exposures to risk factors and harm. CRD4 (11) reports of three for general use (30–32). In addition, at least three others (33–35) have been developed specifically for meta-analysis (36). Two studies (35,37) have demonstrated associations between their relative risks (RRs) and quality scores. The guidelines in CRD4 suggest the same checklists can often be used to assess the strength of evidence from observational studies investigating treatments which are of benefit (rather than risk factors) as the same issues are important (11).

Other checklists

Checklists are also available for studies which assess the accuracy of a diagnostic test. A different checklist is needed because these studies are affected by several different and more complicated issues. Also, separate checklists are available for reviews of economic evaluations. In instances when one is considering non-comparative studies, such as case series, checklists which assess articles on prognosis can be used. Details of all these scales and checklists can be found in CRD4 [(11), p. 31–7].

Incorporating study quality into a meta-analysis

Once a formal assessment of study quality has been made, the next question is: should the measure of quality be incorporated into the analysis, or just used as a threshold value for including/excluding studies?

This is a problematic question which has produced differing opinions, some of which are highlighted below.

[5] The checklists described by Spitzer *et al.* (34) and Cho and Bero (10) are noteworthy as they are applicable to both experimental and observational studies.

A trade-off between the precision and the accuracy of the estimate of effect may exist if studies of variable quality are to be combined. Whilst inclusion of more studies may allow a more precise estimate (tighter CIs), if this is done by including studies of dubious validity, then it may be at the expense of accuracy (3).

There has been doubt as to the interpretation of a quality score. Detsky *et al.* ask:

> 'Is the quality of a trial a continuous characteristic or is there a threshold effect of quality? .... It seems highly unlikely that these scales would result in a linear or monotonically increasing relationship to true quality.' (3)

They conclude:

> '... the relationship between quality scoring systems and the degree to which the study results approximate the truth should be viewed with some caution.' (3)

Indeed, no general relationship has been found that links quality score and magnitude of outcome (17), although Colditz *et al.* consider the idea to be attractive (38).

Below are the outlines of methods proposed to incorporate an assessment of study quality into a meta-analysis.

Graphical techniques

It has been suggested (3) that a plot of the point estimate and 95% CI for each studies treatment effect against quality score (derived from a scoring system), can be investigated to see if there is any trend between the two variables.

An equivalent way of investigating, essentially the same thing, is to include a variable for study quality in a logistic regression model (these are covered in chapter 11 on meta-regression); however, this formal test would often lack power due to too few studies in many meta-analyses (3).

Weighting

Rather than weight by sample size (see chapter 9), one could weight each of the individual estimates by a variable which measures the perceived quality of the study (3). For the log odds ratio scale, this can either be done by hand, or any statistical package capable of performing weighted logistic regression.

In doing this one has to be aware that:

> 'Although actual estimates, such as the pooled odds ratio, are affected only by the relative weights used, the width of the confidence intervals is affected by the weights used' (3)

For example if a study is weighted by a quality score of 0.5, it is equivalent to an unweighted study with half the sample size. Detsky *et al.* go on to comment that:

> 'Although a widening of the confidence intervals is probably called for, the amount of the widening that automatically results by weighting (in logistic regression) is completely without empirical support' (3).

The amount of widening can be modified by multiplying scores by a constant. Increasing all the scores could be justified by arguing that a score of 1 is almost impossible to achieve. An alternative procedure would be to divide each trials score by the mean, by doing this CI widening does not result.

Detsky *et al.* still do not consider this form of analysis as satisfactory as the below explanation will testify:

> 'weighting by either the unadjusted or adjusted scores to control the increase in the width of the confidence intervals is difficult to defend, since there is no *a priori* reason why this process should alter Type 1 and Type 2 error rates in such a way as to move the aggregate effect size estimate "closer to the truth". This is because, while weighting study estimates by the precision has very desirable optimality properties, quality scores are not direct measures of precision.' (3)

The authors go on to comment that, what is desired is a method of determining the relationship between quality scores and precision (bias). So, ideally weights would be determined by sample size, inherent binomial variation, quality induced variation and quality induced bias (the last of these, however, would be difficult to ascertain).

Excluding studies

Another approach is to exclude the studies of poor(est) quality altogether. This can be viewed as an extreme form of weighting, giving the poorest studies no weight at all. Light justified this approach by arguing:

> 'if it is clear that a certain study is fundamentally flawed, say with obvious numerical errors, I find it hard to argue for its inclusion. I do not believe that wrong information is better than no information' (39)

To determine what classifies as a poor quality study, a threshold value needs to be produced. If a scoring system such as those outlined in this chapter is used, figures such as the mean, the mean plus one standard deviation or the median can be used as this threshold (3). Alternatively simpler criteria can be used, such as whether randomisation was fairly performed, or whether there were blinded outcome assessments (3). There does not appear to be a consensus as to the optimal rigour used in deciding whether to reject studies. Some authors recommend inclusion of all but the very worst of studies (40) (a quality weighting scheme could still be applied to remaining studies to be included), while others advocate the exclusion of all but the best studies. One of the supporters of the latter approach is Slavin, who has promoted an approach to pooling called best evidence synthesis (41), where all but the methodologically most adequate studies are excluded. An outline of this approach is given on page 216.

Sequential methods

Detsky *et al.* (3) suggest this method. It can be viewed as a form of sensitivity analysis (see chapter 27). A cumulative meta-analysis based on a quality score is conducted, i.e. trials are combined sequentially from the highest to lowest quality and a pooled estimate is calculated for each new addition and plotted. The authors state: 'An investigation of this graph will then provide an opportunity to discern the effect of quality on estimated effect size'. The authors go on to comment that this method as several advantages: i) it uses quality scores simply to rank order trials for the exploration of quality effects, and as such is free of further assumptions about the relationship between scores and 'true' rigour, ii) the method basically draws on standard techniques of regression 'diagnostics' and iii) the method is conservative, in that controlling for extra-binomial variation means that the CIs will tend to be wider than is the case with conventional methods of aggregating individual effect size estimates. Cumulative meta-analyses are the subject of chapter 25, where this methodology is discussed further.

Sensitivity analysis

Incorporating study quality via weighting in the main analysis has come under criticism. Shadish and Haddock make the case for leaving the incorporation of study quality for a sensitivity analysis:

> '... weighting schemes that are applied at the earlier points seem to be based on three assumptions: (a) Theory or evidence suggests that studies with some characteristics are more accurate or less biased with respect to the desired inference than studies with other characteristics, (b) the nature and direction of that bias can be estimated prior to combining, and (c) appropriate weights to compensate for the bias can be constructed and justified.' (42)

They conclude by stating that a quality weighting scheme does not meet the above conditions so it should be applied with caution. They go on to suggest investigating quality weighting schemes after combining results without such weighting, as a form of sensitivity analysis. The results with, and without this weighting can be compared. To further support this method they comment:

> 'In fact, such explorations are one of the few ways to generate badly needed information about the nature and direction of biases introduced by the many ways in which studies differ from each other.' (42)

Wortman (1) suggests another, more specific, form of sensitivity analysis. He proposes a method to estimate the amount of bias from patients in a randomised trial switching treatment groups. This method makes a simplification and assumes that the sickest patients cross over to the other arm of the trial. By making this assumption an estimate of the amount of bias in an effect size introduced, by a given rate of attrition, is possible (43).

Multivariate analysis

By using the quality score as a covariate in regression models (see chapter 11) to explain heterogeneity (see chapter 8) of study effects, one can take study quality into account. Fredenreich (44) comments this is preferable to weighting studies by their quality because it minimises the influence of quality scoring bias. For an example of the use of this method, see (22).

Bayesian approach

The Bayesian statistical framework to meta-analysis (which is explained in chapter 13) can also incorporate study quality into the analysis. This is done by including prior opinions, relating quality and study bias, provided by one or several 'experts', in the model. Analyses can be performed for each set of ratings to study the dependence of conclusions on individual opinion (to the quality of the studies). If the conclusions are stable over this 'community of opinion', then, the meta-analysis is likely to have substantial impact. If conclusions are sensitive to the range of opinions, then no consensus has been reached (45). Quantifying the degree of sensitivity is itself important. It is important that the assessors should not know study results. However, it is hard to draw the line between 'inputs' and 'results'; for example, study attributes and baseline data are proposed to be available to the assessors. However, follow-up rates are a more difficult issue, as they can indicate a well designed and conducted study, but they may also may indicate an effective treatment (45). In a similar manner there may be *a priori* beliefs regarding the eligibility of evidence from studies within different designs, e.g. randomised and non-randomised. This particular situation is dealt with in chapter 26.

Practical implementation

There are several practical issues to consider when assessing the quality of studies. The first issue is whether to blind the assessors to aspects of the studies. The problems of masking the results and conclusions, necessary for a Bayesian analysis, have already been discussed on pages 28–9. There has been suggestions (e.g. from TC Chalmers) that for assessing the quality of a trial that only the methods and results sections should be presented, with the authors and setting masked, and even the names of the treatment groups deleted to reduce assessor bias (3). Jadad (28) has recently investigated the effects of blinding, and found evidence to suggest that blinded assessment produced significantly lower and more consistent scores than open assessment. This is the first piece of evidence to support what was previously seen as a purely speculative and elaborate precaution (3).

Another problem is that some large and complex trials report the details of study methodology in separate earlier publications. Detsky *et al.* (3) argue that looking at this material would probably increase quality score of the trial above the score it would achieve when considering it in isolation.

As for the way the actual assessing is carried out, the procedure described by Detsky *et al.* seems sensible:

> 'In the past, we have followed a specific protocol, beginning with a training session for quality assessors to review the items in the scale and practice with a sample of studies of variable quality. We have also insisted that the quality assessment be done by a pair of reviewers who then check their results against each other and discuss any discrepancies.' (3)

The researcher should be aware that when a quality evaluation is done, there may be too few studies deemed of good enough quality to pool. This is an acceptable conclusion, indeed, it has already been stated that no information is better than misleading information.

Another point to note is that when the synthesis is being reported, a list of trials analysed and a log of rejected trials should be given (26).

Finally, a few comments on when quality scoring is important (3):

- If all trials are of uniformly high quality considerations will be relatively unimportant.
- In RCTs with hard outcome measures and simple interventions, study quality will have less of an impact on estimated effect sizes.
- 'Assuming "quality counts", it stands to reason that the issue must be formally recognized in meta-analytic techniques whenever there is evidence of variation in the quality of the design and conduct of individual trials.' (26)

Further research

- Evidence of how methodology effects biases.
- If studies are to be excluded guidelines for deciding which ones to exclude are needed.
- Currently one can only weight by quality if the same checklist was used for each study, i.e. one could not do so if different study types are being combined.
- Investigate any relationships between components of quality score and an average amount of bias in study result (or at least its direction).
- The use of the Internet to provide further study details not included in journal reports etc.
- As well as the use of scales and methodology for incorporating the results into a meta-analysis, the issue of whether to include unpublished/non-peer reviewed data and how this should be assessed needs addressing.
- Guidelines for how to proceed when information about studies necessary for scoring is not included in reports.

Moher *et al.* report:

> 'Even if the scales available vary in their size, complexity, and level of development, it would be useful to ascertain whether different scales, when applied to the same trial, provide similar results. This information could guide quality assessors in their choice of scale. There would be little advantage in using a 20-item scale to assess trial quality if similar results could be obtained by using a 6-item scale.'

> 'Future efforts in assessing quality may be best spent in developing scales with appropriate rigor.' 'We also need to address whether, as part of a meta-analysis,

efficacy and safety analyses should be conducted with and without quality scores.' Moher *et al.* also provide a table entitled 'Specific issues to address in the development of a scale to assess trial quality.' (9)

If significant heterogeneity is present (see chapter 8) then quality should be one of the possible factors examined to see whether it explains it. Through further research, a fuller understanding of this relationship may be obtained.

There is little discussion in the literature concerning whether random or fixed effects should be used in conjunction with quality scores (if used to weight), if other heterogeneity is present. Detsky *et al.* (3) mention using a generalised linear model approach which is similar to the random effects model (see chapter 10) that takes extra variation into account.

When excluding studies, Detsky (3) gives several suggestions on how to calculate a cut off value. Little empirical evidence is given to justify this, thus an investigation into the robustness of this value would be desirable.

Summary

This chapter has considered both the assessment and use of quality scores in meta-analysis. Whilst a number of methods have been proposed for assessing study quality (of primary studies) in a meta-analysis, no consensus appears to have developed as to which method is most appropriate, or indeed whether such an exercise is appropriate at all. As far as the use to which such quality scores can be put, a number of possibilities exist, but in specific situations the meta-analysist should not be totally reliant on any one method, in addition that is to an unadjusted analysis.

References

1. Wortman PM, Cooper H, Hedges LV, editors. Judging research quality. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 97–110.
2. Glass GV. Primary, secondary and meta-analysis of research. *Educ Res* 1976;**5**:3–8.
3. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol* 1992;**45**:255–65.
4. Thacker SB. Meta-analysis. A quantitative approach to research integration. *JAMA* 1988;**259**:1685–9.
5. Naylor CD. Two cheers for meta-analysis: problems and opportunities in aggregating results of clinical trials. *Can Med Assoc J* 1988;**138**:891–5.
6. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;**140**:290–6.
7. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science (review). *Epidemiol Rev* 1992;**14**:154–76.
8. Moher D. Assessing the quality of randomized controlled trials: implications for the conduct of meta-analyses. NHS HTA 93/52/04.
9. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials – an annotated- bibliography of scales and checklists. *Controlled Clin Trials* 1995;**12**:62–73.
10. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA* 1994;**272**:101–4.
11. Deeks J, Glanville J, Sheldon T. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. Centre for Reviews and Dissemination, York: York Publishing Services Ltd, Report 4, 1996.
12. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994;**272**:125–8.
13. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations in the use of antithrombotic agents. Antithrombotic Therapy Consensus Conference. *Chest* 1992;**102**:305S–11S.
14. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical-trials. *N Engl J Med* 1983;**309**:1358–61.
15. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I: control of bias and comparison with large co-operative trials. *Stat Med* 1987;**6**:315–25.
16. Khan KS, Daya S, Collins JA, Walter SD. Empirical-evidence of bias in infertility research – over-estimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertil Steril* 1996;**65**:939–45.
17. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clin Trials* 1990;**11**:339–52.
18. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: Surgical. *Stat Med* 1989;**8**:455–66.
19. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med* 1989;**8**:441–54.

20. Khan KS, Daya S, Jadad AR. The importance of quality of primary studies in producing unbiased systematic reviews. *Arch Int Med* 1996;**156**:661–6.

21. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273**:408–12.

22. Sowden AJ, Deeks JJ, Sheldon TA. Volume and outcome in coronary artery bypass graft surgery: True association or artefact? *BMJ* 1995;**311**:151–5.

23. Wortman PM. Evaluation research: a methodological perspective. *Annu Rev Psychol* 1983;**34**:223–60.

24. Cook TD, Campbell DT. Quasi-experimentation: design & analysis issues for field settings. Boston: Houghton Mifflin, 1979.

25. Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. *Controlled Clin Trials* 1981;**2**:31–49.

26. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analysis of randomized controlled trials. *N Engl J Med* 1987;**316**:450–5.

27. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomised controlled trials: current issues and future directions. *Int J Technol Assess Health Care* 1996;**12**:195–208.

28. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay H. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clin Trials* 1996;**17**:1–12.

29. Begg C, Cho M, Eastwood S, Horton R, Moher O, Olkin I. Improving the quality of reporting of randomised controlled trials: the CONSORT statement. *JAMA* 1996;**276**:637–9.

30. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127–39.

31. Levine M, Walters S, Lee H, Haines T, Holbrook A, Moyer V. IV: How to use an article about harm. *JAMA* 1994;**271**:1615–9.

32. Department of Clinical Epidemiology and Biostatistics MM, Health Sciences Centre. How to read clinical journals, IV: to determine etiology or causation. *Can Med Assoc J* 1981;**124**:985–90.

33. Longnecker MP, Berlin JA, Orza MJ, Chalmers TC. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 1988;**260**:652–6.

34. Spitzer WO, Lawrence V, Dales R. Links between passive smoking and disease: a best evidence synthesis: a report of the working group on passive smoking. *Clin Invest Med* 1990;**13**:17–42.

35. Berlin JA, Colditz GA. A meta-analysis of physical activity in the prevention of coronary heart disease. *Am J Epidemiol* 1990;**132**:612–28.

36. Morris RD. Meta-analysis in cancer epidemiology. *Environ Health Perspect* 1994;**102**:61–6.

37. Yusuf S, Peto R, Lewis J, Collins R, Sleight P, *et al.* Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Prog Cardiovasc Dis* 1985;**27**:335–71.

38. Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: commentary. *Am J Epidemiol* 1995;**142**:371–82.

39. Light RJ. Accumulating evidence from independent studies – what we can win and what we can lose. *Stat Med* 1987;**6**:221–31.

40. Abramson JH. Meta-analysis: a review of pros and cons. *Pub Health Rev* 1990;**9**:149–51.

41. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis (review). *J Clin Epidemiol* 1995;**48**:9–18.

42. Shadish WR, Haddock CK. Cooper H, Hedges LV, editors. Combining estimates of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 261–84.

43. Yeaton WH, Wortman PM, Langberg N. Differential attrition: Estimating the effect of crossovers on the evaluation of a medical technology. *Evaluation Review* 1983;**7**:831–40.

44. Friedenreich CM. Methods for pooled analyses of epidemiologic studies (review). *Epidemiology* 1993;**4**:295–302.

45. Louis TA, Zelterman D. Cooper H, Hedges LV, editors. Bayesian approaches to research synthesis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 411–22.

46. Macarthur C, Foran PJ, Bailar JC. Qualitative assessment of studies included in a meta-analysis: DES and the risk of pregnancy loss. *J Clin Epidemiol* 1995;**48**:739–47.

Chapter 7

Simple methods for combining studies

This chapter deals with two methods of combining studies, namely vote-counting methods and combining *p*-values. Both of these methods are very simple; neither of them combines treatment effects size estimates, nor produces a pooled estimate. For more complex methods for data synthesis, see chapter 10 onwards.

Vote-counting methods

Introduction

Vote-counting procedures are one of the simplest forms of pooling study results. Essentially, only the direction of the result from each study is considered, whether that be an elevated risk or a negative correlation. This method ignores the magnitude of effect and, for the majority of the methods, also the significance of the result. For these reasons, vote-counting procedures are only recommended as a last resort, when effect magnitudes and significance levels are not available, or, as a compliment to one of the other methods described later in this part of the report. An instance where this would be sensible is where treatment estimates/significance levels are only available on a proportion of the studies; here, effect sizes could be combined on the possible subset of studies but a vote-counting procedure could be carried out on all studies (1).

The conventional vote-counting procedure

Light and Smith (2) in 1971 were among the first to describe formally, the ‘taking a vote’ procedure. Put simply, each study is tallied to one of three categories, namely showing a positive relationship, negative relationship or no specific relationship in either direction, depending upon the effect size. The category with the highest count is assumed to give the best estimate of the direction of the true relationship.

Clearly, this method could not be simpler, unfortunately it has been criticised for the following reasons:

- The sample size, and therefore the precision of each estimate from the primary studies is not incorporated into the vote (2).
- It does not provide an effect size estimate (3).
- It has very low power for the range of sample sizes and effect sizes most common in the social sciences. When the effect sizes are medium to small, the procedure frequently fails to detect an effect (4).
- The power of this test decreases as the number of studies increases (4).

In conclusion, this method cannot be recommended; it has been described as naive, with no statistical rationale and can lead to erroneous conclusions (4,5). For a thorough explanation of the limitations, see page 48 of (6).

The sign test

Again, this is a simple procedure involving a non-parametric statistical test. The rationale behind the test is that if there is no treatment effect then the chance of a study showing a positive effect is 0.5.

Hence the null hypothesis is:

H_0: probability of a positive result on average in the population (p) = 0.5,

and the alternative,

H_A: $p > 0.5$.

Let U = the number of positive results in k independent studies being considered. Then an estimator of p is $\hat{p} = U/k$.

Tables for the binomial distribution are then consulted to calculate how extreme $\hat{p}$ is, and whether to reject the null hypothesis.

Again, this test also has its disadvantages: it does not incorporate sample size into the vote, and it does not produce an effect estimate.

CIs based on equal sample sizes

Unlike the previous two methods, this method gives an estimate of the treatment effect. However, this vote-counting method is only possible if one assumes that all studies to be combined have the same sample sizes and the numbers in both arms of each study are also the same. Clearly this is very restrictive and unlikely to be the case in most instances. Hedges and Olkin (4) recommend treating studies as if they have the same sample size, if in reality they are not very different. Hedges and Olkin (6) give details on what to set as this fixed sample size, when in reality they differ.

This method produces a CI for the treatment effect of interest. From this, inferences can be made about the effectiveness of the intervention.

This method is quite involved, and thus is not described in this report. A clear account with several examples, using different scales is given elsewhere (7).

Because unequal sample sizes are the rule in research synthesis rather than the exception, the counting estimators are likely to be most useful for providing quick approximate estimates rather than as the analytic tool for the final analysis (6).

CIs based on unequal sample sizes

This method extends the methodology to handle unequal sample sizes for the primary studies. This method was first described by Hedges and Olkin (6). Another description of the method is given by Bushman (1). The method involves maximum likelihood (ML) calculation, and is considerably more complex than that for equal sample sizes. The interested reader is referred to either of these texts for more detail.

It should also be noted that both the above methods rely on a reasonably large sample of studies to obtain accurate estimates (6).

Results all in same direction

If all the results are in the same direction, the method of ML (see above) cannot be used to obtain an estimate of p. Instead, if all the results are in the same direction, we can obtain a Bayes estimate (see chapter 13): see Hedges and Olkin [(6), p. 300], or [(1), p. 211] for details.

Conclusion

As mentioned previously, these methods should only be used for the main analysis as a last resort, when treatment effects are not available for at least a proportion of the studies. However, it is difficult to know how to define a positive result, e.g. it could be one that is significant ($p = 0.05$) or one where it is just positive ($p > 0.5$). Light and Smith (2) discuss an alternative approach based on rejecting inferior studies and state that if this method is taken to the extreme, only one study will be left to vote, i.e. the study deemed to be best will give the final result. Hedges and Olkin (6) state that 0.05 is a good practical choice, as a paper may state that result reached statistical significance even if it does not give any other details. On the other hand, taking a positive result to be $p > 0.5$ allows synthesis in situations where the data are so sparse that only the direction of the result needs to be known.

With increased awareness of their importance, treatment estimates from studies should be obtainable from reports. Even if they are not, ways often exist for obtaining them, for example deriving them from other results in reports (e.g. see pages 148–52 and several of the methods of chapter 20), or by contact with the authors (see pages 14–15). Ideally, therefore, this method should not be used unless absolutely necessary.

Combining *p*-values/ significance levels

Introduction

Methods of combining probability (p-)values from independent significance tests have a long history (8). Several of these methods are closely related to the vote-counting techniques (9) outlined at the beginning of this chapter.

p-*Value definition*

A definition of a p-value can be given as follows:

> 'The probability of finding a test statistic (i.e. a set of sample data) as unusual or extreme as that calculated given that the null hypothesis is true.'
>
> 'Observed p-values derived from continuous test statistics have a uniform distribution under the null hypothesis regardless of the test statistics or distribution from which they arise.' (6)

These facts underlie all the tests in the following sections.

When to use this method

Like vote-counting procedures, these methods do not produce an effect size estimate. Because of this, analyses based on effect magnitude measures are usually preferable (9). Hasselblad (10) suggests two situations when combining p-values might be appropriate: 1) when some studies do not report any effect measures but do report p-values; 2) when the study designs or treatment levels may be so different that combining effect measures would be inappropriate. However, Hasselblad goes on to comment that, like vote-counting techniques: '...

methods of combining p-values are seldom used as meta-analytic tools' (10). They could be used as a supplement to combining treatment effects (see part D) when not all treatment effects are available.

Stating the null and alternative hypotheses for combined significance tests

This section presents the formal hypotheses that are being tested when using the methods presented in this section. This section can be skipped without loss of continuity.

Let T_i represent the effect of interest in the ith study. Then:

$$H_0\text{: } T_i = 0, \text{ for } i = 1,\ldots k.$$

So, for the joint null hypothesis to be true, all the individual null hypotheses must be true.

In words, this means that the treatment effects in all the primary studies have to be zero.

A possible alternative hypothesis is:

$$H_A\text{: } T_i \neq 0$$

Under this alternative, the population parameters are not required to have the same sign. This is very general, and does not inform about the specific structure of variability (9). Alternatively

$$H_A\text{: } T_i \geq 0, \text{ for } i = 1,\ldots k, \text{ with } T_j > 0, \text{ for at least one } j$$

This is used if one knows the effect cannot be negative, such as a correlation, or if one is not interested in negative values such as a variance test where negative values are evidence of zero variance.

Methods for combining *p*-values and significance levels

All the methods described below can be described as non-parametric, as they do not rely on any parametric form of the underlying data only on the p-values (6). This section is essentially a summary of the review of Becker (9), which should be consulted if more details are required.

Minimum p method

This method was proposed by Tippett in 1931 (11). One rejects the null hypothesis if any of the p-values (from the k studies) are less than α. Where α is computed as

$$1 - (1 - \alpha^*)^{1/k}$$

and α* is the present significance level for the combined significance test (e.g. traditionally set α* to 0.05 = 5%). Put formally, one rejects H_0 if:

$$\text{Min}(p_1,\ldots,p_k) = p_{[1]} < \alpha = 1 - (1 - \alpha^*)^{1/k} \qquad (7.1)$$

It can be noted that this method is a special case of Wilkinson's method described in 1951 (12), for $r = 1$ and

$$\alpha = 1 - (1 - \alpha^*)^{1/\text{k}}$$

Also, the Beta distribution with 1 and k degrees of freedom can be used to obtain a level α* test based on the minimum p ($p_{[1]}$).

A generalisation suggested by Wilkinson (12) is to use the rth smallest p-value (13):

H_0 is rejected if:

$$P_{[r]} < \text{C}_{\alpha,k,r}$$

where $\text{C}_{\alpha,k,r}$ is a constant that can be obtained from the Beta distribution [these are tabulated in (6), p. 37].

The advantage of this method is that it does not rely on the most extreme result, and therefore is more resistant to outliers in the data than Tippett's original method.

Sum of z's method

This method was first described by Stouffer *et al.* in 1949 (14). The combined significance test is based on the sum of $z(p_i)$ values (sometimes known as z-scores).

The test statistic is:

$$\sum_{i=1}^{k} z(p_i) \big/ \sqrt{k} \qquad (7.2)$$

This is compared with critical values of a standard Normal distribution (9).

Sum of logs method

This method was first described by Fisher in 1932 (8).The test statistic is expressed:

$$-2\sum_{i=1}^{k} \log(p_i) \qquad (7.3)$$

This is compared with the 100(1 – α*)% critical value of the χ^2 distribution with $2k$ degrees of freedom (df).

Fisher's method can be shown to be derived from a proposed method of ranking vector observations (15), and in particular the choice of $p = 0.37$ is a critical one; p-values below this value suggest the null hypothesis is more likely to be false and p-values above 0.37 suggest it is more likely to be true.

Logit method

This method was proposed by George in 1977 (16), and uses a test statistic defined by:

$$-\sum_{i=1}^{k}\log(p_i/1-p_i)\left[k\pi^2(5k+2)/3(5k+4)\right]^{-1/2} \qquad (7.4)$$

This test statistic is approximately distributed t with $5k + 4$ df.[1] Thus it is compared to the $100(1 - \alpha^*)\%$ critical value of the t distribution with $5k + 4$ df.

Other methods of combining significance levels

Hedges and Olkin (6) briefly discuss several other methods and modifications for combining p-values and significance levels.

Pearson suggested combining p-values via the product: $(1 - p_1)\ldots(1 - p_k)$. David in 1934 considered combining $P_1^*,\ldots,P_k^*$ where $p_i^* = \min[P_i, 1 - P_i]$ [described in (13)].

Edgington proposed combining the sum of p-values: $S = p_1 + \ldots + p_k$. However, this test has very low statistical power (13). This is because a single large p-value may overwhelm many small p-values. This procedure is believed to be poorer than Fisher's method, but very few numerical investigations have been carried out.

Other methods have been proposed by Lipták, who presented a general form of which both Fisher's and normal procedures are both examples. Lancaster presented another statistic based on the sum of the inverse of χ^2 cumulative distribution functions.

In addition, Draper *et al.* (13) reported that Good (1955) and Mosteller and Bush (1954) proposed weighted versions of the inverse χ^2 procedure and inverse normal procedure, respectively. Lancaster (1961) also suggested an alternative weighting procedure for a generalised inverse χ^2 method.

As the reader can gather, there are many different test statistics available for combining p-values and significance levels. The interested reader should note that Becker [(9) p. 222–23] gives a classified table of 16 test statistics that can be used for this purpose. For more information for the methods not referenced in this section, see (6) and (9).

Combining discrete p*-values*

All the above methods have assumed that the statistical tests to be combined have a continuous test statistic, which in turn leads to a p-value that is uniformly distributed under the null hypothesis. If test statistics with discrete distributions (e.g. test statistics based on discrete data) are used, the combination procedures described in this chapter will have to be modified by incorporating 'corrections for continuity' (6). An alternative approach is making p-values into continuous random variables by adding to them an appropriate uniform random variable (6). However, Draper *et al.* (13) observe that this method is very rarely used in practice. This problem is not discussed further here; see (6) for more details.

Combining lower bounds on Bayes factors as an alternative to p*-values*

Chapter 13 of this report discusses Bayesian methods for research synthesis. It should be noted that the whole concept of p-values is at odds with the Bayesian philosophy.[2] Because of the recognised conflict between Classical and Bayesian perspective, Berger and Mortera (17) investigate the interpretation of a p-value, from a Bayesian perspective. This is done by treating the p-value as the data, and computing corresponding posterior probabilities or Bayes factors (BFs). They go on to compare the use of p-values (combined using z-scores) to BFs and posterior probabilities. They conclude:

> 'The lower bounds on Bayes factors are not meant to be a substitute for actual subjective Bayes factors which can be substantially larger. However, if it is not possible to compute actual subjective Bayes factors, the use of lower bounds is arguably superior to the use of p-values.' (17)

Combining significance levels and p*-values when only the level of significance is known*

In some instances, authors may not give the exact p-value, but report the level of significance (e.g. '$p < 0.05$', or the treatment difference was significant

[1] Could also be thought of as distributed normally with zero mean and variance $k\pi^2/3$.

[2] Despite this, Goutis *et al.* (22) look at the different methods for combining p-values and look to place them in a Bayesian context; however, this has been unsuccessful.

at the 5% level). This is indeed a problem. In this situation, it is impossible to construct a test at a particular level; however, if one substituted $p = 0.05$ (or any other known level of significance) for the offending study, a conservative test can be performed (10).

Miscellaneous methods

Becker (9) comments on the possibility of weighting *p*-values. Two different situations are outlined: 1) to account for prior information about the sizes of effects and 2) to allow for subjective differences (e.g. quality), differences in sample size or degrees of freedom).

Other methods for combining *p*-values have been proposed, several of which are reviewed by Mosteller and Bush (18).

Rosenthal's 'file-drawer' test for publication bias, which is covered on pages 126–32, is based on combining *p*-values. Tests for contrasts using *p*-values have also been put forward by Rosenthal (19,20), as a way to identify where variations between studies lie. This involves using the $z(p)$ as an effect measure in its own right. However, Becker (9) gives evidence suggesting that procedures such as this based on the standard normal approximation may tend to overlook real differences in effects when the null hypothesis is false. An extension to this method is given by Strube (21) to combine significance levels of non-independent studies; this is discussed in chapter 27.

Appraisal of the methods

With so many alternative test statistics available, it would be desirable to consider the power of each, to find if any are generally superior in that respect.[3] Unfortunately, no one test is the most powerful in all situations. However, as Elston observes (15), Littel and Folks paper showed Fisher's method to be asymptotically optimal among essentially all methods of combining independent tests. Hence, perhaps the best advice available is given by Hedges and Olkin (6) who state:

> 'It seems that Fisher's test is perhaps the best one to use if there is no indication of particular alternatives.'

It should not be forgotten that all methods of combining *p*-values have disadvantages (as well as advantages) these are summarised in *Box 2.*

BOX 2 Advantages and disadvantages of combining p-values and significance levels

Advantage:

1 Broad applicability: can combine *p*-values even if the studies are dissimilar, i.e. they do not have to have the same design or health endpoint (10).

Disadvantages:

1) Does not provide very detailed information, i.e. no average effect size estimate is produced (9).

2) Does not weight the studies according to the uncertainty of the sample size (10).

3) Acceptance or rejection can depend more on the choice of the statistic than on the data. (13).

4) The information in a highly informative experiment can be masked, and thereby largely disregarded (13).

(13) illustrates the limitations of these methods by using studies investigating the effect of aspirin on patients admitted to hospital having a myocardial infarction. Tippett's and Fisher's methods give the results of 0.157 and 0.0070, respectively, which are clearly very different. Thus method choice seems critical, in at least some instances; however no real guidelines exist for which method to choose.

There is confusion in the literature as to whether by combining studies via *p*-values weights the studies according to their power to detect a treatment effect. Although *p*-values do contain information relating to sample size and variability, the extent to which this is true in any specific situation will depend on a number of factors, including the type of test used.

Summary

This chapter has considered principally two basic methods for synthesising evidence; vote counting and the combination of *p*-values. Whilst vote counting is one of the simplest methods available, it should only be used if absolutely necessary. By contrast, although the combination of *p*-values does convey some aspect of effect size, there are a number of disadvantages to the use of such a method. As a result, it should only be used with caution, since it may mask some fundamental differences in the studies.

References

1. Bushman BJ, Cooper H, Hedges LV, editors. Vote-counting procedures in meta-analysis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 193–214.

[3] Many studies have been performed: see (9), p. 227 for a bibliography and brief summary.

2. Light RJ, Smith PV. Accumulating evidence: procedures for resolving contradictions among different research studies. *Harvard Educational Review* 1971;**41**:429–71.

3. Glass GV, McGraw B, Smith ML. Meta-analysis in social research. Newbury Park, CA: Sage, 1981.

4. Hedges LV, Olkin I. Vote-counting methods in research synthesis. *Psychol Bull* 1980;**88**:359–69.

5. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.

6. Hedges LV, Olkin I. Statistical methods for meta-analysis. London: Academic Press, 1985.

7. White HD, Cooper H, Hedges LV, editors. Scientific communication and literature retrieval. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 41–56.

8. Fisher RA. Statistical methods for research workers. 4th edn. London: Oliver and Boyd, 1932.

9. Becker BJ, Cooper H, Hedges LV, editors. Combining significance levels. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 215–30.

10. Hasselblad V. Meta-analysis of environmental health data. *Sci Total Environ* 1995;**160–161**:545–58.

11. Tippett LHC. The methods of statistics. 1st edn. London: Williams & Norgate, 1931.

12. Wilkinson BA. A statistical consideration in psychological research. *Psychol Bull* 1951;**48**:156–8.

13. National Research Council. Combining information: statistical issues and opportunities for research. Washington DC: National Academy Press, 1992.

14. Stouffer SA, Suchman EA, DeVinney LC, *et al.* The American soldier: adjustment during army life (vol. 1). Princeton, NJ: Princeton University Press, 1949.

15. Elston RC. On Fisher's method of combining *p*-values. *Biomet J* 1991;**33**:339–45.

16. George EO. Combining independent one-sided an two-sided statistical tests – some theory and applications. University of Rochester, 1977.

17. Berger JO, Mortera J. Interpreting the stars in precise hypothesis-testing. *Int Statist Rev* 1991;**59**:337–53.

18. Mosteller F, Bush RR, Lindzey G, editors. Selected quantitative techniques. In: Handbook of social psychology. Cambridge, MA: Addison-Wesley, 1954, p. 289–334.

19. Rosenthal R, Rubin DB. Comparing significance levels of independent studies. *Psychol Bull* 1979;**86**:1165–8.

20. Rosenthal R. Meta-analytic procedures for social research. Revised edition edn. California: Sage, 1991.

21. Strube MJ. Combining and comparing significance levels from nonindependent hypothesis tests. *Psychol Bull* 1985;**97**:334–41.

22. Goutis C, Casella G, Wells MT. Assessing evidence in multiple hypotheses. *J Am Statist Assoc* 1996;**91**:1268–77.

Chapter 8

Heterogeneity

Introduction (defining heterogeneity and homogeneity with respect to meta-analysis)

It is almost guaranteed, when carrying out any meta-analysis, that the point estimates of the effect size from the different studies being considered will differ, to some degree. This is to be expected, and is at least partly due to sampling error which is present in every estimate being combined. That is to say, if several samples are taken from a population, with the same underlying true effect size, the sample estimates will inevitably vary from one another. When effect sizes differ, but only due to sampling error, it is customary to consider the effect estimates as homogeneous. This source of variation can be accommodated in meta-analysis by using a fixed effects model which is discussed in chapter 9.

It is often the case that the variability in effect size estimates exceeds that expected from sampling error alone. If a formal synthesis is to be undertaken this extra variability requires further consideration. When it is present the effect size estimates are considered heterogeneous. Possible reasons for this heterogeneity are discussed later on pages 41–3.

The subject of the heterogeneity of study results is fundamental in meta-analysis and is the source of much debate in the field of systematic reviews. Colditz, in his review of heterogeneity in the meta-analysis of epidemiological studies states:

> '...heterogeneity and approaches to dealing with it take many forms, and such diversity may leave the reader uncertain about the interpretation of the combined results.' (1)

One should be aware that heterogeneity may exist when all or most studies indicate the same direction of treatment effect (i.e. either harmful or beneficial), but the size of this effect differs, as well as when the trials contradict each other about whether there is any treatment benefit.

The most common test for heterogeneity is outlined, followed by an example of its implementation. This is followed by a discussion of its shortcomings. Then a discussion of the various approaches that, in the past, have been taken to deal with any heterogeneity are given. This is followed with a section discussing how heterogeneity affects the results and interpretation of a meta-analysis. The chapter concludes with a section outlining other, lesser used, tests that can be used to check for heterogeneity.

Test for presence of heterogeneity

As Thompson points out, (2) this test, to check the data are homogeneous, is perversely, usually termed a test of heterogeneity. Although several authors have put forward slightly differing formulas for the test they are, mostly, essentially equivalent, being based on χ^2 or F statistics (3). The one devised by Cochran (4), which is widely used, is given below.

General formal test

The formula given below (8.1) can be applied to all types of treatment effect data commonly combined (for details of the different types normally encountered in medical research see chapters 9 and 14). It tests the hypothesis

$$H_0: \theta_1 = \theta_2 = \ldots = \theta_k$$

where the θ_is are the underlying true treatment effect of the corresponding ith studies; versus the alternative that at least one of the effect sizes θ_I differs from the remainder. Essentially, this is testing whether it is reasonable to assume that all the studies to be combined are estimating a single underlying population parameter (this is one of the assumptions underlying the fixed effect model – see chapter 9).

$$Q = \sum_{i=1}^{k} w_i (T_i - \bar{T}.)^2, \qquad (8.1)$$

where k is the number of studies being combined, T_i is the treatment effect estimate in the ith study, and

$$\bar{T}. = \frac{\sum_i w_i T_i}{\sum_i w_i},$$

is the weighted estimator of treatment effect. w_i, is the weight of that study (usually the inverse of the *i*th sampling variance, but not necessarily) in the meta-analysis (chapter 9 covers the calculation of $\bar{T}$. and needs to be referred to before calculation of Q is attempted).

A computationally convenient form of the above formula is:

$$Q = \sum_{i=1}^{k} w_i T_i^2 - \frac{\left(\sum_{i=1}^{k} w_i T_i \right)^2}{\sum_{i=1}^{k} w_i} \qquad (8.2)$$

Q is approximately distributed by a χ^2 distribution on $k - 1$ degrees of freedom. Hence if the value for Q exceeds the upper-tail critical value of χ^2 distribution with $k - 1$ degrees of freedom, the observed variance in study effect sizes is significantly greater than what we would expect by chance if all studies estimated a common population effect size. Thus, one would reject H_0 in favour of H_A and conclude heterogeneity is present (5).

For practical examples using this test see pages 56–66.

Choosing an appropriate critical value for this test is made difficult due to its low statistical power (6), and is discussed at length in the next section.

Additional technical notes

The weights in Q may vary according to the assumptions made about the sampling variances. For instance, when the sampling variances can be assumed to be equal, then w_i, $i = 1,\ldots,k$, is the inverse of a common sampling variance s^2 (7).

Laird and Mosteller [(8), p. 15] comment that an alternative approach to estimating between study variation is available using one-way analysis of variance (ANOVA). ANOVA type procedures are also used by Hedges and Olkin to investigate variability in a number of situations (9); some of these are discussed on pages 50–2.

Problems with detecting heterogeneity – limitations of the Q statistic

Unfortunately, interpreting this test for heterogeneity is often difficult and not as clear cut as it may first appear. Below is a summary of the problems researchers face using this test.

1. The statistical power (i.e. if there are true differences between studies, how likely are these differences to be detected?) of statistical tests for heterogeneity are, in most cases, is very low due to the small number of combined trials (10). This means heterogeneity may indeed be present even if the Q statistic is not statistically significant. Due to this, for the detection of a treatment-by-clinic interaction in a multiclinic trial (i.e. investigating if the underlying treatment effects for each clinic were heterogeneous), Fleiss (11) recommended using a cut-off significance level of 0.10, rather than the usual 0.05. This has become a customary practice in meta-analysis.
2. Shadish and Haddock state: ‘When within-study sample sizes in each study are very large, however, Q may be rejected even when the individual effect size estimates do not really differ much; in such cases, it may be reasonable to pool effect size estimates anyway.’ (5)
3. Matt and Cook state: ‘The likelihood of design flaws in primary studies and of publication biases and the like makes the interpretation of homogeneity tests more complex. If all the studies being meta-analysed share the same flaw, or if the studies with zero and negative effects are less likely to be published, then a consistent bias results across studies can make the effect sizes appear more consistent than they really are.....Conversely, if all the studies have different design flaws, effect sizes could be heterogeneous even though they actually share the same population effect. Obviously, the causes of heterogeneity that are of greatest interest are of a substantiative rather than methodological nature.’ The authors conclude ‘Consequently, it is useful to differentiate between homogeneity tests conducted before and after the assumption has been defended that all study-level differences in methodological irrelevancies have been accounted for.’ (12)

Clearly, due to all the above reasons, one has to be cautious when interpreting the Q statistic, some have gone as far as to suggest it should not be used as a test at all. Shadish and Haddock to consider that it should be used as a diagnostic tool to help researchers know whether they have ‘accounted for all the variance’ (5). This has led to the below suggestion by Colditz *et al.*:

> ‘Because we cannot believe that the among-study variance can ever be zero and because the tests for homogeneity are weak, we should not uncritically accept homogeneity. Perhaps we should not test for homogeneity, but rather quantify estimates for the between study variance as recommended by the National Research Council.’ (1)

However, the researcher perceives the role of the Q statistic, the main point to bear in mind is that, just because the hypothesis that all the studies are estimating the same true underlying effect is not rejected at the 5%, or even the 10% level, does not mean there is not some degree of heterogeneity present.

If there is any doubt, it would seem sensible to err on the side of caution and treat the data as heterogeneous because carrying out an analysis assuming homogeneity on heterogeneous data will produce an estimate with a CI which is too narrow (i.e. too confident a result). If one is able to determine the factors that cause the heterogeneity in the data, it may be possible to adjust the estimates accordingly thus removing the excessive variation making a homogeneous (fixed effects) analysis possible. Some informal tests for heterogeneity are given below that can be used instead or in conjunction with the formal one of pages 39–40.

Graphical informal tests/explorations of heterogeneity

Since the formal *Q* statistic has low power, one of the following exploratory methods should be considered even when this statistic is non-significant to aid decisions on how to proceed with the synthesis (13). They should be used as exploratory techniques and give an indication between which studies heterogeneity is greatest and indicate possible outliers. It should be noted that Greenland (14) is critical of the subjectivity in interpreting graphical plots (‘one can pull trends out of anything if you look at it long enough’), though he encourages their use up to a point.

Plot of normalised (z) scores

If the *z*-scores

$$(T_i - \bar{T}.)/\mathrm{SE}(T_i)$$

are placed in a histogram; under the hypothesis of only random differences among the studies, this histogram should have an approximately normal distribution. Large absolute *z*-scores can signal important departures of individual studies from the average result (13).

Radial plots (Galbraith diagrams)

These are also known as Galbraith diagrams (15). Here, the *z*-statistics are plotted against the reciprocal of the standard errors (SEs). Galbraith reports:

> ‘If this sort of plot is done then points from a homogeneous set of trials will scatter homoscedastically, with unit standard deviation, about a line through the origin.’ (15)

Additionally, one can look at the resulting plot of points for certain characteristics by plotting levels in different colours. Hence, this plot enables studies whose results depart greatly from the line can be observed as possible outliers.

Forrest plot

These plots are commonly used as a way of presenting the results of a meta-analysis (see chapter 22). The estimates of treatment effects, along with their SEs from each study are plotted on the same axis. From this plot an idea of the distribution of the estimates can be gained.

L’Abbé plot

This plot is described by L’Abbe *et al.* (16). The event rates of the treatment groups are plotted against the event rates for the controls for each trial. If the trials are fairly homogeneous the points would lie around a line corresponding to the pooled treatment effect parallel to the line of identity; large deviations would indicate possible heterogeneity (17).

All these graphical methods can aid the researcher in detecting heterogeneity. It is recommended that some kind of investigative plot should be constructed when carrying out a meta-analysis.

Causes of heterogeneity

As well as investigating for the presence of heterogeneity it is also necessary to consider its underlying cause. It may then be possible to adjust the analysis accordingly (see pages 43–50). Bailey in his paper on how study differences affect the interpretation and analysis of the results (18) presented a table showing causes of heterogeneity. The essence of this is reproduced below (*Table 2*). The lower down the table you go, the less desirable the source (the source may influence the analysis and the interpretation of the results). The various approaches for accommodating heterogeneity are outlined on pages 43–8 and the interpretation and validity of the results when heterogeneity is present is discussed on pages 48–50.

So in summary, heterogeneity may be due to chance, or spurious due to the scale used to measure the treatment effect. It may be due to treatment characteristics which can be investigated and/or patient-level covariates which can only be investigated it the researcher has got IPD; or if none of the above account for it, unexplainable.

***TABLE 2** Levels of explanation of heterogeneity [reproduced in modified form from (18)]*

0. Chance

It could be that, in fact, the studies are homogeneous but the *Q*-statistic at whatever level of significance it was tested wrongly rejected the null hypothesis (i.e. a type one error for the *Q*-statistic)

1. Homogeneity achieved by different definition of treatment effect (e.g. absolute difference)

Non-intuitive as it may seem, it is possible to remove heterogeneity by transforming the data to a different scale: 'if by going to a different definition of a treatment effect, one can eliminate the heterogeneity, then one not go any further (in trying to adjust for it)' (N.B. this definition should be reasonably simple and not contrived.)

2. Heterogeneity accounted for by design factor(s)

   (a) Data-derived explanation.
   (b) Explanation not 'influenced' by data.

It may be that the studies differed in their design and conduct (implementation): randomisation, blinding, stopping rules, different eligibility criteria, different definitions of disease, variations in treatment (see pages 42–3). It could also be explained by patient level covariates (these are only available if one is doing an IPD meta-analysis, see chapter 24). If this is the case, these covariates are not nuisance factors, such as study design etc., but they may describe subgroups of patients for whom the treatment is more/less effective (see pages 209–10). It is important to differentiate between data-derived explanations and explanations derived independently of the data. This is because the first of these may have been found through 'fishing expeditions', i.e. different covariates were investigated till one gave statistical significance. This method is plagued with the problems of multiple testing and type one errors.

3. Unexplainable (and real)

Bailey (19) considers this to be the situation in which he is least comfortable about drawing conclusions. It could be that many different factors each contribute a small amount towards the heterogeneity of the results. The combined effect of such factors may be substantial, but due to lack of data or sample size these factors go undetected. This led Boissel to state (10): 'It is because several sources of heterogeneity exist that low *p*-values from heterogeneity tests make interpretation of meta-analysis results difficult.' Another explanation is that the factors which caused the variation may not have been measured or recorded for the studies being combined.

Specific factors that may cause heterogeneity

Bailey made the comment:

> 'Clearly. the interpretation of heterogeneity of outcome depends heavily on how similar the trials were in terms of treatment, patient population, length of follow-up, outcome measurement used, etc. The more similar the trials seem in other respects, the more disturbing any heterogeneity of outcome becomes, and, therefore, the more prominent a role heterogeneity would play in the basic statistical analysis. Conversely, if differences in study design are large, then heterogeneity of outcome is less surprising. The role of heterogeneity becomes one of trying to sort out or understand differences in outcome based on other differences.' (18)

So, simply by considering how the design and conduct of the studies differ, may lead to an explanation of existing heterogeneity. *Box 3,* modified from Naylor (20), states possible ways in which trials in a meta-analysis may differ.

All these can be considered when carrying out a meta-analysis. It should be stressed that if one is considering studies with different designs then because they are subject to different biases, then this may create heterogeneity.

BOX 3 Ways in which apparently similar trials may differ [modified from (20)]

1. Differences in inclusion and exclusion criteria.
2. Other pertinent differences in baseline states of available patients despite identical selection criteria.
3. Variability in control or treatment interventions (e.g. doses, timing, and brand).
4. Broader variability in management (e.g. pharmacological co-interventions, responses to intermediate outcomes including crossovers, different settings for patient care).
5. Differences in outcome measures, such as follow-up times, use of cause-specific mortality, etc.
6. Variation in analysis, especially in handling withdrawals, drop-outs, and crossovers.
7. Variation in quality of design and execution, with bias of imprecision in individual estimates of treatment effect.

In the section below, we discuss how certain factors from the above list may affect heterogeneity. These are some of the most common, and in some cases specific methodology exists for dealing with them.

Impact of early stopping rules on heterogeneity

For ethical reasons, RCTs are sometimes stopped early if it is clear from interim analyses that one of the treatment arms is clearly superior to the other(s). Hughes *et al.* investigated the effect of stopping a trial early would have on heterogeneity in an overview (21), and concluded that:

> 'If the true treatment effect being studied is small, as is often the case, then artificial heterogeneity is introduced, thus increasing the Type I error rate in the test of homogeneity. This could lead to erroneous use of a random effects model, producing exaggerated estimates and confidence intervals. However, if the true mean effect is large, then between-trial heterogeneity may be underestimated.' (21)

They go on to comment:

> 'When undertaking or interpreting overview, one should ascertain whether stopping rules have been used (either formally or informally) and should consider whether their use might account for any heterogeneity found.' The paper advises repeating heterogeneity assessments excluding trials with early stopping rules. 'Then if no evidence is found, then to attribute the heterogeneity to the use of stopping rules may be reasonable though the reduction in power to detect any real variability between trials needs also to be appreciated'. (21)

Impact of underlying risk on heterogeneity

Thompson *et al.* (22), Brand (23), and Davey Smith and Egger (24) have all pointed out that an important issue is to ascertain whether the treatment benefit varies according to the underlying risk of the patients in different RCTs. Several methods have been proposed to investigate this, these are described on pages 46–8.

Impact of size of dose on heterogeneity

It may be the case, that the studies may have used different dose levels of the intervention under investigation. If this is the case, then common sense dictates that treatment effects may vary due to this. Ways of carrying out a dose–response meta-analysis exist so the dose size is taken into account. These are covered on pages 157–61.

Impact of publication bias on heterogeneity

Publication bias is the subject of chapter 16. The Q statistic test for heterogeneity is affected by publication bias. This is explained by Spector and Thompson (25):

> 'The between study variance, estimated from the Chi-Squared statistic for heterogeneity, is itself imprecise and, being strongly dependent on the inclusion or exclusion of small studies, is susceptible to the effects of publication bias.'

Compliance rate

Gelber and Goldhirsch (26) highlight the problem of compliance in the primary studies. They give mathematical justification of how reduced compliance could change the effect estimate and hence increase heterogeneity.

Length of follow-up

Gelber and Goldhirsch point out that the length of follow-up of a trial may have an influence on the treatment effect (26). They highlight the following issues that need consideration when investigating this factor:

- Treatment effects might be present either early, late or consistently through time.
- Trials with the longest follow-up are selective because they were (possibly) designed and conducted earlier.
- A summary measurement based on an overall risk reduction that assumes constant annual risk ratios might differ from actuarial estimates based on yearly assessment.

Thompson also makes the following observation on modelling duration of the trial: 'A longer trial would include information on events both soon after and a long time after randomization, so any true effect of duration would be diluted in such an analysis.' (2)

For example, in wound care, most ulcers heal eventually but the rate varies by treatment. So, too long a follow-up and use of outcome measure such as percentage of wound healed will dilute the treatment effect. To get round this problem one could use a survival type analysis (see chapter 20).

Investigating sources of heterogeneity – introduction

It cannot be stressed how important investigating possible sources of heterogeneity is. Identifying sources of variation can lead to important insights about healthcare effects.

> 'In a meta-analysis, documenting heterogeneity of effect (by identifying sources of variability in results across studies) can be as important as reporting averages. Heterogeneity may point to situations in which an intervention works and those in which it does not. Finding systematic variation in results and identifying factors that may account for such variation, in this way, aids in the interpretation of existing data and the planning and execution of future studies.' (1)

Considering this potential for meta-analysis to explore heterogeneous results led Anello and Fleiss (27) to define two sorts of meta-analysis. They consider when there is little or no heterogeneity, and the aim of the analysis it to improve an estimate of effect or test a hypothesis. This sort of analysis could be described as an 'analytic' meta-analysis. When the goal is to resolve controversy, or pose and answer new questions, the main concern of the meta-analysis is to explain the variation in the effect sizes. The authors call this an 'exploratory' meta-analysis, where the characteristics of the different studies become the focus of the analysis. They further suggest this leads to the idea that protocols for a meta-analysis should reflect its goals and how the results are to be used.

It is the aim of this chapter to outline methods to do this. It may not always be easy (or possible), not least due to lack of data; indeed Thompson *et al.* state:

> 'Although many authors have stressed the clinical and scientific importance of investigating potential sources of heterogeneity when conducting a meta-analysis, such investigation can be unrewarding unless the number of trials is large or individual patient data are available.' (22)

In a similar vein, Dickersin has commented:

> 'it is in situations where one or a few studies seem divergent that the meta-analyst faces his or her most serious and interesting challenges.' (3)

It is worth noting that analysis can still proceed when heterogeneity has not been explained, but efforts should be made first to do so. However, it should also be stressed that the conclusion that the results of the studies are too heterogeneous to combine and interpret meaningfully is a very valid one, and one should not combine for the sake of it (this is discussed further on pages 48–50).

Change scale of outcome variable

It may be sufficient simply to change the scale the study outcomes are measured on, to remove heterogeneity (28). Chapters 9 and 14 introduce the most common scales used, and chapter 15 discusses the relative merits of each. As well as changing the type of scale used, a transformation such as taking logarithms is common practice, though there is sometimes a trade-off between statistical homogeneity and clinical interpretability.

Include covariates in regression model

A regression analysis can be performed to examine whether the heterogeneity between studies can be explained consistently by one or several factors across all studies. Several different factors have been investigated in the past; some of these were discussed on pages 42–3. It may be that some of the studies had, on average, older patients and thus the treatment response differed systematically because of this. Another variable often considered is whether the patients in the trials were of comparable health at the start of the trials, i.e. differences in treatment effect may be due to differences in initial baseline risk. Other examples include differences in length of treatment and differences in treatment application. More controversially, systematic differences may be due to the quality of the trials (this is dealt with in detail in chapter 6). Other factors may be identified that are unique to the topic under investigation.

Full details of how to investigate factors such as these via a meta-regression model are given in chapter 11. For the moment, it is enough to consider which variables are appropriate to include for modelling. Indicator variables for any study characteristic can be constructed, and in addition, scales to calculate overall study quality have been devised (see chapter 6). However, no relationship between study quality and treatment effect have been observed thus far (29). Special techniques are available for investigating dose–response (see pages 157–61) and baseline risk (see pages 46–8) to take into account the continuous nature of these factors. Heterogeneity due to different study types can be investigated via meta-regression as well as newer techniques such as cross-design synthesis (see chapter 26).

If the covariate is a well established correlate, introducing it as routine is justified. If on the other hand it is a non-standard variable we have no more than exploratory data analysis, unless the association is very strong (1). Colditz states:

> 'In addition, when we have few studies, introducing several covariates may use up most of the degrees of freedom. Of course, when several covariates are under consideration, many possible sets of them may have been considered, so problems of being unable to make an honest estimate of residual uncertainty.' (1)

It is worth noting that epidemiological studies vary in their design and conduct, generally more than RCTs, for this reason the scope of the methods is greatest for epidemiological studies (3).

When this type of analysis identifies variables that explain the variation between studies, and one is confident that an 'acceptable' level heterogeneity (above that expected purely by random error) is explained, then one can report the results obtained from the meta-regression. If this is not the case, see page 46 for an alternative model.

Exclude studies

One can test the influence of each study on the heterogeneity results by comparing its contribution to the Q statistic to the χ^2 distribution on 1 df (this is an approximate test) (9). One could exclude study/ies that contribute most variation. This procedure can be justified by reasoning that the first stages of summarising results of any data analysis can involve removing outliers or extreme results. However, one has to be aware that this could introduce bias into the estimates. Colditz *et al.* (1) ran a simulation experiment investigating the effect of removing extreme studies and concluded: '... if the observations had been drawn from standard normal distributions, then removing an extreme quarter of them in samples of the size being used in these studies (derived from a survey of meta-analyses in epidemiology that had removed outliers) or larger would create a bias of about 0.4 of a standard deviation (units of study standard deviation, not the smaller mean)' (1). The authors also noted that by removing largest (or smallest) 25% of data reduced the variance by more than 40%. Colditz *et al.* went on to comment:

> 'It is our impression that scientists generally frown on deleting observations unless there is an assignable cause that has been systematically and fairly appraised for every study, not just the outliers. Thus, we think setting aside studies without cause is generally dangerous for inference and should be discouraged. It can easily lead to overassurance about the precision of the results and suppression of among study variation.' (1)

However, in the case where the data are suspected to be contaminated with errors, Colditz *et al.* (1) conclude it is acceptable to trim data to get a value with substantial meaning.

If one considers removing studies is justified, and by doing so heterogeneity is removed then one can proceed, if desired, with a fixed effect analysis (see chapter 9). The effect of doing so can always be explored in a sensitivity analysis (see pages 209–10).

Analyse groups of studies separately

One may conclude that the studies are too heterogeneous to sensibly combine. When this happens there may be one or several groups of studies that seem similar and thus a decision to combine just these can be made. This could be looked at as a more general case (see above), where all but the most extreme study/ies were combined. This type of analysis is sometimes called subgroup analysis.

Yusuf *et al.* (29) categorise subgroups according to whether they are defined by characteristics measured before randomisation of by those measured after randomisation. Emphasis is placed on the need for subgroup analysis to be defined *a priori*. 'When a subgroup is defined *post hoc*, we have no more than exploratory data analysis and so we recommend that the results be described without testing for statistical significance and that investigators look to other data sets to replicate the finding, since spurious results are less likely to be replicated.' (30)

Gelber and Goldhirsch (26) also discuss subset analyses in meta-analysis and make a distinction between two situations that occur, namely:

1. Analysing all the data and including covariates with the aim of detecting therapeutic effects within subsets of patients (or include study characteristic covariate to investigate how this affects outcome, i.e. explain heterogeneity). (This is really equivalent to the meta-regression methodology discussed on pages 44–5).
2. Separate analyses of subsets of studies. 'Studies being pooled generally differ with respect to treatments applied, control groups, patient eligibility, quality control, study conduct and follow up maturity. Separate comparisons within subsets defined by these features will be misinterpreted unless confounding factors are recognized.'

It should be noted that subgroup analyses are usually secondary analyses (and could be part of a sensitivity analysis – see pages 209–10), that is carried out in addition to the analysis of all the studies. There is a problem of potential over-interpretation of subgroup analyses in medical statistics generally, and thus caution should be applied when interpreting such analyses. Page 209 deals with subgroup analyses in more depth.

Use random effects model

Rather than explain or explicitly adjust for variation between studies, one can pool studies using a random effects model which allows for variation in the underlying effect size between studies to be taken into account. This is often used when the source of variation cannot be identified. Chapter 10 is devoted to this type of meta-analysis model. Colditz *et al.* discuss the use of this model [DerSimonian and Laird popularised its use (7), hence the name] at length (1):

> 'The DerSimonian and Laird statistic for estimating effects has an attractive aspect in its handling of

> homogeneity and heterogeneity that differs substantially from the usual method of testing hypotheses. The usual test asks whether the observed heterogeneity is more extreme than can be accounted for by random fluctuations when allowing some small level of significance, such as 5 percent. If a more extreme result is observed, the investigator declares the set of studies to be heterogeneous. If the observed heterogeneity does not exceed the chosen significance level, the investigator ordinarily decides to act as if the homogeneous case holds even if there is considerable evidence against it. Therefore, the effects are estimated as if all studies had the same mean value, thus leading to the fixed effects model with weights inversely proportional to the variances within the separate studies.
>
> The DerSimonian and Laird statistic instead balances its decision around the **average** value of the observed heterogeneity that would occur when all studies had the same mean (the homogeneous case). If the observed heterogeneity is **less** than **average** for the ideal situation with no true heterogeneity, the investigator uses the same formula that the hypothesis tester would use when the test does not reject homogeneity. On the other hand, if the observed heterogeneity **exceeds** the **average** associated with no heterogeneity, then the investigator uses a different formula for estimating effects that has weights more appropriate to a situation with heterogeneity between the studies, as described below.
>
> The DerSimonian and Laird test that decides which formula to use has roughly a 50 percent significance level, not a 5 percent level. Statistics that change their formulas like this in response to the data are sometimes called **adaptive**. The DerSimonian and Laird formulae respond more smoothly to the actual situation than the testing hypothesis approach. The changed weights themselves are also responsive to the degree of heterogeneity observed, with more heterogeneity leading to more nearly equal weights assigned to the studies. Thus, the procedure adapts continuously as the observed heterogeneity increases.'

It should be stressed, however, that by using a random effects model, no investigation of the causes of heterogeneity is made, so the researcher is none the wiser as to why the study results vary. This conflicts with the view of Greenland (14) that:

> 'I maintain that the primary value of meta-analysis is in the search for predictors of between-study heterogeneity. If use of random effects models makes a difference, the analysis is incomplete. The analyst should carefully search for the source of the discrepancy between the fixed and the random effects interval estimates. The random-effects summary is merely a last resort, to be used only if on cannot identify the predictors or causes of the between-study heterogeneity.'

The whole idea of random effects has been controversial in meta-analysis; see pages 76–8 for a synopsis of various arguments put forward advocating and criticising its use.

Mixed-effect models

If an investigation into the sources of heterogeneity has been carried out and one or more variables appear to account for a proportion of the variation, but evidence that some level of heterogeneity (above the level of random variation) remains, then a random effects term can be included in the model to account for this 'residual' heterogeneity. This model is called a mixed-effect model as it can be viewed as a combination of a meta-regression and a random effects model. This model is the subject of chapter 12. This model seems a sensible compromise and has led to the suggestion that in reality, there will always be unexplained heterogeneity. Thus a random effects term should always be included to account for this.

Use of new models

Other, newer, methods of combining data do exist. Two of these are Bayesian meta-analysis and cross-design synthesis. Each of these has its own way of dealing with heterogeneity. These are discussed in chapters 13 and 26, respectively.

Methods for assessing heterogeneity of underlying risk

The issue that studies may appear heterogeneous because of differences in the baseline risk of the patients was introduced on page 43. If such a relationship exists, its nature could crucially affect decisions about which patients should be treated (22). Such a relationship may even delineate which patients may not benefit from medical interventions, in that the treatment effect may be in the opposite direction for patients at low and high risk (a qualitative interaction) (22).

The usual way of investigating baseline risk is to consider the observed risk of events in the control group (or sometimes the average risk in the control and treatment groups) (23). This variable can be used to adjust the pooled estimate via a regression model (see pages 44–5 and chapter 11). It is necessary to adjust for potential overdispersion in such models, that is residual heterogeneity in the treatment effect not explained by the covariate (baseline risk), otherwise the SE of the estimated slope will be too small.[1]

However, Senn (31) showed that this type of analysis is flawed. The drawback of the regression method is the structural dependence involved (regressing the treatment difference on either the risk in the treatment or control groups or a combination of the two). The origin of the phenomenon lies in the fact that the baseline forms part of the definition of the difference. This can lead to a spurious correlation between extent of treatment effect and the level of response in a placebo group.[2]

Sharp *et al.* (17) discuss this problem of regression to the mean further. They review three conventional approaches relating treatment effect to the proportion of events in the control group, and suggest alternative analyses to get round this problem. These are summarised below.

1. Graph of treatment effect against proportion of events in control group

Note that the problem is not solved by this method. One can plot a graph of the odds ratio of an event (log scale) against proportion of events in the control group (log odds scale) for each trial. Each study can be marked with a circle; the size of the circle relating to the size of that particular study. However, if one calculates a weighted regression line for this plot – one has the problem of regression towards the mean.

Use of this technique:

- is not an appropriate method, and will always be biased
- will be less misleading, that is, less biased, if the trials are mostly large, or the variation in true underlying risks is large.

2. Graph of treatment effect against average proportion of events in the control and treated groups

One can plot the odds ratio of an event (log scale) against the average proportion of events in the control and treated groups (log odds scale). In the example presented in Sharp *et al.* (17) this gave a different conclusion from method 1. However, the authors explain that this approach relies on the assumption that the true treatment effect does not vary between trials; departures from this assumption will lead to bias in the size and direction of any observed association. Again, this method does not solve the problem of regression to the mean.

Use of this technique:

- is appropriate only if the true treatment effect is constant across trials
- will be less misleading if the variation in true underlying risks is large.

3. L'Abbé plot: proportion of events in the treated group against proportion of events in the control group

This plot was proposed as a graphical means of exploring possible heterogeneity (16) (see page 41). If a weighted regression line is fitted to the plot then again due to regression towards the mean this can be misleading.

Use of this technique:

- is a useful exploratory graphical method as an adjunct to a standard meta-analysis plot
- is not appropriate for defining groups in which treatment is or is not effective.

Sharp *et al.* go on to discuss a clinically more useful alternative:

> 'Given that a patient's "underlying risk" is known only to the clinician through certain measured characteristics, a clinically more useful alternative to the problematic analyses we have described is to relate treatment benefit to measurable baseline characteristics. These characteristics, or some combination of them, would act as a surrogate measure of the patient's risk.........An extension of this idea would be to combine several prognostic variables into a risk score........Such a combination would avoid the problem of *post hoc* data dredging which arises when many variables are considered separately and would best be based on data from sources other than the trials which form the meta-analysis for treatment effects, such as prospective studies.' (17)

Other work has been carried out on this subject; McIntosh (32) presented a method to examine population risk as a source of heterogeneity by representing clinical trials in a bivariate two-level hierarchical model, and estimate model parameters by both ML and Bayes procedures (chapter 13).

[1] A method for doing this in the statistical package GLIM is given by Thompson *et al.* (22).

[2] Thompson *et al.* (22) give the mathematical justification for this.

More recently Thompson *et al.* (22) present a solution to the problem using Bayesian methods. This method uses a Bayesian approach implemented using Gibbs sampling (see chapter 13 for further details). This analysis can be extended to include other trial level covariates and patient level ones, when IPD are available. Their method uses the log odds ratio scale (see pages 56–63), and they state that using other scales is possible in principle but currently difficult in practice.

The method of McIntosh (31) assumes bivariate normality of true treatment effects and control group risks across trials, and using a normal approximation for binary outcome data. Thompson *et al.* (22) find these assumptions questionable, especially that the true control groups risks will be normally distributed across trials in a meta-analysis. They state that the robustness of the results to apparently strong assumptions needs investigating. Cook and Walter (33) have presented another method which does not depend on bivariate normality assumptions, and used an unconditional ML approach. Thompson *et al.* (22) compare their Bayesian approach to this and find the results do differ. Further research is needed to ascertain which is the best method.

It should be noted that if individual patient data are available (see chapter 24), it is possible to relate treatment effects to individual patient covariates in an attempt to investigate heterogeneity. As Thompson *et al.* state:

> 'This analysis would not suffer the problems discussed for 'underlying risk', and would moreover be directly useful to the clinician considering treatment for an individual patient.' (22)

This is because underlying risk itself is not a measurable quantity, a clinician only knows about underlying risk through the patient's measurable characteristics.

Thompson *et al.* (22) go on to suggest the development of a prognostic score based on patient covariates and relate treatment effects to this score for individual patients. Such an analysis would remove the need for considering 'underlying risk' directly. They suggest the prognostic score would best be based on data other than that from the trials which form the meta-analysis for treatment effects. Note that the score of risk used should where possible be one which clinicians can use so as to determine which of their patients are likely to benefit sufficiently from an intervention.

The validity of pooling studies with heterogeneous treatment effects

So far, this chapter has outlined ways to detect, and up to a point, deal with heterogeneity in study estimates. It would be wrong, however, to give the impression that heterogeneity between studies can always be dealt with satisfactorily and without controversy. Indeed it is one area in which opposes to meta-analysis lay much criticism. It is very alluring that meta-analysis gives an answer no matter what data are being combined. This issue of whether the results of separate trials are homogeneous enough to be meaningfully combined [termed combinability by Sacks (34)] is real and problematic. It has been argued that producing an overall combined estimate for heterogeneous studies is wrong and leads to a result which is misleading, and impossible to interpret, a much used quote is that it is equivalent to: 'combining apples and oranges and the occasional lemon' (35). However, there are certainly no clear guidelines outlining how variable study results have to be before it is deemed invalid to combine them. Blair *et al.* state:

> 'The decision as to whether estimated differences are large enough to preclude combination or averaging across studies should depend on the scientific context, not just statistical significance. For example, a 25% difference among relative risks may be considered unimportant in a study into a very rare cancer, but important in a study of a more prevalent disease.' (36)

Berlin (37) discusses a meta-analysis with excessive heterogeneity, and concludes that despite no conclusion being able to be drawn from the studies one could provide clinical insight and generate hypotheses. He states:

> 'the decision about whether to calculate a quantitative summary of the data is not always straightforward, and different investigators could legitimately arrive at different decisions.' (37)

Below are personal viewpoints and advice concerning the validity of combining heterogeneous study results.

Fleiss observes:

> 'Some statistical reviewers at the US Food and Drug Administration have strongly criticised the pooling of results from controlled clinical trials in which there is heterogeneity of treatment effect, i.e. sizeable differences exist between studies in their estimates of the effect of treatment, and have suggested that it is valid to combine results only from studies in which the estimates are sufficiently close one to another [sic]'. (38)

However:

> '...not all FDA reviewers are in agreement as to how strict the statistical criteria should be for deciding that different studies are combinable (39).' (37)

This leads Fleiss to question whether the FDA 'reviewers would accept as evidence for efficacy the finding of a statistically significant pooled effect even if the meta-analysis was restricted to studies that were combinable.' (38)

Pater (re-iterating Bob Wittes) takes the following view:

> '...the degree of heterogeneity you are willing to tolerate depends upon the question you're trying to answer. If the question you're trying to answer is the very pragmatic one of how to treat patients, then the degree of heterogeneity you might be willing to tolerate may not be as great as if you are trying to answer some general question about the biologic effect of treatment because we can't give patients 'chemotherapy', we can't give patients 'CMF'. We have to give patients a treatment regimen.' (40)

DeMets (41) questions the meaning attached to the overall results of a meta-analysis when there is heterogeneity across studies. Simon comments:

> 'When the studies differ substantially, one must recognise that the average results may not be representative of the components making up the average.' (42)

Greenland goes one stage further, suggesting:

> 'when there is substantial unexplained variance after covariates have been taken into account, there should be no attempt to pool results and summarise them.' (14)

These comments may give the reader the impression that the existence of heterogeneity is a real drawback for the meta-analysist. However, it has been argued that the fact that meta-analysis can be used to confront heterogeneity and is one of its strengths. Naylor (43) reasons that the generalisability of several small trials, with diverse study populations, may be greater than that of a single trial, especially when the large trial may have involved a carefully selected subset of patients.

Peto comments:

> 'it is precisely when studies differ with respect to the magnitude and perhaps even the direction of the treatment effect that the formal methods of meta-analysis are needed to summarize in an unbiased manner all of the information available to date.' (44)

In a similar fashion, Hedges states when studies conflict, the meta-analysis simply has more to explain (8). Olkin reasons:

> 'If studies that go into a meta-analysis are clones of one another, then we would be able to make statements with a high certainty about a small segment of the population. By the same token, if there is too much diversity, then our degree of certainty is considerably lower, but our conclusions refer to a larger segment of the population.' (45)

Rubin (45) takes a different outlook again on meta-analysis and heterogeneity of study results. He states he is 'concerned not with the summary, but with the forecast one might make for the outcome of a study that may differ from all the studies in hand. Essentially, he would hope to estimate a response surface that gave different results for differently constructed studies, so as, for example, to maximize output for a program.' (46) (see pages 214–15 for more details on this approach).

So, depending on ones aim, it seems that heterogeneity is both the meta-analysts friend and enemy! The thoughts of Bailey may offer some practical help (18) on how to proceed, when the studies are heterogeneous:

In determining the role of inter-study variation it is important to consider three factors:

1. Which question one is trying to answer.
2. The degree of similarity or dissimilarity of design.
3. The degree to which heterogeneity of outcomes can be explained.

Three questions one may be interested in are:

1. Whether the treatment can be effective in some circumstances.
2. Whether treatment is effective on average.
3. Whether treatment was effective on average in trials in hand.

Bailey concluded that under the assumption of no qualitative interaction, the answers to these question coincide. A qualitative interaction between outcome and study can be defined as one where the sign of the outcome changes, i.e. an intervention appears harmful and beneficial in different trials. This is in contrast with a quantitative interaction, where it is only the magnitude (and not the sign) of the effect which changes. Peto considers qualitative interactions 'unusual but not impossible' (13). Fleiss *et al.* (37) believe may be more common than is currently appreciated.

Pocock and Hughes, address the issue of whether fixed or random effects should be used:

> 'A sensible overall conclusion is that neither the fixed effect nor the random effects model can be trusted to give a wholly informative summary of the data when heterogeneity is present. Perhaps the presentation of both approaches reveals the inevitable uncertainty inherent in an overview with heterogeneity.' (47)

Dickersin and Berlin (3) add that if a random and fixed effects analysis come to different conclusions then one can conclude heterogeneity is a problem.

Finally, Boissel *et al.* (10) state three basic causes of a low *p*-value for the heterogeneity test and offer practical advice on how to deal with them:

- random variation (chance)
- inadequacy of the treatment effect model
- interaction between treatment and trials.

In such a situation, it is advisable first to proceed with the association test and the estimate of treatment effect; and second, to consider performing a further analysis.

There are three possibilities: 1) to exclude those trials for which possible sources of inconsistency have been identified on the basis of either medical or methodological grounds (a special case should be made for heterogeneity coming from sets of trials with qualitative interaction); 2) to use a different model of treatment effect; 3) to include the cause of heterogeneity as a covariable in the analysis either at the trial level or at the patient level provided that individual records are available. (In practice, the degree of emphasis accorded to the question of heterogeneity will depend on the objective of the meta-analysis. If the purpose is merely to detect that the treatment has some significant effect, one need not worry unduly about heterogeneity, however low the *p*-value.)

Other tests/investigations of heterogeneity

(This section can be skipped if desired without loss of continuity.)

The below are an outline of other tests/tools that can be used in the investigation of heterogeneity. They are not used as frequently as the methods outlined on pages 39–41 and have been put here for reference purposes. Many other tests exist; Paul and Donner (48) compare nine of these using the odds ratio scale, in a simulation study.

Likelihood ratio test

This method is described by Hedges and Olkin in their book (9). It can be used as an alternative to the *Q* statistic (page 40), but is computationally much more difficult to calculate for no gains, so the authors do not recommend its use, though recently Biggerstaff and Tweedie (49) derived a likelihood ratio test (LRT) for the general comparison of meta-analytic models. Hardy and Thompson (50) show how the maximum likelihood estimates (MLEs) required for such a test can be calculated, either via a relatively straightforward iterative procedure or by direct maximisation of the likelihood in packages such as S-plus (51) and as shown by Senn (52) in Mathcad[3] (see chapter 10).

Odd man out method

The odd man out method (53) is really a completely different approach to meta-analysis that has not been widely adopted. Dickersin and Berlin give a concise explanation:

> 'The areas of overlap of confidence intervals from individual studies are used to construct summary 'confidence regions.' These regions are within the graphic display and include information about both the influence of individual studies and the overall results.' (3)

This method has been used in some meta-analyses to reduce heterogeneity. There are questions to its validity, however, because it excluded trials according to their results, not their design. In such circumstances it may be more helpful to investigate why the results are different, rather than simply exclude studies.

Exact test

Zelen (54) devised an exact test for heterogeneity. The StatXact software of Metha, Patel, and Grey provides an implementation so it is easy to use (55). This is a relatively new development. Emerson (55) recommends its use generally, and it can be particularly useful when studies are small or events are rare. Klein *et al.* (56) have used this test for a meta-analysis.

Tests for heterogeneity when the data is sparse

Tests for heterogeneity have been described that have been modified for sparse data (11,57). The

[3] Produced by Mathsoft.

authors are not aware of their use in meta-analysis other than being noted by Huque and Dubey (58).

Extensions of the Q statistic

Formalising the Q statistic

Hedges and Olkin [(9), p. 153] take a very formal approach to testing. They conclude to likening exploring heterogeneity with the analysis of variance. So

$$Q_T = Q_B + Q_W$$

where Q_T is the total heterogeneity, Q_B the between-classes, and Q_W the within-classes heterogeneity. Splitting the heterogeneity up in this way, testing homogeneity across classes and within classes is possible. Most of the examples given (9) are in education; however, Hedges (59) derives fixed effects estimates from an ANOVA model and similarly give homogeneity tests based on the model. This method breaks heterogeneity down into between and within groups, where the groups are defined by study characteristics.

Using the Q statistic to find outliers

If the contribution each trial makes to the overall Q test statistic is investigated then it may be possible to identify outliers. A formal, but approximate comparison of each q_i^2 to a χ^2 (1 df) distribution can be made, provided the number of trials, *k*, is not too small [(9), p. 256].

Q for vector of correlated estimates

Hedges and Olkin [(9), p. 210] present a method of testing the heterogeneity of a vector of correlated estimates. One may have these when combining multiple outcomes from the primary studies (see chapter 23). See original reference for details.

Test for qualitative interaction

Peto (44) argues quantitative study by treatment interactions (where the treatment effect varies in magnitude across studies, but not in direction) are inevitable, and that it is only important to test for qualitative interaction, where the treatment effect varies in direction across studies. Gail and Simon propose a test (60) for qualitative interaction based on a likelihood ratio statistic (LRS) [outlined by Schmid *et al.* (61), p. 109]. The implications of a qualitative interaction are that it suggests a treatment is beneficial on certain subsets of patients an not others.

Estimating the degree of heterogeneity between event rates using likelihood

Recently Matuzzi and Hills (62) presented a simple way of testing for the presence of heterogeneity and estimating its extent using likelihood. The authors state that this is more powerful than the conventional χ^2 statistic on $N-1$ degrees of freedom. This test, to our knowledge has not been applied to meta-analysis; however, it seems as though this would be possible, clearly more investigation is needed.

Goodness of fit of linear models

If linear or logistic regression models are used, lack of homogeneity of treatment effect can be tested by computing tests of goodness of fit of the model with only main effects for study and treatment, or by testing the interaction of study and treatment (63).

Test for homogeneity of disattenuated effect sizes

Hedges and Olkin report:

> 'If the reliabilities $\rho(T_i, Y_i')$ of the measures used in a series of studies differ, then this differential reliability will attenuate effect sizes to a different degree in each study. Thus even if the disattenuated effect sizes are perfectly homogeneous, the attenuated effect sizes will be heterogeneous.' [(9), p. 136]

A formula is presented to test for heterogeneity, corrected for reliability.

Tests of homogeneity for correlation coefficients

Hedges and Olkin [(9), p. 235] give a test for the homogeneity of *z*-transformed correlations:[4]

$$Q = \sum_{i=1}^{k} (n_i - 3)(z_i - z_+)^2 \qquad (8.3)$$

where z_+ is the weighted average correlation [see (9) for details].

An LRS for correlations is also presented [(9), p. 236]

$$\text{LRS} = -2\left(\sum_{i=1}^{k} n_i \log \frac{1 - r_i^2}{(1 - r_i\hat{\rho}^2)} + N\log(1 - \hat{\rho}^2) \right) \qquad (8.4)$$

[4] Spector and Levine (65) conducted an investigation to determine the Types I and II error rates of the U (equivalent to equation 8.3) statistic test for heterogeneity using the combined estimate for correlation coefficients of Schmidt and Hunter (66) [see (67) also for a similar investigation].

Heterogeneity tests previously used for case–control studies but could be applied in a meta-analysis

Klein *et al.* (57) compare seven tests of homogeneity of the odds ration under various sample size configurations using Monte Carlo methods. The paper assumes the data comes from a single stratified case–control study, however these methods could be applied to meta-analysis by considering study as the stratifying variable.

Heterogeneity within studies

This chapter has considered heterogeneity exclusively at the study level; this is appropriate if one is only concerned with pooling only summary results from each trial. If however, IPD are available for the studies being combined, one is able to also investigate within study variation (as well as between study variation).

Interesting application

Pladevall-Vila *et al.* (64) conducted a meta-analysis investigating the effect of oral contraceptives on rheumatoid arthritis. The *Q* statistic was highly significant and it was clear that heterogeneity was present. To investigate this, the study used many of the techniques described in this (and in other) chapters to assess this heterogeneity. Techniques used include: funnel plot for publication bias, odd man out method, random and fixed effects, subgroup analysis, sensitivity analysis, quality scores, and meta-regression. This paper provides a good illustration of these methods.

Further research

Generally no guidelines on which method/s are superior, and which methods should be used in practice exist. Investigations into the exact and likelihood based tests for heterogeneity should be undertaken, to determine if benefit over more standard methods exists, and if so under what conditions.

- Investigation of the importance of a) the number of studies and b) the size of the individual studies, on the power of the test for heterogeneity.
- Investigating baseline risk, which method(s) is/are superior and should be recommended.
- Investigation into a critical value beyond which one should not consider combining studies.
- Investigation of the relationship between publication bias and heterogeneity.

Summary

In conclusion, we are some way off agreeing upon the best strategy for dealing with heterogeneity. It seems essential to look for it and test for it and sensible to explore possible reasons for its presence. When a sizeable amount of unexplained heterogeneity is still present after this, a judgement has to be made on whether it is appropriate to combine the results; if so with what model; and what conclusions can be drawn from it. Presently these decisions require a large degree of subjectivity on the part of the reviewer. Whatever approach is used, 'it is invalid to delete from the set of studies to be meta-analysed those whose results are in the 'wrong direction' for the opportunity for bias in identifying the 'deviant' studies is too great.' [Fleiss (38)]

References

1. Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: commentary. *Am J Epidemiol* 1995;**142**:371–82.
2. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction (review). *Stat Methods Med Res* 1993;**2**:173–92.
3. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science (review). *Epidemiol Rev* 1992;**14**:154–76.
4. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;**10**:101–29.
5. Shadish WR, Haddock CK, Cooper H, Hedges LV, editors. Combining estimates of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 261–84.
6. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet* 1991;**338**:1127–30.
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177–88.
8. Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care* 1990;**6**:5–30.
9. Hedges LV, Olkin I. Statistical methods for meta-analysis. London: Academic Press, 1985.
10. Boissel JP, Blanchard J, Panak E, Peyrieux JC, Sacks H. Considerations for the meta-analysis of randomized clinical trials: summary of a panel discussion. *Controlled Clin Trials* 1989;**10**:254–81.
11. Fleiss JL. Analysis of data from multiclinic trials. *Controlled Clin Trials* 1986;**7**:267–75.
12. Matt GE, Cook TD, Cooper H, Hedges LV, editors. Threats to the validity of research synthesis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 503–20.

13. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.

14. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;**140**:290–6.

15. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;**7**:889–94.

16. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. *Ann Int Med* 1987;**107**:224–33.

17. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *BMJ* 1996;**313**:735–8.

18. Bailey KR. Inter-study differences – how should they influence the interpretation and analysis of results. *Stat Med* 1987;**6**:351–60.

19. Green SJ, Fleming TR, Emerson S. Effects on overviews of early stopping rules for clinical-trials. *Stat Med* 1987;**6**:361.

20. Naylor CD. Meta-analysis of controlled clinical trials. *J Rheumatol* 1989;**16**:424–6.

21. Hughes MD, Freedman LS, Pocock SJ. The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* 1992;**48**:41–53.

22. Thompson SG, Smith TC, Sharp SJ. Investigation underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* 1997;**16**:2741–58.

23. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med* 1992;**11**:2077–82.

24. Davey Smith G, Egger M. Commentary on the cholesterol papers: statistical problems. *BMJ* 1994;**308**:1025–7.

25. Spector TD, Thompson SG. Research methods in epidemiology. 5. The potential and limitations of meta-analysis. *J Epidemiol Commun Hlth* 1991;**45**:89–92.

26. Gelber RD, Goldhirsch A. The evaluation of subsets in meta-analysis. *Stat Med* 1987;**6**:371–88.

27. Anello C, Fleiss JL. Exploratory or analytic meta-analysis: should we distinguish between them? *J Clin Epidemiol* 1995;**48**:109–16.

28. Fleiss JL. The statistical basis of meta-analysis (review). *Stat Methods Med Res* 1993;**2**:121–45.

29. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clin Trials* 1990;**11**:339–52.

30. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;**266**:93–8.

31. Senn S. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med* 1994;**13**:293–6.

32. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med* 1996;**15**:1713–28.

33. Cook RJ, Walter SD. A logistic model for trend in 2 x 2 x kappa tables with applications to meta-analyses. *Biometrics* 1997;**53**:352–7.

34. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analysis of randomized controlled trials. *N Engl J Med* 1987;**316**:450–5.

35. Furberg CT, Morgan TM. Lessons from overviews of cardiovascular trials. *Stat Med* 1987;**6**:295–303.

36. Blair A, Burg J, Foran J, Gibb H, Greenland S, Morris R, *et al.* Guidelines for application of meta-analysis in environmental epidemiology. ISLI Risk Science Institute. *Regul Toxicol Pharmacol* 1995;**22**:189–97.

37. Berlin JA. Commentary: summary statistics of poor quality studies must be treated cautiously. *BMJ* 1997;**314**:337.

38. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127–39.

39. Huque MF. Experiences with meta-analysis in NDA submissions. *Proc Biopharm Sect Am Statist Assoc* 1987:28–33.

40. Meier P. Proceedings of methodologic issues in overviews of randomized clinical-trials – commentary. *Stat Med* 1987;**6**:329–31.

41. Demets DL. Methods for combining randomised clinical trials: strengths and limitations. *Stat Med* 1987;**6**:341–8.

42. Simon R. Overviews of randomised clinical trials. *Cancer Treat Rev* 1987;**71**:3–5.

43. Naylor CD. Two cheers for meta-analysis: problems and opportunities in aggregating results of clinical trials. *Can Med Assoc J* 1988;**138**:891–5.

44. Peto R. Why do we need systematic overviews of randomised trials? *Stat Med* 1987;**6**:233–40.

45. Olkin I. Meta-analysis: current issues in research synthesis. *Stat Med* 1996;**15**:1253–7.

46. Rubin D, Wachter KW, Straf ML, editors. A new perspective. In: The future of meta-analysis. New York: Russell Sage Foundation, 1992, p. 155–65.

47. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med* 1990;**9**:657–71.

48. Paul SR, Donner A. A comparison of tests of homogeneity of odds ratios in k 2x2 tables. *Stat Med* 1989;8:1455–68.

49. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med* 1997;**16**:753–68.

50. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med* 1996;**15**:619–29.

51. Becker RA, Chambers JM, Wilks AR. The new S language. A programming environment for data analysis and graphics. Pacific Grove, California: Wadsworth & Brooks/Cole, 1988.

52. Senn S. Meta-analysis with Mathcad. *ISCB News* 1996;**20**:4–5.

53. Walker AM, Martin-Moreno JM, Artalejo FR. Odd man out: a graphical approach to meta-analysis. *Am J Public Health* 1988;**78**:961–6.

54. Zelen M. The analysis of several 2 x 2 contingency tables. *Biometrika* 1971;**58**:129–37.

55. Emerson JD. Combining estimates of the odds ratio: the state of the art (review). *Stat Methods Med Res* 1994;**3**:157–78.

56. Klein S, Simes J, Blackburn GL. Total parenteral nutrition and cancer clinical trials. *Cancer* 1986;**58**:1378–86.

57. Jones MP, O'Gorman TW, Lemke JH, Woolson RF. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics* 1989;**45**:171–81.

58. Huque MF, Dubey SD. A meta-analysis methodology for utilizing study-level covariate-information from clinical-trials. *Commun Statist Theory Methods* 1994;**23**:377–94.

59. Hedges LV, Cooper H, Hedges LV, editors. Fixed effects models. The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 285–300.

60. Gail M, Simon R. Testing for qualitative interaction between treatment effects and patient subsets. *Biometrics* 1985;**41**:361–72.

61. Schmid JE, Koch GG, LaVange LM. An overview of statistical issues and methods of meta-analysis. *J Biopharm Stat* 1991;**1**:103–20.

62. Matuzzi M, Hills M. Estimating the degree of heterogeneity between event rates using likelihood. *Am J Epidemiol* 1995;**141**:369–74.

63. Rosenthal R, Cooper H, Hedges LV, editors. Parametric measures of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 231–44.

64. PladevallVila M, Delclos GL, Varas C, Guyer H, Brugues Tarradellas J, Anglada Arisa A. Controversy of oral contraceptives and risk of rheumatoid arthritis: meta-analysis of conflicting studies and review of conflicting meta-analyses with special emphasis on analysis of heterogeneity. *Am J Epidemiol* 1996;**144**:1–14.

65. Spector PE, Levine EL. Meta-analysis for integrating study outcomes: a Monte Carlo study of its susceptibility to Type I and Type II errors. *J Appl Psychol* 1987;**72**:3–9.

66. Hunter JE, Schmit FL. Methods of meta-analysis: correcting error and bias in research findings. Newbury Park, California: SAGE Publications, 1990.

67. Sackett PR, Harris MM, Orr JM. On seeking moderator variables in the meta-analysis of correlational data: a Monte Carlo investigation of statistical power and resistance to type I error. *J Appl Psychol* 1986;**71**:302–10.

Part D:

Results III – statistical analyses (classical and Bayesian)

Chapter 9

Fixed effects methods for combining data

Introduction

Using a fixed effect model to combine treatment estimates assumes no heterogeneity between the study results to be combined; that is to say the studies are all estimating one single true value underlying all the study results. Hence, all observed variation in the treatment effects between the studies is considered due to sampling error alone. Clearly, in many instances this may not seem realistic; by simply eyeballing the data differences observed may appear larger than those expected solely by sampling error. However, the decision will not always be so clear cut, and for this reason the formal test of heterogeneity given on pages 39–41 (or any of its equivalents) can (and should) be used as a guide to when the use of a fixed effect model is appropriate. So, the methods presented in this chapter could be considered for use in a special case, i.e. when no heterogeneity is present.

The general approach, which can be adapted to most data types, is presented, followed by illustrative examples using two common scales of measurement used in evidence based medicine, namely, odds ratios (ORs) and standardised effect sizes (continuous outcomes). Methods specific to ORs have also been developed, namely, the Mantel–Haenszel, Peto and ML methods, these are also covered. Again, examples are given for each method. Chapter 14 deals with the other dichotomous (binary) and continuous scales of measurement used in medical research, together with a section on ordinal data. Chapter 15 discusses issues concerning these different scales.

All the methods presented in this chapter, with the exception of the MLE method, are conceptually simple, and can be calculated without the use of computer software. That is not to say that analysis cannot be facilitated by the use of a computer. MLE methods, however, require computer intensive methods for their implementation.

General fixed effect model – the inverse variance-weighted method

Fixed effect estimates can generally be calculated for all data types using the same general formula presented here. The inverse variance-weighted method was first described by Birge (1) and Cochran (2) in the 1930s and is conceptually simple. Each study estimate is given a weight directly proportional to its precision (i.e. inversely proportional to its variance).

For $i = 1,\ldots, k$ independent studies to be combined, let T_i be the observed effect size, θ_i the population effect size with variance v_i, for the ith study. We assume all population effect sizes are equal i.e. $\theta_1 =\ldots= \theta_k = \theta$ for a fixed effect model; that is to say, the studies are all estimating one single true value underlying all the study results. A general formula for the weighted average effect size for these studies is thus:

$$\bar{T}. = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \qquad (9.1)$$

The weights that minimise the variance of $\bar{T}.$ are inversely proportional to the conditional variance in each study (3), i.e.

$$w_i = \frac{1}{v_i} \qquad (9.2)$$

The explicit variance formulae depend on the effect measure being combined. The sections on the OR (pages 56–63), continuous outcome (pages 63–6), and combining other effect sizes (chapter 14) give the necessary formula for v_i. For an exhaustive list of these see (4,5).

An approximate [exact if the effect size is normally distributed (6)] $100(1 - \alpha)\%$ CI for the population effect size is given by:

$$\bar{T}. - z_{\alpha/2}\sqrt{\left(1\Big/\sum_{i=1}^{k} w_i\right)} \le \theta \le \bar{T}. + z_{\alpha/2}\sqrt{\left(1\Big/\sum_{i=1}^{k} w_i\right)} \qquad (9.3)$$

where $z_{\alpha/2}$ is the appropriate critical value of the normal distribution. This is essentially all that is required to synthesise study treatment effects at the most basic level. For a more thorough coverage of the inverse variance-weighted method see (3).

Technical note

Li *et al.* (7) show that the variance estimation formula for the standard inverse variance-weighted method can sometimes be biased and too sensitive to the minimum of the estimates of the variances in the K studies. If the minimum happens to be wrongly reported to have a very small value, its influence would be great, leading to a badly underestimated value of the true pooled variance. This paper gives mathematical justification for this and goes on to suggest an adjusted variance formula.

Combining binary outcomes from studies using the OR

If the outcome from a study, such as an RCT, is binary (e.g. failure/success or death/survival etc.) the results can be presented in the form of *Table 3* below. The OR can then be calculated by the formula

$$\frac{a/(a+c)}{b/(b+d)} \tag{9.4}$$

but the slightly simpler approximation

$$\frac{ad}{bc} \tag{9.5}$$

is often used (and will be through the course of this report).

TABLE 3

	Failure	Success
New treatment	*a*	*b*
Control	*c*	*d*

This measure gives a relative measure of risk in the form of the ratio of two odds (8). An OR of < 1 when comparing a new treatment to the control would indicate an improvement on the new treatment; while a ratio greater than one would imply the new treatment was less effective than the control.

For the purposes of combining results, it is common, and recommended, to first transform the data and work with log ORs instead. The main reason for this being, only the finite interval from 0 to 1 is available for indexing a lower risk in the treatment population, but an infinite interval from 1 up is, theoretically, available for indexing a higher risk in the treatment population. Transforming the scale in this way removes this constraint. A further advantage (but of lesser importance) of doing this is that the log(OR) takes on the value zero when no relationship exists, rather than one, which is intuitively more appealing (5). The estimate and corresponding CI obtained can then be converted back onto an OR scale by taking anti-logarithms. The large sample variance of the natural log of the OR is:

$$v_{Ln(OR)} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{9.6}$$

Thus formula (9.6) can be used to calculate weights for the inverse variance-weighted method. An important problem that needs addressing is that (9.6) is undefined if there are no events in either of the treatment arms (i.e. one or more of $a, b, c, d = 0$). When this occurs the inverse variance-weighted method cannot be used. One way to get round this is to take the advice of Gart and Zweifel (9), who suggest it good practice to add 0.5 to each cell frequency before proceeding with the analysis. They suggest, this reduces bias caused by one or more small cells; it can be seen as a continuity correction factor for converting discrete data to a continuous scale.[1]

Example 1: combining ORs

Effect on mortality of lowering ones' serum cholesterol level

This dataset was introduced in chapter 5. Of the 34 RCTs described (10), only the seven using patients largely without pre-existing cardiovascular disease, i.e. intervention used as primary prevention, will be considered. This is primarily to reduce the amount of computation required for the purposes of illustration. For clarity data from these seven trials are reproduced in *Table 4*.

The point estimate of the OR(OR) for each study can be calculated as before (e.g. study identification (ID) 16: OR = (174 × 244)/(250 × 178) = 0.954). The 95% CIs for these estimates are calculated on a log scale using the formula:

$$\begin{aligned} \ln(\overline{\text{OR}}) = -1.96\big(\text{SE}(\ln(\overline{\text{OR}}))\big) &\le \ln(\text{OR}) \\ &\le \ln(\overline{\text{OR}}) + 1.96\big(\text{SE}(\ln(\overline{\text{OR}}))\big) \end{aligned} \tag{9.7}$$

[1] There is some discussion in the literature as to the exact nature that a continuity correction should take. This aspect has been considered specifically in (25).

***TABLE 4** Results of the seven primary studies investigating the effect cholesterol lowering on mortality to be combined*

Study ID	Number of subjects in the treatment arm (nt)	Number of subjects in the control arm (nc)	Number of deaths in treatment arm (dt) = *a*	Number of deaths in the control arm (dc) = *c*	Number still alive in treatment arm (nt–dt) = *b*	Number still alive in the control arm (nc–dc) = *d*
16	424	422	174	178	250	244
20	1149	1129	37	48	1112	1081
24	4541	4516	269	248	4272	4268
28	1906	1900	68	71	1838	1829
29	2051	2030	44	43	2007	1987
30	6582	1663	33	3	6549	1660
31	5331	5296	236	181	5095	5115

***TABLE 5** Point estimate and approximate 95% CI from the seven primary studies*

Study ID number	Estimate of OR $(a \times d)/(b \times c) = \overline{\text{OR}}$ [95% CI]
16	0.95 [0.73,1.25]
20	0.75 [0.48,1.16]
24	1.08 [0.91,1.29]
28	0.95 [0.68,1.34]
29	1.01 [0.66,1.55]
30	2.79 [0.85,9.10]
31	1.31 [1.07,1.59]

***TABLE 6** Relative weightings of the seven studies to be combined*

Study ID	var(ln(OR))	SE(ln(OR))	$w = 1/\text{SE}^2$
16	0.0195	0.1400	51.38
20	0.0497	0.2229	20.13
24	0.0082	0.0907	121.68
28	0.0300	0.1729	33.47
29	0.0465	0.2156	21.28
30	0.3644	0.6037	2.74
31	0.0102	0.1010	98.48

where $\text{SE}(\ln(\overline{\text{OR}}))$ is calculated by taking the square-root of (9.6).

(e.g. study ID 16: = $v_{Ln(OR)} = 1/250 + 1/174 + 1/244 + 1/178 = 0.0195$. $\ln(\overline{\text{OR}}) = \ln(0.95) = -0.051$ giving a 95% CI of $-0.051 \pm 1.96\sqrt{0.0195} = [-0.31,0.22]$

95% CI for $(\overline{\text{OR}}) = [e^{(-0.31)}, e^{(0.22)}] = [0.73,1.25]$)

Table 5 displays the point estimates along with their 95% CI for the other studies.

It can be seen from *Table 5* that three of the point estimates are < 1, indicating a reduced risk for people on the treatment (cholesterol reducing) and four are > 1 indicating an increased risk for the treatment. However, with the exception of study 31, every CI includes one. From this, one would conclude no significant treatment effect was detected (when considering each study separately). Study 31's 95% CI spans 1.07–1.59, indicating evidence of an statistically significant increased risk for the patients in the treatment arm. By combining these studies it is hoped an estimate, which is more generalisable than any of the individual study results (because studies using different populations are being combined) and more precise (due to increased numbers) can be produced.

As noted previously, when combining ORs, it is desirable to work on the log odds scale. The weighting of each study (w) in the combined estimate needs calculating. Using the inverse variance-weighted method, this is simply equal to $1/\text{v}_{Ln(OR)}$ (e.g. for study 1, $W = 1/0.0195 = 5.13$). *Table 6* shows the values for the rest of the studies.

Before combining the results of studies (using the inverse variance-weighted method), it is necessary to check that the results are in fact homogeneous.

Using the test for heterogeneity given on pages 39–41:

$$Q = \sum_{i=1}^{k} w_i \ln\left(\overline{\text{OR}}\right)_i^2 - \frac{\left(\sum_{i=1}^{k} \ln\left(\overline{\text{OR}}\right)_i\right)^2}{\sum_{i=1}^{k} w_i}$$

$$= \left[(51.38 \times (-0.05^2)) + \ldots + (98.48 \times 0.27^2)\right] - \frac{\left[(51.38 \times (-0.05)) + \ldots + (98.48 \times 0.27)\right]^2}{(51.38 + \ldots + 98.48)}$$

$$= 10.1854$$

This is compared to a χ^2 statistic on six ($n - 1$ studies) degrees of freedom. This value gives a corresponding p-value of 0.117. As stated on pages 39–41, this test has low power and thus a significance level of $p < 0.1$, is usually taken as a critical value. In this case $p > 0.1$, and is therefore non-significant, though only marginally so. It is clearly possible that study results may vary by a greater amount than chance alone would permit. In chapter 10, this same analysis is repeated using methods which take the between study variation into account. This section also discusses under what conditions each type of analysis is appropriate. For now, it is sufficient to consider no heterogeneity being present and proceed with a fixed effects analysis (for illustrative purposes.)

Combining the results using formula (9.1), gives a point estimate of the log(OR)

$$\text{Ln}(\overline{T}_{OR}) = \frac{\left[\left(51.38 \times (-0.05)\right) + \ldots + (98.489 \times 0.27)\right]}{(51.38 + \ldots + 98.48)}$$

$$= 0.085$$

with an estimated SE (9.2)

$$\text{SE}\left(\ln(\overline{T}_{OR})\right) = \sqrt{1/(51.38 + \ldots + 98.48)}$$

$$= 0.054$$

Converting back to an OR scale

$$\overline{T}_{OR} = \exp(0.088)$$

$$= 1.09$$

Calculating the 95% CI for this combined estimate, using formula (9.7)

$$\text{lower limit} = \exp(0.088 - (1.96 \times 0.054))$$

$$= 0.98$$

$$\text{upper limit} = \exp(0.088 + (1.96 \times 0.054))$$

$$= 1.21$$

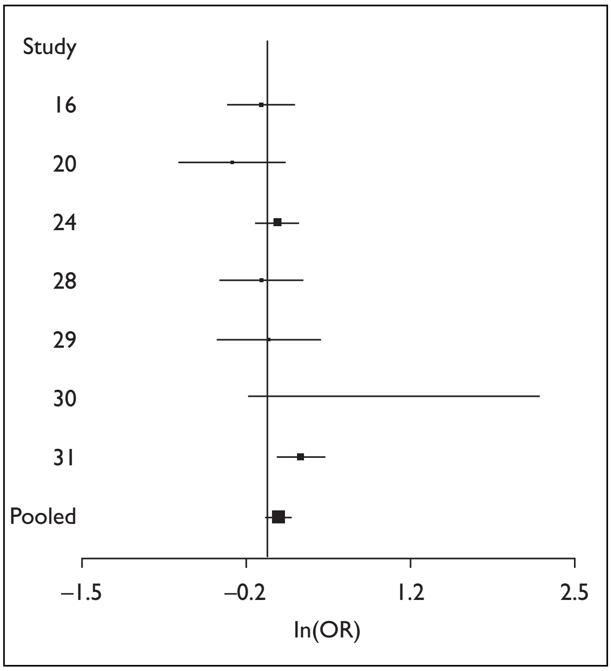


***FIGURE 1** Plot of seven cholesterol trials together with pooled result using a fixed effects model*

The results of this analysis are displayed graphically in *Figure 1*. This is a very common way of displaying the results of a meta-analysis. Each studies point estimate together with its 95% CI is displayed. The size of the box representing each point estimate is proportional to the size and hence the weight of that study in the analysis. The estimate at the bottom of the diagram is centred on the combined point estimate together with the lower and upper bounds of its 95% CI.

The combined OR is slightly greater than 1, however, because its corresponding CI includes 1, a conclusion that no evidence of a treatment effect exists is drawn from the combined results of the seven studies. Although the point estimate is small, its CI is tight, due to large numbers and only just crosses unity. The possibility that cholesterol lowering treatment may actually be harmful as a primary intervention cannot be completely ruled out.

Other methods for combining ORs

Other methods, specific to combing ORs, are available. Under most conditions the estimates obtained from each method should be very similar to one another. However, when the data is sparse, results may differ and some traditional methods may break down altogether. For this reason, new computer intensive methods have been developed to be used in situations where the traditional methods have questionable validity (11). Both the standard methods and the newer ones are outlined. For a more extensive coverage see (11).

Mantel–Haenszel method for combining ORs

This method was first described by Mantel and Haenszel (12) for the use in combining ORs for stratified case–control studies. Later Mantel (13) reported that the method could be used for a wider class of problems, including prospective studies (6). For use in meta-analysis, the study number functions as the stratification variable (14). The formula is given below:

$$\bar{T}_{MH(OR)} = \frac{\sum_{i=1}^{k} a_i d_i \big/ n_i}{\sum_{i=1}^{k} b_i c_i \big/ n_i} \qquad (9.8)$$

where a_i, b_i, c_i and d_i are the four cells of the 2×2 table for the ith study as given on pages 56–63 and n_i is the total number of people in the ith trial.

A variance estimate for the estimated summary OR, $\bar{T}_{MH(OR)}$, is required in order to calculate a CI around this point estimate. The formula commonly used[2,3] was derived by Robins, Breslow and Greenland (15) and Robins, Greenland and Breslow (16). This formula computes a variance estimate for the log of $\bar{T}_{MH(OR)}$, and is notated:

$$v_{\mathrm{MH(ln(OR))}} = \frac{\sum_{i=1}^{k} P_i R_i}{2\left(\sum_{i=1}^{k} R_i\right)^2} + \frac{\sum_{i=1}^{k} (P_i S_i + Q_i R_i)}{2\left(\sum_{i=1}^{k} R_i\right)\left(\sum_{i=1}^{k} S_i\right)} + \frac{\sum_{i=1}^{k} Q_i S_i}{2\left(\sum_{i=1}^{k} S_i\right)^2} \qquad (9.9)$$

where $P_i = (a_i + d_i)/n_i$, $Q_i = (b_i + c_i)/n_i$, $R_i = a_i d_i / n_i$, and $S_i = b_i c_i / n_i$.

A $100(1 - \alpha)\%$ CI is thus given by:

$$\exp\left[\ln(\bar{T}_{MH(OR)}) - z_{\alpha/2}(v_{MH(OR)})^{1/2}\right] \leq \theta \leq \exp\left[\ln(\bar{T}_{MH(OR)}) + z_{\alpha/2}(v_{MH(OR)})^{1/2}\right] \qquad (9.10)$$

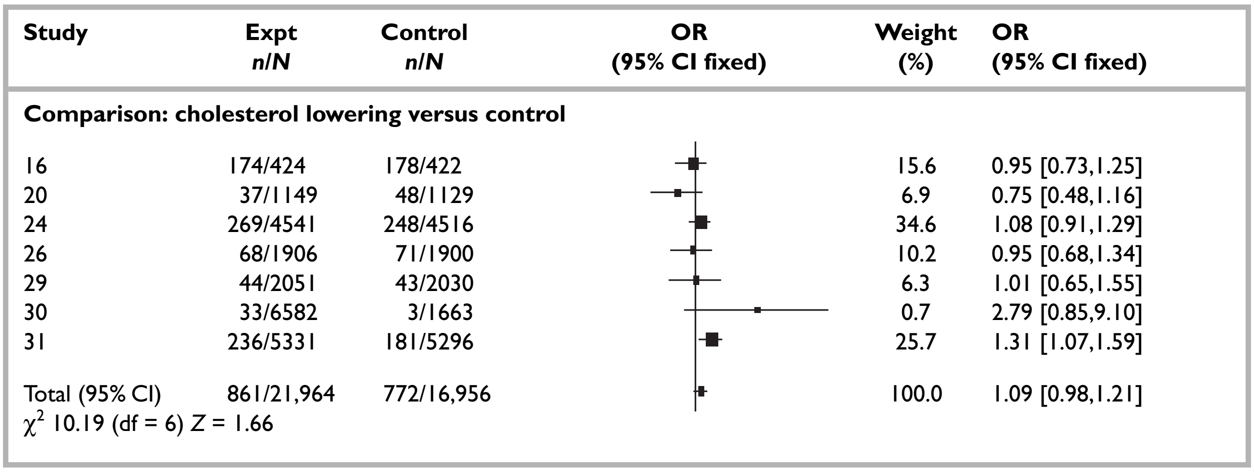
Applying the Mantel–Haenszel method to the cholesterol lowering data (Figure 2)

Taking the seven primary studies used above and combining them using the Mantel–Haenszel estimate (9.8) and calculating a 95% CI using the Robins, Breslow and Greenland formula, (9.9), presented above:

$$\bar{T}_{MH(OR)} = \frac{\sum_{i=1}^{7} a_i d_i \Big/ n_i}{\sum_{i=1}^{7} b_i c_i \Big/ n_i} = \frac{\left[\frac{(174 \times 244)}{846} + \ldots + \frac{(236 \times 5115)}{10{,}627}\right]}{\left[\frac{(250 \times 178)}{846} + \ldots + \frac{(5095 \times 181)}{10{,}627}\right]}$$

$$= 1.09$$

FIGURE 2 *Graphical plot of the combined studies using the Mantel–Haenszel estimate (obtained using Meta View, part of the Cochrane systematic review software package)*



Study	Expt *n*/*N*	Control *n*/*N*	OR (95% CI fixed)	Weight (%)	OR (95% CI fixed)
Comparison: cholesterol lowering versus control					
16	174/424	178/422		15.6	0.95 [0.73,1.25]
20	37/1149	48/1129		6.9	0.75 [0.48,1.16]
24	269/4541	248/4516		34.6	1.08 [0.91,1.29]
26	68/1906	71/1900		10.2	0.95 [0.68,1.34]
29	44/2051	43/2030		6.3	1.01 [0.65,1.55]
30	33/6582	3/1663		0.7	2.79 [0.85,9.10]
31	236/5331	181/5296		25.7	1.31 [1.07,1.59]
Total (95% CI)	861/21,964	772/16,956		100.0	1.09 [0.98,1.21]
χ^2 10.19 (df = 6) *Z* = 1.66					

[2] Several others have been put forward, these are further explored in (15,16,35). Emerson (11) also discusses the variance estimator at length. Sato (36) developed a method that works directly on the odds ration scale (opposed to ln(OR)). Simulations have shown that this works as well as the method of Robins given above and may have slight advantage for matched pair data arising in epidemiology studies. Pigeot (37) has developed another approach using the jack-knife.

[3] Fleiss (23) comments that above variance estimator 'is remarkable in that it is valid both when the study's design calls for matching and when it calls for stratification'.

Calculating the variance of the natural log of the above estimate from formula (9.9)

$$v_{MH(\ln(\mathrm{OR}))} = \frac{\sum\limits_{i=1}^{7} P_iR_i}{2\left(\sum\limits_{i=1}^{7} R_i\right)^2} + \frac{\sum\limits_{i=1}^{7}(P_iS_i + Q_iR_i)}{2\left(\sum\limits_{i=1}^{7} R_i\right)\left(\sum\limits_{i=1}^{7} S_i\right)} + \frac{\sum\limits_{i=1}^{7} Q_iS_i}{2\left(\sum\limits_{i=1}^{7} S_i\right)^2}$$

$$= \frac{[(0.51 \times 52.60)+\ldots+(0.50 \times 86.78)]}{2(52.60+\ldots+86.78)^2}$$

$$+ \frac{[(0.51 \times 50.18 + 0.49 \times 52.60)+\ldots+(0.50 \times 113.59 + 0.50 \times 86.78)]}{2(52.60+\ldots+86.78)(50.18+\ldots+113.59)}$$

$$+ \frac{[(0.49 \times 50.18)+\ldots+(0.50 \times 113.59)]}{2(50.18+\ldots+113.59)^2}$$

$$= \frac{169.72}{2(337.60)^2} + \frac{354.30}{2(337.60)(368.84)} + \frac{182.42}{2(368.84)^2}$$

$$= 0.0028$$

Using formula (9.10) a 95% CI for $\overline{T}_{MH(OR)}$ is given by:

$$\exp[0.088 - 1.96(0.0028)^{1/2}] \leq \theta \leq \exp[0.088 - 1.96(0.0028)^{1/2}$$

$$= [0.984,1.213]$$

If this estimate and CI is compared to that obtained using the inverse variance-weighted method, it can be seen that in this case both methods give nearly identical answers and the conclusions drawn here are the same as those on pages 56–8.

Peto method for combining ORs

This method was first described by Peto in 1977 (17) and more thoroughly by Yusuf *et al.* (18). It can be regarded as a modification of the Mantel–Haenszel method presented above. An advantage it has over the Mantel–Haenszel method that it can still be used when some of the cells in the table are zero; and is easy to calculate. Unfortunately, this method is capable of producing serious under estimates (5), when the OR is far from unity. This is most unlikely to be a problem in clinical trials, but could be in the meta-analysis of epidemiological studies (19) (see chapter 19).

Defining n_i as the number of patients in the ith trial and n_{ti} as the number in the treatment group of the ith trial. Let d_i equal the total number of events from both treatment and control groups, O_i the number of events in the treatment group, E_i the 'expected' number of events in the treatment group (in the ith trial), calculated: $E_i = (n_{ti}/n_i)\,d_i$. For each study two statistics are calculated: 1) O–E, the difference between the observed and the number expected to have done so under the hypothesis that the treatment is no different from the control, E. 2) v, the variance of the difference O–E.

For K studies the pooled estimate of the OR is given by (20):

$$\overline{T}_{PETO(OR)} = \exp\left[\sum_{i=1}^{K}(O_i - E_i) / \sum_{i=1}^{K} v_i\right] \qquad (9.11)$$

where $v_i = E_i[(n_i - n_{ti})/n_i][(n_i - d_i)/(n_i - 1)]$.

An estimate of the approximate variance of the natural log of the estimated pooled OR is provided by:

$$\mathrm{var}(\ln \overline{T}_{PETO(OR)}) = \left(\sum_{i=1}^{K} v_i\right) \qquad (9.12)$$

A 100(1 – α)% CI is thus given by:[4]

$$\exp\left[\frac{\sum\limits_{i=1}^{k}(O_i - E_i) \pm z_{\alpha/2}\sqrt{\sum\limits_{i=1}^{k} v_i}}{\sum\limits_{i=1}^{k} v_i}\right] \qquad (9.13)$$

Applying Peto's method to the cholesterol lowering data

For this method it is necessary to calculate the marginal values for each 2 × 2 table to be combined. *Table 7* illustrates this for the first study (ID 16).

From *Table 7,* the values needed to calculate Peto's method can be calculated.

TABLE 7 *2 x 2 table including marginal values for study ID 16*

	Dead	Alive	Total
Lowering cholesterol treatment	174	250	424
Control	178	244	422
Total	352	494	846

[4] This is not symmetric (34).

For study ID 16:

$$O_i = 174$$

$$E_i = \left(\frac{424}{846}\right)352 = 176.42$$

$$v_i = 176.42\left[\frac{(846 - 424)}{846}\right]\left[\frac{(846 - 352)}{(846 - 1)}\right] = 51.45$$

Table 8 presents these values for the other six studies.

***TABLE 8** Intermediate values needed to calculate the Peto estimate*

Study ID	O_i	E_i	v_i	$O_i - E_i$
16	174	176.42	51.45	–2.42
20	37	42.87	20.46	–5.87
24	269	259.21	121.88	9.79
28	68	69.61	33.49	–1.61
29	44	43.72	21.29	0.28
30	33	28.74	5.77	4.26
31	236	209.19	100.17	26.81

Note: weighting equal to var($\ln\overline{T}_{PETO(OR)}$)

Entering the values from *Table 8* into equation (9.11) gives the combined estimate

$$\overline{T}_{PETO(OR)} = \exp\left[\frac{(174 - 176.42)+\ldots+(236 - 209.19)}{(51.45)+\ldots+(100.17)}\right] = 1.09$$

As equation (9.12) shows, the variance of this estimate is given by the sum of the v_is

$$\mathrm{var}(\ln\overline{T}_{PETO(OR)}) = (51.45+\ldots+100.17) = 354.52$$

Hence a 95% CI is given by (9.13)

$$\exp\left(\frac{31.24 \pm 1.96\sqrt{354.52}}{354.52}\right)$$

$$95\%\ \mathrm{CI}\ [0.98,1.21]$$

Figure 3 shows a plot of studies combined using the Peto method.

This result is exactly equal to that given by the Mantel–Haenszel estimate in the previous section. Hence, in this example, all three methods led to exactly the same conclusions. This is not always the case however; on pages 62–3, instances are discussed when the results of these methods may differ and examines which methods are superior in those instances.

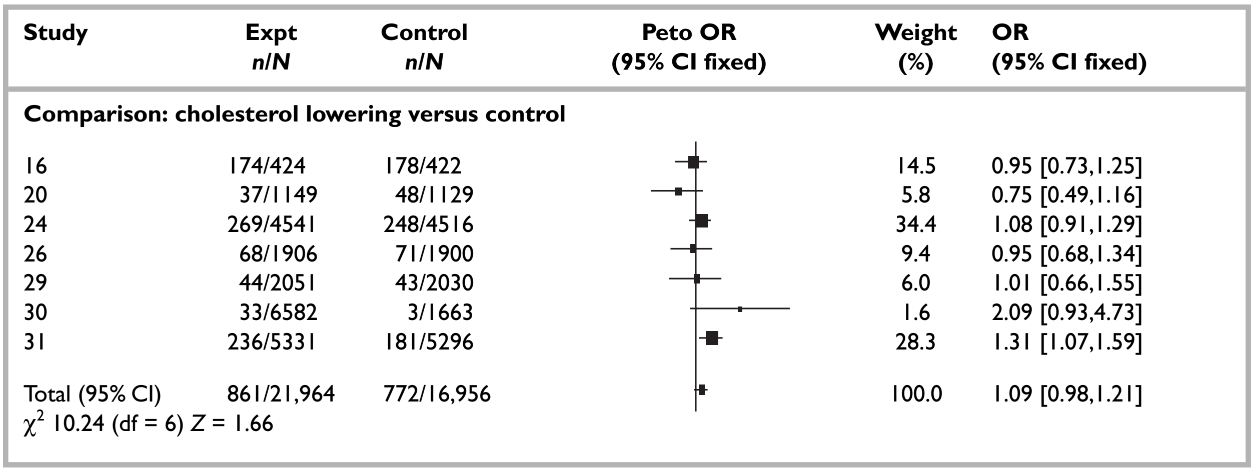
Combining ORs via ML techniques

ML techniques use iterative procedures and therefore need a computer for their implementation.

MLEs are difficult to compute exactly, but they are the most efficient for large sample sizes. Unfortunately, there is no way of knowing how large the sample sizes must be for this property to hold (6). The MLE of θ is based on the likelihood of the *k* studies and can be denoted (6):

$$L \propto \prod_{i=1}^{k} \theta_{ci}^{b_i}(1 - \theta_{ci})^{d_i}\,\theta_{ti}^{a_i}(1 - \theta_{ti})^{c_i} \qquad (9.14)$$

***FIGURE 3** Plot of studies combined using the Peto method (obtained using Meta View)*



Study	Expt n/N	Control n/N	Peto OR (95% CI fixed)	Weight (%)	OR (95% CI fixed)
Comparison: cholesterol lowering versus control					
16	174/424	178/422		14.5	0.95 [0.73,1.25]
20	37/1149	48/1129		5.8	0.75 [0.49,1.16]
24	269/4541	248/4516		34.4	1.08 [0.91,1.29]
26	68/1906	71/1900		9.4	0.95 [0.68,1.34]
29	44/2051	43/2030		6.0	1.01 [0.66,1.55]
30	33/6582	3/1663		1.6	2.09 [0.93,4.73]
31	236/5331	181/5296		28.3	1.31 [1.07,1.59]
Total (95% CI)	861/21,964	772/16,956		100.0	1.09 [0.98,1.21]

χ^2 10.24 (df = 6) Z = 1.66

subject to:

$$\text{OR}_{Umle} = \theta_{ti}(1 - \theta_{ci}) / \theta_{ci}(1 - \theta_{ti}) \qquad (9.15)$$

This is an unconditional estimate. Emerson (11) reports that Breslow found that unconditional MLE, which had earlier been investigated by Gart, is not consistent for estimating the OR when the number of counts remained bounded.

Conditional MLEs also exist; they use uses the conditional distribution of the data in each table, given the fixed values for the total counts in the margins. The conditioning leads to an estimator that is consistent and asymptotically normal (11). For formulae see (21). In a study investigating their relative merits, it was found superior to the unconditional MLE, and equal or superior to the Mantel–Haenszel estimator in both bias and precision (21). However, both theory and simulation suggest that (conditional) MLE does not stand up as well as the Mantel–Haenszel estimator under departures from the assumption of independent trials (11).

Emerson reports (11) new non-iterative procedures (including jackknife) (that are asymptotically optimal under the classical assumptions of independence and homogeneity of ORs) have been developed. He comments:

> 'Although these estimators seem to be competitive with the conditional maximum likelihood estimators under the classical assumptions, further research is needed to determine whether any of them exhibit the robustness of the Mantel–Haenszel estimator.' (11)

This is a very brief outline of these methods, a recommended starting point for further investigation is the excellent review by Emerson (11).

Exact methods of interval estimation

The above methods for interval estimation are all asymptotic; their justification assumes either that the counts are large or that the number of strata is large. Exact methods do exist that are not restrained in this way, and are based on exact distribution theory. Although these methods have long been available in principle, modern computer power (using network algorithms) now makes them routinely available. A detailed description of these methods are beyond the scope of this report, the interested reader is referred to (11) for a review of this topic.

More methods for combining ORs

It is pointed out that other methods do exist for combining ORs, again Emerson (11) would be an excellent starting point for further investigation.

Discussion of the relative merits of each method

Having a number of different approaches to combine ORs at the researcher's disposal, it would be desirable to have guidelines indicating when a particular method is most appropriate, and when an alternative procedure would be preferred.

The Peto method has come under strong criticism. It has been demonstrated that this method is capable of producing seriously biased ORs and corresponding SEs when there is severe imbalance in the numbers in the two groups being compared (22). Bias is also possible when the estimated OR is far from unity (23). Having several alternative methods available, Fleiss went on to comment (23) that there is no compelling reason for the Peto method to be employed. Fleiss (24) also describes conditions under which the inverse-weighted and the Mantel–Haenszel method are to be preferred: If the number of studies to be combined is small, but the within-study sample sizes per study are large, the inverse-weighted method should be used. If one has many studies to combine, but the within-study sample size in each study is small, the Mantel–Haenszel method is preferred.

A comparison between the Mantel–Haenszel and (conditional and unconditional) ML techniques has been carried out. Generally, if the sample sizes of the studies are large (all cells ≥ 5) the methods will give almost identical results. If there are cells with counts of < 5 then there will be differences between the methods but these will be small. In conclusion, as there seem to be no clear benefits to be reaped from the difficult computation of the ML method, using the inverse-weighted and Mantel–Haenzel methods when indicated would seem the best strategy in most cases. If, however, samples sizes are small for individual studies exact methods may be preferred (22).

Another factor that needs considering is whether any of the cells have zero events. Recently Sankey *et al.* (25) carried out an assessment of the use of the continuity correction (adding 0.5 to each cell) for sparse data in meta-analysis, using the Mantel–Haenszel estimate. They report:

> 'A study with a 0 cell in the treatment group produces a point estimate of 0.0 for the OR and contributes only to the denominator of the Mantel–Haenszel summary measure. When zero events are observed in the control group, the study odds ratio estimate is undefined and it contributes only to the numerator of the summary measure. Studies with 0

> total observed events contribute no information to the Mantel–Haenszel odds ratio. These studies are also not included in the *Q*-statistic to test for homogeneity, and hence do not add a degree of freedom to the associated chi-squared statistic.' (25)

Thus a study with zero total events is completely excluded from the analysis if no continuity correction is used. It has been argued that dropping them in this way is acceptable because they provide no information on the magnitude of the treatment effect (26). However, Sankey *et al.* consider this as unappealing as a trial with zero events from 200 subjects would be equally non-informative as a trial with only 20 subjects, and hence conclude:

> '... a meta-analysis involving sparse data should usually employ the continuity correction. The only observable exception to this would be if one prefers to use the fixed effect Mantel–Haenszel summary measure and there is strong evidence suggesting that very little heterogeneity exists among component studies. In this situation, the uncorrected method performs very well and the only problem facing the investigator is explaining why studies with zero total events have been excluded from the analysis. In all other sparse data the correction should be employed. The evidence shows that it is at least as good as the uncorrected method, and in some cases clearly superior.' (25)

Recently, another factor has been identified that may be important when carrying out a fixed effects meta-analysis. Mengersen *et al.* (27) compared the ways in which CIs for ORs were calculated for individual studies. They compared the calculation of the ORs in epidemiological studies investigating the effect of exposure to environmental smoke on lung cancer. An exact test (Fisher's) was compared to the Mantel–Haenszel method and the logit variance approximation (used in this instance to calculate OR from each individual stratified study as opposed to across studies to combine estimates). They concluded:

> 'exact methods might increase estimated confidence interval widths by 5–20% over standard approximate (logit and Mantel–Haenszel) methods, and that these methods themselves differ by this order of magnitude.' (27)

Emerson, however, gives a slightly different impression:

> 'Simulations suggest that exact methods do not clearly outperform those associated with Mantel–Haenszel, except perhaps with highly unusual configurations of data' (11)

This is a new concern in meta-analysis and one that may need addressing further due to these conflicting reports.

Finally, Emerson (11) gives formal guidelines on the procedure that should be followed when combining ORs. To the authors of this report's knowledge this has not yet been applied to meta-analysis methodology; however, there seems little reason why it should not. These guidelines are reproduced in *Box 4.*

Combining treatment effect estimates measured on a continuous scale

There are many different continuous scales used to measure outcome in the medical literature e.g. lung function, pulse rate, weight and blood pressure. A property they all have in common is that they are all measured on a positive scale. For this reason, it is common practice to use a logarithmic transformation on the data and then use normal distribution theory (6). Usually the parameter of interest is the difference in effect size between the treatment and control groups. If it can be assumed that all the studies estimate the same parameter and the estimates of continuous-outcome measures are approximately normal, then the inverse variance-weighted method can be used directly, combining the data in their original metric. If different studies measured their outcomes on different scales then synthesis is still possible, but the data first needs standardising. However, it should be noted by doing this the resulting estimate may be difficult to interpret clinically. Both methods are described followed by an example.

Combining data in its original metric

If the data are approximately normal and the outcomes of all the studies to be combined are measured on the same scale, then this method is appropriate. The measure of treatment effect is given by:

$$T_i = \mu_{ti} - \mu_{ci} \tag{9.16}$$

where μ_{ti} and μ_{ci} are the mean responses in the *i*th study for the treatment and control group, respectively.

The variance of this treatment difference is:

$$\text{var}(T_i) = \sigma_i^2(1/n_i^t + 1/n_i^c) \tag{9.17}$$

where n^t is the within-study sample size in the treatment group, n^c is the within-study sample size for the control group, and σ_i^2 is the assumed common variance.

BOX 4 Guidelines for combining ORs [reproduced from (11)]

1. Calculate the Mantel–Haenszel estimate of the common OR, unless a combination of extreme values in all tables leads to degeneracy. This estimate performs well in a wide variety of circumstances. It can withstand departures from standard assumptions of independent subject responses and homogeneity of ORs across strata at least as well as other methods. It performs well for many tables having small counts unless the data give degeneracy.
2. Use the Robins *et al.* estimate of variance of the log-OR to provide a confidence interval for the Mantel–Haenszel estimate. This method of interval estimation gives relatively short intervals with coverage close to the nominal level (usually 95%) in a wide variety of circumstances. The method works except in unusual situations for which each table has an extreme configuration of counts, and it can be carried out on a hand-held calculator.
3. Calculate the conditional ML estimate of the OR when the total count is under 1000, or when the tables show severe imbalance in their marginal counts. For example, when the total count is more than 1000 but one of the four marginal totals is a single-digit number in all tables, we would calculate the conditional ML estimate as a check on the Mantel–Haenszel estimate. If the sum over all tables of the counts in any single position is 0, the estimate is left undefined.
4. Use exact methods to provide a confidence interval for the conditional, ML estimate of a common OR. We recommend using the mid-*P* adjustment when giving an exact confidence interval, because it tends to give shorter and thus more informative intervals while retaining the desired level of coverage.
5. When the exact analyses give results that differ substantially from those of the Mantel–Haenszel methods, we recommend that both analyses be reported. We also recommend including a brief discussion of the potential reason for the discrepancy – a collection of tables that is very close to giving degeneracy of the estimates, substantial heterogeneity of sample ORs across the tables, or strong imbalances among the marginal totals of the 2 × 2 tables.
6. Recommendations (3) and (4) require the use of special computer software; the needed software is incorporated in several commercially available statistical packages for microcomputers including StatXact, Egret, Statcalc, and Systat.
7. We recommend against reporting other analyses: those associated with the Peto method, those using the empirical logit, and those based on unconditional ML techniques.*

* *However, Whitehead and Jones point out: 'One potential problem with the ML method (which is the same as for the Mantel–Haenszel method in the binary case) is that it cannot be calculated if one of the cells in the 2× 2 table is zero. There is only a problem with the Peto method if there are no successes in total or no failures in total.' (38). This implies there may be a use for the Peto method in meta-analysis.*

Synthesis can then proceed using the inverse variance-weighted method described on pages 55–6.

Standardised mean differences

If the normal distribution assumption seems reasonable, but the studies estimate different parameters, the method of standardised mean differences should be used instead. The effect size of an experiment, *d*, is defined as:

$$d = T_{i(STD)} = (\mu_i^t - \mu_i^c)/s_i^* \qquad (9.18)$$

where μ_i^t and μ_i^c are the sample means of the treated and control arms, respectively, and s_i^* is the estimate of the standard deviation of the *i*th study. s_i^* can be defined in different ways, each of which will yield a different estimate. Common and intuitive choices for s_i^* are s_i^{t*}, and s_i^{c*}, which are the standard deviations of the treatment and control group, respectively. Alternatively, a pooled standard deviation combining both s_i^{t*} and s_i^{c*} could be used.[5]

Hedges and Olkin [(28), p. 78] suggest using the pooled estimate for the standard deviation, if it is reasonable to assume equal population variances. They go on to show that this estimate has both smaller bias and variance than using, s_i^{c*}, the control standard deviation as suggested by Glass (29). [See Hedges and Olkin (28) and Rosenthal (4) for a thorough treatment of the alternative measures of effect difference variance]. The formula for this pooled sample standard deviation is:

$$s_i = \sqrt{\left(\frac{(n_i^t - 1)(s_i^t)^2 + (n_i^c - 1)(s_i^c)^2}{n_i^t + n_i^c - 2}\right)} \qquad (9.19)$$

where n_i^t and n_i^c are the treatment and control group sample sizes, respectively.

The estimate, *d*, has small sample bias, and a correction equation has been derived, for formulae and a thorough account of this topic [see Hedges and Olkin (28) p. 81].

The variance of this estimate of effect difference is difficult to compute exactly, however if the underlying data can be assumed to be normal the conditional variance of $T_{i(STD)}$ can be estimated as:

[5] Another alternative, discussed by Rosenthal (4), used when the *S*s of the two groups differ greatly, is to transform the data to make the *S*s more similar. Such transformations require having access to the original data.

$$v(d) = \frac{n^T + n^C}{n^T n^C} + \frac{d^2}{2(n^T + n^C)} \qquad (9.20)$$

where n^T and n^C are the numbers in the treatment and control groups respectively, and d is the observed standardised mean difference (note: the is have been omitted).

More exact methods are possible using a computer intensive method; for a discussion see (28).[6]

Fleiss (23) states that if the sample sizes in the two treatment groups (n^T and n^C) are both large, and the population variances are equal then the simpler variance approximation

$$v(d) = \frac{n^T + n^C}{n^T n^C} \qquad (9.21)$$

can be used. Whichever variance estimate is used for the standardised mean difference from each study, synthesis can proceed using the inverse variance-weighted method.

If the data appear to be non-normal-skewed they can often be transformed to achieve, at least, an approximately normal distribution. If this is the case one can proceed using the above methods on the transformed data.

One drawback to doing this is that different answers will be obtained for the transformed data if the normality assumption was not met. For this reason, a non-parametric estimate was developed by Kraemer and Andrews (30) and extended by Hedges and Olkin (31), which is unaffected by monotonic transformations of the observations. These methods are presented in their entirety in (28), p. 92.

If the data are censored in any way, such as is often the case for survival data, special methods are needed. These are covered under chapter 20 on survival data.

However, Greenland refutes the use of standardised effect measures stating:

> 'By expressing effects in standard deviation units, one can make studies with identical results spuriously appear to yield different results; one can even reverse the order of strength of the results.' (32)

Other measures of the difference between two groups do exist, though are not used as commonly. A large selection of these are discussed by Rosenthal (4). The other type of continuous outcome not mentioned here is the correlation coefficient; this is dealt with in chapter 14, covering other scales of measurement.

The effect of mental health treatment on medical utilisation – combining treatment effect estimates measured on a continuous scale

Table 9 [modified from (23), *Table 1* p. 125] presents data from five comparative studies selected from more than 50 analysed by Mumford *et al.* for the effect of psychotherapy on patients hospitalised for medical reasons (33). The outcome measure was, in some studies, the number of readmissions to hospital, and in other studies, the number of days in hospital. Clearly two different scales are being used here, so it is necessary to combine standardised treatment estimates using the methods described in the previous section.

TABLE 9 *Data for five studies of the effect of mental health treatment on medical utilisation [adapted from (34)]*

	Psychotherapy			Control			
Study	n_1	$\bar{X}_1$	sd_1	n_2	$\bar{X}_2$	sd_2	s^*
1	13	5.0	4.7	13	6.5	3.8	4.27
2	30	4.90	1.71	50	6.10	2.3	2.10
3	35	22.5	3.44	25	24.9	10.65	7.91
4	20	12.5	1.47	20	12.3	1.66	1.57
5	8	6.50	0.76	8	7.38	1.41	1.13

* *s is the square root of the weighted average of* sd_l^2 *and* sd_2^2

Before carrying out a fixed effect analysis, it is wise to test the homogeneity assumption. This is done using the test outlined on pages 39–41.

Summary statistics derived from *Table 9* are given in *Table 10.*

For the five studies presented above:

$$Q = 11.109 - 20.137^2/56.75 = 3.96$$

[6] Hedges and Olkin [(28), p. 82] compare four different estimators of effect difference and conclude the only real differences exist when there are less than 16 degrees of freedom, which is unrealistic in practical applications.

***TABLE 10** Summary statistics for meta-analysis derived from Table 9 [adapted from (34), with corrections]*

Study	Y*	W**	WY	WY^2
1	0.351	6.50	2.282	0.801
2	0.571	18.75	10.706	6.113
3	0.303	17.50	5.303	1.607
4	–0.127	10.00	–1.270	0.161
5	0.779	4.00	3.116	2.427
Sum		56.75	20.137	11.109

* Y *is the standardised difference between the two means in* Table 9; Y = $(\bar{X}_2 - \bar{X}_1)/s$

** W *is the study-specific weighting factor,* W = $n_1n_2/(n_1 + n_2)$

This is not statistically significant (compared to 10% critical value of the χ^2 distribution with 4 df), therefore one can proceed with the fixed effects analysis.

Combining the weighted average of the five effect sizes using standardised treatment effect gives:

$$\bar{T}_{STD} = \frac{20.137}{56.75} = 0.355$$

with SE:

$$SE(\bar{T}_{STD}) = \frac{1}{\sqrt{56.75}} = 0.133$$

Calculating a 95% CI for the common underlying effect size

$$\text{lower limit} = 0.355 - 1.96 \times 0.133 = 0.09$$

$$\text{upper limit} = 0.355 + 1.96 \times 0.133 = 0.62$$

Because the CI excludes the value 0, one may reject the hypothesis that θ = 0, suggesting a benefit from the use of psychotherapy. *Figure 4* summarises this analysis.

Further research

Guidelines for which method to use in given situations when combining on the OR scale, i.e. which methods are valid under which circumstances, is there a role for the Peto method?

Use and implications of the exact methods; should they be used? If so, under what conditions?

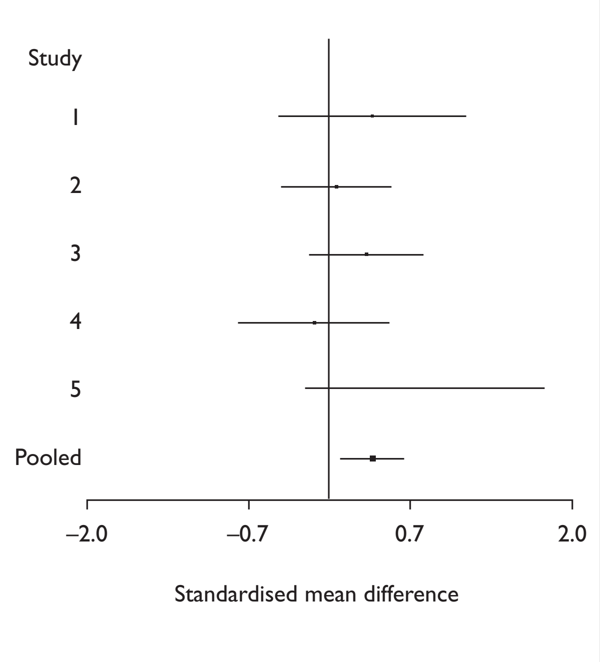


***FIGURE 4** Results of pooling five studies of the effect of mental health treatment on medical utilisation*

Clear guidelines on how to proceed when zeros are present in 2 × 2 tables to be combined; including clear advice on the exact form of any continuity correction factors that should be used.

Summary

This chapter has considered the so called fixed effect approach to meta-analysis. This assumes that all the studies in a meta-analysis are estimating the same underlying unknown true intervention effect. A variety of estimation methods have been proposed for such models, whilst in many situations they give qualitatively similar results, in some circumstances differences can be serious. In terms of binary data, problems with a number of methods occur if there are zero events in any treatment arms in any study. In such circumstances there has been some empirical work reported on the various methods advocated for overcoming this problem. Meta-analysts should report precisely what methods have been used in such circumstances.

References

1. Birge RT. The calculation of errors by the method of least squares. *Phys Rev* 1932;**16**:1–32.

2. Cochran WG. Problems arising in the analysis of a series of similar experiments. *J R Statist Soc* 1937;**4**:102–18.

3. Shadish WR, Haddock CK, Cooper H, Hedges LV, editors. Combining estimates of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 261–84.

4. Rosenthal R, Cooper H, Hedges LV, editors. Parametric measures of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation; 1994, p. 231–44.

5. Fleiss JL, Cooper H, Hedges LV, editors. Measures of effect size for categorical data. The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 245–60.

6. Hasselblad VIC, Mccrory DC. Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making* 1995;**15**:81–96.

7. Li YZ, Shi L, Roth HD. The bias of the commonly-used estimate of variance in metaanalysis. *Commun Statist Theory Methods* 1994;**23**:1063–85.

8. Meinert CL. Clinical trials dictionary: terminology and usage recommendations. Baltimore, Maryland: The Johns Hopkins Center for Clinical Trials, 1996.

9. Gart JJ, Zweifel JR. On the bias of various estimators of the logit and its variance, with application to quantal bioassay. *Biometrika* 1967;**54**:471–5.

10. Smith GD, Song F, Sheldon TA, Song FJ. Cholesterol lowering and mortality: the importance of considering initial level of risk. *BMJ* 1993;**306**: 1367–73.

11. Emerson JD. Combining estimates of the OR: the state of the art (review). *Stat Methods Med Res* 1994;**3**:157–78.

12. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;**22**:719–48.

13. Mantel N. Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure. *J Am Statist Assoc* 1963;**58**:690–700.

14. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science (review). *Epidemiol Rev* 1992;**14**:154–76.

15. Robins J, Breslow N, Greenland S. Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 1986;**42**:311–23.

16. Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel–Haenszel OR. *Am J Epidemiol* 1986;**124**:719–23.

17. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II: Analysis and examples. *Br J Cancer* 1977;**35**:1–39.

18. Yusuf S, Peto R, Lewis J, Collins R, Sleight P, *et al.* Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Prog Cardiovasc Dis* 1985;**27**:335–71.

19. Spector TD, Thompson SG. Research methods in epidemiology. 5. The potential and limitations of metaanalysis. *J Epidemiol Comm Health* 1991;**45**:89–92.

20. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Stat Med* 1989;**8**:141–51.

21. Hauck WW. A comparative study of conditional ML estimation of a common OR. *Biometrics* 1984;**40**:1117–23.

22. Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Stat Med* 1990;**9**:247–52.

23. Fleiss JL. The statistical basis of meta-analysis (review). *Stat Methods Med Res* 1993;**2**:121–45.

24. Fleiss JL. Statistical methods for rates and proportions. 2nd edn. New York: Wiley, 1981.

25. Sankey SS, Weissfeld LA, Fine MJ, Kapoor W. An assessment of the use of the continuity correction for sparse data in metaanalysis. *Commun Statist Simulation Computation* 1996;**25**:1031–56.

26. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. *Stat Med* 1991;**10**:1665–77.

27. Mengersen KL, Tweedie RL, Biggerstaff BJ. The impact of method choice in meta-analysis. *Aust J Stats* 1995;**37**:19–44.

28. Hedges LV, Olkin I. Statistical methods for meta-analysis. London: Academic Press, 1985.

29. Glass GV. Primary, secondary and meta-analysis of research. *Educ Res* 1976;**5**:3–8.

30. Kraemer HC, Andrews G. A non-parametric technique for meta-analysis effect size calculation. *Psychol Bull* 1982;**91**:404–12.

31. Hedges LV, Olkin I. Nonparametric estimators of effect size in meta-analysis. *Psychol Bull* 1984;**96**:573–80.

32. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.

33. Mumford E, Schlesinger HJ, Glass GV, Patrick C, Cuerdon T. A new look at evidence about reduced cost of medical treatment utilization following mental health treatment. *Am J Psychiatry* 1984;**141**:1145–58.

34. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127–39.

35. Phillips A, Holland PW. Estimators of the variance of the Mantel–Haenszel log-odds-ratio estimate. *Biometrics* 1987;**43**:425–31.

36. Sato T. Confidence limits for the common OR based on the asymptotic distribution of the Mantel–Haenszel estimator. *Biometrics* 1990;**46**:71–80.

37. Pigeot I. A jacknife estimator for a combined OR. *Biometrics* 1991;**47**:420–3.

38. Whitehead A, Jones NMB. A meta-analysis of clinical trials involving different classifications of response into ordered categories. *Stat Med* 1994;**13**:2503–15.

Chapter 10

Random effects methods for combining data

Introduction

An assumption made when implementing a fixed effect model is that all the studies are estimating the same underlying effect size (i.e. H_0: $\theta_1 = \theta_2 = \dots \theta_k$). Pages 39–41 reported the test for heterogeneity, which tests this hypothesis. When a low *p*-value is obtained from this test, the above assumption is violated and doubts exit as to whether the fixed effect model is wholly appropriate. Thompson comments:

> 'With a fixed effect method, the confidence interval for the overall treatment effect reflects the random variation within each trial but not potential heterogeneity between trials. In terms of extrapolation on future patients, the confidence interval is therefore artificially narrow' (1).

Pages 43–8 suggested ways of dealing with heterogeneity, one of which was to use include the cause of heterogeneity (such as age of population, dose level of treatment etc.) as a covariate in the analysis. Meta-regression techniques for doing this are given in chapter 11. If no variables available appear to explain, or only partly explain, the apparent heterogeneity, a different model for the treatment effect is required. The random effects model described in this chapter presents a way of modelling this extra variation, when no covariates are included. This methodology is extended in chapter 12 to cover the inclusion of variables, that partly explain the heterogeneity, within a random effects framework, these are usually called mixed models.

It has been clearly established, that the test of heterogeneity has low power (see pages 39–41); thus, even when a result not significant at the 5% level is returned, there is a good chance there may still be a degree of underlying heterogeneity. For this reason, in certain circumstances, by considering other evidence, such as descriptions of study designs, study populations, dose levels, and graphical plots of effect size (see pages 39–48), the assumption of one fixed effect size underlying all the studies may still seem unrealistic. In this situation, random effects models can be used. In fact, some people consider that by the very nature of biomedical experiments, some degree of heterogeneity is always present, for this reason random effects models should be used as a matter of course.

Random effects models are not without their critics though, and their appropriateness has been a matter of considerable debate over the past decade. A summary of some of these arguments, both advocating and rejecting their use, is given at the end of the chapter.

Concept behind random effects models

A way is needed of taking into account the extra variation incurred, when assuming the studies are estimating different (underlying) effect sizes. These underlying effects are assumed to vary at random within the model presented. More specifically, to make modelling possible, they are assumed to vary according to a given distribution. In addition, the variation caused by sampling error described in the fixed effects model is still present. A random effects model has to take into account both these forms of uncertainty. (NB: It may not be truly random – there may be a clear reason for the differences that could be explained by a single covariate; however, this may not have been available for certain studies, so the relationship went undetected. Alternatively the heterogeneity could have been the result of the effect of many, even hundreds of factors, each of which contributed only a small amount to the variation, so detecting them was impossible.)

The, now standard, model that allowed for random variation of the underlying effect size between studies was described in 1986 by DerSimonian and Laird (2). Their model assumes that the study specific effect sizes come from a random distribution of effect sizes with a fixed mean and variance. This assumption has caused much dispute; it suggests that each study comes from an infinite sample of similar studies, a concept some people feel is unrealistic. It should be noted, however, that random effects models have a long history in other fields of application.

So, the total variation of the estimated effect size can be broken down into two exclusive components:

Variance of estimated effects = random effects variance + estimation variance

If the random effects variance was zero, the above model would reduce exactly to the fixed effects model described in chapter 9.

Expressed algebraically, where T_i is an estimate of effect size and θ_i is the true effect size:

$$T_i = \theta_i + e_i \qquad (10.1)$$

where e_i is the error with which T_i estimates θ_i.

and

$$\mathrm{Var}(T_i) = \tau^2_\theta + v_i \qquad (10.2)$$

where τ^2_θ is the random effects variance and v_i is the variance due to sampling error.

Algebraic derivation for random effects models

Random effects models are more complex than those for fixed effects, and the formulae presented are similarly more involved. Specialised software will be required, in many instances, to implement these.

Formulae can be derived using two different approaches, both of which are outlined in (3) and are reproduced here.

Weighted method

Firstly starting with the general inverse weighted variance model first presented on pages 55–6.

$$\bar{T}. = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \qquad (9.1)$$

where

$$w_i = \frac{1}{v_i} \qquad (9.2)$$

Recalling the test for heterogeneity from pages 39–41, this can be seen as measuring study-to-study variation in effect size.

$$Q = \sum_{i=1}^{k} w_i (T_i - \bar{T})^2 \qquad (8.1)$$

Under the assumption that the studies are a random sample from a larger population of studies, there is a mean population effect size, say $\bar{\varphi}$, about which the study-specific effect sizes vary (4). This is the parameter we primarily wish to estimate.

Let denote, $\hat{\tau}^2$, the variance of the studies effect sizes (an estimate for τ^2_θ), a quantity yet to be determined. Further define $\bar{w}$ and s^2_w to be the mean and variance of the weights (ws):

$$\bar{w} = \sum_{i=1}^{k} w_i \Big/ k \qquad (10.3)$$

and

$$s^2_w = \frac{1}{k-1}\left(\sum_{i=1}^{k} w_i^2 - k\bar{w}^2\right) \qquad (10.4)$$

Further, define:

$$U = (k-1)\left(\bar{w} - \frac{s^2_w}{k\bar{w}}\right) \qquad (10.5)$$

The estimated component of variance due to interstudy variation in effect size, $\hat{\tau}^2$, is calculated as:[1]

$$\hat{\tau}^2 = 0 \qquad \text{if } Q \le k-1$$

and

$$\hat{\tau}^2 = \left(\mathrm{Q} - (\mathrm{k}-1)\right)/\mathrm{U} \qquad \text{if Q} > \mathrm{k}-1 \qquad (10.6)$$

Now, the adjusted weights for each of the studies are calculated, define $w_i{*}$ as:

$$w_i{*} = \frac{1}{[(1/w_i) + \hat{\tau}^2]} \qquad (10.7)$$

(i.e. The random effects study weighting is given by the reciprocal of the sum of the between and within study variances.)

[1] An alternative, equivalent expression for the estimate of τ^2_θ given in several textbooks and papers [e.g. (5,34)] is given by:

$$\hat{\tau}^2 = \max\left[0, \left[Q - (m-1)\right] \Big/ \left[\sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i\right]\right]$$

and is sometimes known as the weighted estimate (3).

The treatment point estimate and $\alpha/2$ CI for $\bar{\theta}$, the mean treatment effect of all studies, can then be computed by:

$$\bar{T}._{RND} = \sum_{i=1}^{k} w_i^* T_i \bigg/ \sum_{i=1}^{k} w_i^* \qquad (10.8)$$

$$\mathrm{var}(\bar{T}._{RND}) = \sum_{i=1}^{k} 1/w_i^* \qquad (10.9)$$

$$\bar{T}._{RND} - z_{\alpha/2} \bigg/ \sqrt{\left(\sum_{i=1}^{k} w_i^*\right)} \le \bar{\theta} \le \bar{T}._{RND} + z_{\alpha/2} \bigg/ \sqrt{\left(\sum_{i=1}^{k} w_i^*\right)} \qquad (10.10)$$

Note: The CI given here assumes normality, unlike the rest of the derivation.

It is worth noting that if the test for heterogeneity is significant the random effects CI for the treatment effect will always be larger than for a fixed effects analysis, on the same data, due to the extra level of variability being accounted for by including σ_θ^2 in the formulae.

Alternative derivation – the unweighted method

Start with the ordinary (unweighted) sample estimate of the variance of the effect sizes, T_1, …, T_k, computed as:

$$s^2(T) = \sum_{i=1}^{k} \left[(T_i - \bar{T}^2) \big/ (k-1) \right] \qquad (10.11)$$

The expected value of $s^2(T)$ (i.e. the unconditional variance we would expect to be associated with any particular effect size) is:

$$E\left[s^2(T)\right] = \sigma_\theta^2 + (1/k)\sum_{i=1}^{k} \sigma^2(T_i \backslash \theta_i) \qquad (10.12)$$

To estimate $\sigma\,(T_i\backslash\theta_i)$, one needs to use v_i, which varies depending on which scale estimates are being combined on. For ORs (using Mantel–Haenszel method) equation (9.9) can be used, for standardised effect sizes (9.20). For other scales see chapter 14.[2]

Using these estimates equation (10.12) can be solved to obtain an estimate for the variance component:

$$\hat{\sigma}_\theta^2 = s^2(T) - (1/k)\sum_{i=1}^{k} v_i \qquad (10.13)$$

If this value is negative it is set at 0.

(I.e. this is competing with, $\hat{\tau}^2 = \hat{\sigma}_\theta^2$, where

$$\hat{\tau}^2 = \max\left[0, \left[Q - (m-1)\right] \bigg/ \left[\sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i^2 \bigg/ \sum_{i=1}^{k} w_i\right]\right]\Bigg)$$

The two methods outlined above are both non-iterative. Solutions are possible via ML and restricted maximum likelihood (REML); these are outlined below. Both use iterative algorithms and hence are more computer intensive.

Solving the formula using the normal–normal model (ML and REML estimate solutions)

If it is assumed that each of the underlying effect parameters, the θ_js, come from a normal distribution, with mean μ and variance τ^2, [and $\mathrm{T_i}$ is $\mathrm{N}(\theta_\mathrm{i}, \mathrm{s}_\mathrm{i}^2)$ (2)] then the likelihood is proportional to (5):[3]

$$L \propto \exp\left[-\sum_{i=1}^{k} \left[(\hat{\theta}_\mathrm{i} - \mu)^2 \big/ (\tau^2 + \sigma_i^2) + \ln(\tau^2 + \sigma_i^2) \right] \big/ 2 \right] \qquad (10.14)$$

Approximate solutions to this model have been given by DerSimonian and Laird (2)[4] and Hedges (6). Also, it is possible to calculate MLEs directly, or Bayesian estimates can be calculated with the specification of a prior (see chapter 13 on Bayesian methods).

Summary of methods

In summary, there are four different ways to carry out a random effects meta-analysis. Two of the methods are non-iterative, and have been called the weighted and non-weighted approaches. Two are iterative both of these require the extra assumption that the underlying distribution of study effect sizes are normally distributed (though all four need this assumption to construct CIs). These are referred to as the MLE and the REML estimate. The likelihood to be maximised is slightly modified using REML (from that of MLE), to

[2] Hedges and Olkin (38) note that more exact estimates of conditional variability under the random effects model exist, however their use makes little practical difference.

[3] For an alternative derivation of a likelihood based random effects model see (35), p. 144.

[4] DerSimonian and Laird used the EM algorithm (13) (which is an iterative procedure for computing MLEs appropriate when the observations can be viewed as incomplete data) to calculate MLE [equations given by Rao *et al.* (39)] and REML [equations reviewed by Harville (40)] solutions. In REML estimation, the likelihood to be maximised is slightly modified to adjust for μ and τ^2 being estimated from the same data (2).

adjust for the fact that the underlying mean and variance are being estimated from the same data. The REML are the iterative equivalent to the weighted estimators (2). Obtaining solutions for these latter two approaches is more difficult than for the non-iterative ones.

Discussion of the merits of each method

With four different methods of estimation to choose from it would be desirable to establish guidelines on which method to use in a given situation. The four methods were compared by DerSimonian and Laird (2) who re-analysed eight meta-analyses using all four methods. They commented that 'The weighted method and the REML estimation procedures consistently yield slightly higher values of $\hat{\tau}^2$ (the random effects variance) than the ML procedure. This is because both these procedures adjust for $\bar{T}._{RND}$ and $\hat{\tau}^2$ being estimated from the same data where as the MLE procedure does not.' In addition, 'Comparing the unweighted method of moments with the other three methods, we find that the estimates for $\hat{\tau}^2$ from this method differ, and sometimes differ widely, from the estimates of the other three methods but without any consistent pattern. The estimates of $\bar{T}._{RND}$ and its SE from the unweighted method also differ from the estimates of the other three methods.' (2)

So it seems that the unweighted differs considerably from the other three. Shadish and Haddock (3) comment that the relative merits of each of the above methods have not been widely stated, the main difference between them being that the weighted method gives a non-zero estimate of the variance component only if the homogeneity statistic Q is larger than its expected value under the null hypothesis. In conclusion, DerSimonian and Laird suggested 'that the weighted noniterative method is an attractive procedure because of the comparability of its estimates with those of the ML methods and because of its relative simplicity.' (2)[5]

However, all these methods have one disadvantage that is clearly explained by DerSimonian and Laird:

> 'in all our work we assume that the sampling variances are known, although in reality we estimate them from the data. Further research needs to be done in this area as there are alternative estimators that might be preferable to the ones we use. For instance, if the sample sizes in each study are small, then sampling variances based on pooled estimates of the proportions in the treatment and control groups might be better than the ones based on estimates of proportions from individual studies. Another alternative is to shrink the individual proportions towards a pooled estimate before calculating the variances. Further investigation is needed before one single method emerges as superior.' (2)

Very recently, new estimates have been developed which take this uncertainty into account. These are discussed on pages 73–6 (extensions to the basic model).

Finally, it should be noted that Sankey *et al.* (7) recommend using the continuity correction (adding a half to cells) for sparse when the OR scale is being used to carry out a random effects analysis.

Examples of combining data using a random effects analysis

Example: effect on mortality of lowering serum cholesterol level

On pages 41–3, several fixed effect analyses were carried out using only the seven primary studies in the dataset. The test for heterogeneity for these studies led to a test statistic of $Q = 10.19$, which has a corresponding p-value of 0.117. This result, although not formally significant, led to concern that there may be a degree of heterogeneity between the studies, especially when the low power of the test is considered.

The same studies are combined below, this time using a random effects model. The weighted non-iterative approach is used in this example.

The first step is to calculate the mean and variance of the within-study weights. The weighted values were worked out for the fixed effects analysis and are displayed in *Table 6*. Using equation (10.1):

$$\bar{w} = \frac{(51.38+\ldots+98.48)}{7} = 49.88$$

and equation (10.12)

$$s_w^2 = \frac{1}{7-1}(29130.91 - 7 \times 49.88^2) = 1952.34$$

[5] It would appear that the weighted non-iterative approach has become the most commonly used random effects model in meta-analysis. In many papers this method may simply referred to as the DerSimonian and Laird model.

Calculating U defined in equation (10.11):

$$U = (7 - 1)\left(49.88 - \frac{1952.34}{7 \times 49.88}\right) = 265.73$$

Calculating the estimated component of variance due to between-study variation in the values of the ORs ($\hat{\tau}^2$) from (10.12)

$$Q = 10.185 \text{ (calculated in chapter 9)}$$

therefore, $Q > 6$ (k – 1), so

$$\hat{\tau}^2 = (10.185 - (7 - 1))/265.73 = 0.016.$$

Now the weights $w_1^*,\ldots,w_7^*$ used in the random effects model can be calculated using equation (10.13).

So, for the first study

$$w_1^* = \frac{1}{[(1/51.37) + 0.016]} = 28.20$$

Table 11 displays these weights for the other studies.

It is instructive to examine how the relative weighting has changed between the fixed and the random effects models. It can be seen that using the random effects model the larger studies have been down weighted while the relative weighting of the smaller studies is increased. This trend generally holds true for all meta-analysis.

The pooled point estimate of the OR together with its associated 95% CI can be calculated from equations (10.12) and (10.11), respectively (remember, we are working on the natural log scale).

$$\bar{T}_{RND(ln(OR))} = \begin{matrix}[(28.20 \times (-0.051))+\ldots+ \\ (38.23 \times 0.270)]/(28.20+\ldots+38.23)\end{matrix} = 0.06$$

$$(\mathrm{SE}(\bar{T}_{RND(ln(OR))}) = 1/\sqrt{\left(\sum_{i=1}^{7} w_i^*\right)} = 0.078)$$

Calculating an approximate 95% CI for the combined log OR:

$$0.06 - 1.96 \times 0.078 \leq \ln(\bar{\theta}) \leq 0.06 + 1.96 \times 0.078$$

$$= [-0.09,0.21]$$

Converting back to OR scale gives:

$\bar{T}_{RND(OR)} = 1.06$ with approximate 95% CI [0.91,1.24]

A plot of these results is given in *Figure 5*.

This result can be compared with those obtained from fitting a fixed effects model, say the Mantel–Haenszel estimate obtained on pages 56–63. There the point estimate was 1.09, slightly higher than that of the random effects model above (1.06). Comparing CIs, using the Mantel–Haenszel method gave 0.98–1.21, whilst using a random effect model gave 0.91–1.24. The random effects derived interval is thus wider incorporating both higher and lower values than that of the corresponding fixed effects one. This is a typical result, as previously mentioned, the extra width is due to the between study variation being taken into account in the random effects analysis. The conclusion, is thus similar to that given earlier; the treatment effect is non-significant, but the result is more conservative.

TABLE 11 *Weighting of studies used in the weighted non-iterative random effects model*

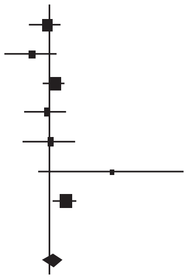
Study ID	T_i	ln(T_i)	w_i (% of total)	w_i^* (% of total)
16	0.95	–0.051	51.37 (14.7)	28.20 (17.3)
20	0.75	–0.288	20.13 (5.8)	15.23 (9.3)
24	1.08	0.077	121.68 (38.9)	41.29 (25.3)
28	0.95	–0.051	33.47 (9.6)	21.80 (13.4)
29	1.01	0.010	21.28 (6.1)	15.87 (9.7)
30	2.79	1.026	2.74 (0.8)	2.62 (1.6)
31	1.31	0.270	98.48 (28.2)	38.23 (23.4)

Extensions to the basic model

Accounting for extra uncertainty

Though the random effects model gives wider CIs than that of a corresponding fixed effect analysis, concerns have been raised that it is still too narrow and hence insufficiently conservative. Recent methodological advances have attempted to address this problem. It has been pointed out that the uncertainty due to $\hat{\tau}^2$ being estimated from the data has not been taken into account when estimating, $T_{\cdot RND}$, the overall treatment effect (8–10). Two approaches have been put forward to deal with this.

FIGURE 5 *Plot of combined cholesterol trials using a random effects model (obtained using Meta-View software developed by the Cochrane Collaboration)*

Study	Expt	Control	OR	Weight	OR
	n/*N*	*n*/*N*	(95% CI fixed)	(%)	(95% CI random)
Comparison: cholesterol lowering versus control					
16	174/424	178/422		17.3	0.95 [0.73,1.25]
20	37/1149	48/1129		9.3	0.75 [0.48,1.16]
24	269/4541	248/4516		25.4	1.08 [0.91,1.29]
26	68/1906	71/1900		13.3	0.95 [0.68,1.34]
29	44/2051	43/2030		9.7	1.01 [0.66,1.55]
30	33/6582	3/1663		1.6	2.79 [0.85,9.10]
31	236/5331	181/5296		23.5	1.31 [1.07,1.59]
Total (95% CI)	861/21,964	772/16,956		100.0	1.09 [0.91,1.24]
χ^2 10.24 (df = 6) Z = 1.66					

Firstly, Hardy and Thompson (8) propose a random effects model which gives a CI for the parameter $\hat{\tau}^2$ (the random effects variance). It also gives a CI for $\mathrm{T.}_{\mathrm{RND}}$ which takes into account the fact that $\hat{\tau}^2$ has to be estimated from the data. It uses a profile likelihood approach to calculate confidence regions, which assumes normality of the data. The approach yields a wider CI than the standard random effects approaches. The paper concludes that the proposed method is preferred when $\hat{\tau}^2$ has an important effect on the overall estimated treatment effect. A sensitivity plot of $\hat{\tau}^2$ against $T._{RND}$ is given to investigate the robustness of $T._{RND}$ to changes in the value of $\hat{\hat{\tau}}^2$; this can be used to provide insight into whether the likelihood method is required or whether the simpler standard random effects analysis using a moment estimator of the between-study variance is adequate. This method can be applied to continuous, ordinal and survival outcome measures as well as binary.[6]

In addition, Hardy and Thompson also comment (8): '(This method) still assumes that the individual study variances are known, when in practice they too must be estimated. The full likelihood, in the case of binomial data, includes the conditional distribution of each 2 × 2 frequency table given its margins' (11). If a full likelihood method were pursued, the CIs for the overall treatment effect would be expected to be even wider. Except when all the trials are small, some have advocated that the additional uncertainty would not be expected to have a great impact on the results and so pursuing a full likelihood approach is unnecessarily sophisticated for most practical purposes.[7]

Secondly, Biggerstaff and Tweedie (9) address the same problem by developing a variance estimator for Q, that leads to an interval estimation of τ^2, utilising an approximating distribution for Q. They also developed asymptotic likelihood methods for the same estimate. This information is then used to give a new method of calculating the weight given to the individual studies which takes into account variation in these point estimates of τ^2. In the given examples, these new weights are between the standard fixed and random effects in down-weighting the results of large studies and up-weighting those of small. (A past concern has been that when τ^2 is large the standard random effects model gives too much weight to the relatively small studies.)

These new weights will differ greatest from those of the standard random effects model, when the number of studies to be combined is small. 'If 20

[6] This method is implemented using S+ code. The paper also comments that for continuous scales, one could use a linear mixed model vie the SAS procedure PROC MIXED, this could be used for IPD (see chapter 27), if common variances are assumed. Senn has also shown how it can be implemented simply by the software package Mathcad (41) (produced by Mathsoft).

[7] The authors give an example in which the combined treatment effect CI goes from 0.37–0.95 in their approach to 0.37–0.97 in the full likelihood approach. See other developments below and (11) for details of the full likelihood approach.

or more studies are to be combined, then the weights should be similar to those in the standard random effects model.' (9)[8,9]

Bigerstaff (10), builds on the work of (9) in investigating interval estimates for τ^2. He compares, through simulation studies, the methods given in (9) with several new ones.

It should be noted that Bayesian methods exist which take into account this extra uncertainty (12) (see chapter 13).

Complete likelihood approach

Van Houwelingen proposed two goals in his paper (11), firstly to present a likelihood based approach to random effects which avoids use of approximating normal distribution and can be used when the assumptions of normality are violated.[10] Solutions are obtained via the EM algorithm (13). Secondly, he extends this method to a bivariate random effects model, in which the effects in both groups are supposed random. In this way, inference can be made about the relationship between improvement and baseline effect. This is a non-parametric procedure that is recommended by Hardy and Thompson (8) when the normality assumption is violated.

Using sample survey methods

Schmid *et al.* (14) mention a technique, based on the use of survey sampling methods. This uses a model assuming that a sample of observations has been taken from within each of a sample of studies, themselves chosen from a population of studies. The approach differs from the random effects model by not involving an explicit estimate of the subject or study variance. Instead, a robust estimate of the variance of the treatment effect is computed and is used to produce test statistics about those effects.

Methodology for non-independent studies

Emerson *et al.* (15) state that the DerSimonian and Laird random effects method (weighted non-iterative) inversely weights using the sum of the between-study variance and the conditional within study variance. They go on to reason:

> 'Because these weights are not independent of the risk differences, the procedure sometimes exhibits bias and unnatural behaviour.' (15)[11]

Their paper proposes a modified weighting scheme that uses unconditional within-study variance to avoid this source of bias. 'The modified procedure has variance closer to that available from weighting by ideal weights when such weights are known.' They also state: 'In combining studies, this procedure represents a compromise between an unweighted (equally weighted) mean and an *n*-weighted (sample-size weighted) mean; and it avoids the correlation between the risk differences and their weights.' (15)

Using trimmed means

Emerson *et al.* (16) present a trimmed versions of meta-analytic estimators for the risk difference. They incorporate this into a random effects model, and by doing so state that the model can resist the impact of a few anomalous studies. They compare four trimmed procedures [on different models including the one given above (15)] and found that a trimmed (20% most extreme data removed) DerSimonian and Laird (weighted non-iterative) method offers best performance over a wide range of simulation designs and sample. However, they conclude that none of the methods, whether trimmed or untrimmed, is uniformly preferable. It should be noted that this method ignores the information which may be given by the outliers, and removes any possibility of investigating why their results are so extreme.

Combining sibpair linkage studies

Li and Rao (17) published a paper which proposes a random effects model for combining results from independent quantitative sibpair linkage studies. This is an extension of the standard random effects methodology presented in this chapter. Weighted and empirical Bayes (EB) (see chapter 13) solutions are both presented. This is a first step into this area and the authors comment that more work is needed and report more research is being done on this topic.

[8] Software for implementing this method is given in (9).

[9] The authors comment that the approximating distribution for τ^2 has immediate application in EB methodology (see chapter 13).

[10] It is interesting to note that the requirement of normality in random effects meta-analysis is often brushed over and not investigated.

[11] The reader should be aware that the issue of lack of independence is only a problem in a very limited set of cases [see (15) for more details and chapter 26 on multiple effect sizes].

The normality assumption

Raudenbush and Bryk (18) describe techniques for assessing normality when the number of studies is reasonably large. It is difficult to assess whether the normality assumption as been violated with a small number of studies. Seltzer (19) developed robust estimation procedure that allows the analyst to assume that random effects are t rather than normally distributed.

Other methods

Other extensions of the methodology presented in this chapter are given in other chapters of this report, where they fit more naturally. For example, Berlin *et al.* (20) discusses dose–response models for fixed and random effects; these are dealt with in chapter 19. It is worth pointing out that Whitehead and Whitehead (21) presented a unified methodology for meta-analysis (general parametric approach), so the random and fixed effects models of chapter 9 and this chapter could be incorporated in one model. The interested reader is referred to the original paper for more information (21).

Comparison with fixed effects

Empirical evidence

At certain points throughout this chapter, comparisons between the fixed and random effects model have been made. Investigations into the differences in results produced by the two methods have been carried out.

Berlin *et al.* (22) compared the results of 22 meta-analyses by reanalysing them using both the Peto fixed effect method (pages 60–1) and the random effects model described of DerSimonian and Laird (pages 70–2). Eight of the studies showed evidence of heterogeneity, in three of these different conclusions would have been drawn about the treatment effect for both methods (23). In each of these three cases, the Peto method suggested a beneficial treatment effect while the DerSimonian and Laird method did not. In all the other studies, including ones showing no evidence of heterogeneity, both methods lead to the same conclusion.

Mengersen *et al.* (24) carried out a meta-analysis of the effect of passive smoking on lung cancer, and investigated how the results differed using different methods. They state that different conclusions may have been drawn if only fixed or random effect methods had been used.

Raudenbush (25) highlights the below advantages and disadvantages of random effects models.

Advantages of random effects:

1. Conceptualisation is consistent with standard specific aims of generalisation.
2. Allows a parsimonious summary of results when the number of studies is not very large.
3. Can use random effects model with no covariates as a baseline value for which the goodness of fit of regression models can be judged against. (i.e. can calibrate how much variation certain covariates explain).

Disadvantages/drawbacks of random effects:

1. Need to estimate sigma from the data (presuming one does not use the recently proposed methods; see pages 73–5).
2. Need to make the normality assumption (again assuming new methods are not being used; see page 75).

In addition, Greenland has pointed out (26) that random effects models are more sensitive to publication bias. The reason for this is as follows. As previously reported, in a random effects analysis large studies will be downweighted and small ones given increased weight. So, any tendency not to publish small statistically non-significant studies will lead to a greater proportion of spuriously strong associations among small published studies than among large published studies. 'Thus, by giving more weight to small studies, the random effects summary will give more weight to spuriously strong associations and so produce a more biased summary estimate if publication bias is present.' (26)

To summarise, random effects will always give a CI that is at least as large, and usually larger than a fixed effects model because it allows for variation between studies. The greater the degree of heterogeneity the greater the difference in the CIs will be.

When should random effects models (rather than fixed effect models) be used? Researchers' opinions

There seems no simple answer to this question. Several authors suggest guidelines to the use of fixed and random effect models, most of whom also acknowledge widely differing points of view exist between practitioners in the field. Shadish and Haddock (3) consider the answer to be partly statistical, partly conceptual and rarely indisputable. However, many believe that if there exists evidence of heterogeneity, that cannot be explained (using the techniques of chapter 8), this extra variation needs to be accounted for when estimating the pooled estimate and CI (2), and that the fixed effect methods will give and over confident result.

Also, it is recognised that the heterogeneity test lacks power (chapter 8), so the chance of a type two statistical error is quite large, suggesting the studies are homogeneous when in fact there is a degree of heterogeneity. This implies that just because the studies appear homogeneous a random effects model may still be worth considering as it cannot be assumed that true homogeneity exists (1). It is worth noting at this point that there must become a time when the heterogeneity between studies is so large that the random effects model is not adequate, and the question of whether the results should be combined at all has to be addressed (for further discussion, see chapter 8).

Another corollary question needing to be addressed is: should a decision be made *a priori* as to which modelling strategy is to be adopted?

It should be noted that both the results of fixed and random effects can be reported, this is justified by viewing it as a form of sensitivity analysis. If the two methods differ, one can conclude heterogeneity must be a problem and stress the random effects estimate, or go on to investigate possible causes of heterogeneity.

Hasselblad and McCroy (5) comment: 'there are those who would argue that the unexplained variation must be explained before any conclusions can be drawn. Others argue that the only appropriate model is the hierarchical one because Mother Nature is never consistent across studies.'

Raudenbush (25) suggests the choice may depend in part on the number of studies available. He reasons that if only a few studies exist (for an extreme he says two), between study variation will be very poorly estimated and thus fixed effects will be the sensible choice. If more were available (say several hundred), the fixed effects approach would make little sense because the treatment by studies interaction test would have great power, virtually ensuring rejection of the null hypothesis (see original for a more thorough explanation).

Comments on random versus fixed effects

Below (in no particular order) are comments from leading researchers and practitioners of meta-analysis on their beliefs about the applicability of fixed and random effect models:

Thompson:

> 'any set of studies is inevitably clinically heterogeneous by virtue of differences in study design, patient selection , or treatment policy.' (1)

and (slightly edited):

> 'However the random-effects method is no panacea for heterogeneity. Formal interpretation relies on the peculiar premise that the trials done are representative of some hypothetical population of trials, and on the unrealistic assumption that the heterogeneity between studies can be represented by a single variance, and that the between trial distribution is normal. Moreover, for the interpretation of the overall θ as applying to future trials or patients there is the necessary but intangible assumption that the trials included in the meta-analysis are "representative" of the future The results are also often strongly dependent on the inclusion or exclusion of small trials, which may themselves reflect publication bias. The random effects methods may therefore give undue weight to small studies, emphasising poor evidence at the expense of good. (An additional technical consideration is that the estimate of σ^2, being made usually from relatively few trials, is very imprecise. Given these problems, one can only view the random effects analysis as replacing the implausible assumption of the fixed effect analysis by untenable assumptions of its own.)' (1)

Peto (of random effects):

> 'I think that this is actually wholly wrong as an approach to the overviews and trials. I think that it does answer a question. But it's a very abstruse and uninteresting question. It's trying to say "what would happen if we chose another treatment at random from the universe of treatments that we could choose another population at random from the universe of populations". I think this is not an important question.' (27)

Meier formally disagreed with Peto above at the conference and put a case for random effects. 'inter-study variation is a key feature of the data and should contribute to the analysis' (28)

Thompson:

> '....... the assumption that the true treatment effects are the same for all the trials, that is an assumption of homogeneity. In any meta-analysis this is a simplistic and implausible assumption.' (29)

also: (talking of fixed effect analysis) '.. the derived confidence interval for the overall odds ratio is too narrow in terms of extrapolation to future trials or future patients.' (29)

also: 'An intuitively appealing aspect of the random effects analysis is that, by taking into account a component of between-trial variability, it appropriately introduces a degree of statistical caution that is not present in the fixed effect analysis.' (29)

and: 'A more useful way to consider the random effects method is as a type of sensitivity analysis, to investigate how much the overall conclusions change as the assumptions underlying the statistical methodology also change. (In fact the random effects analysis can be viewed as simply changing the percentage of weight allocated to each trial, as compared with the fixed effect analysis.)' [quoted from (30)].

Fleiss gives an example of the problem of heterogeneity:

> '.. if in one meta-analysis there are two published studies with ORs of and 6.0, if in another there are two published studies with ORs of 2.0 and 3.0, and if all four values of V (the variance of the logarithm of the OR) are equal to 0.01, then in both studies the value of the pooled OR will be 2.45 and in both studies the approximate 95% confidence intervals extend from 2.13 to 2.81.' (31)

Fleiss:

> 'Bailey (32) suggests that, when the research question concerns whether the treatment will have an effect, on the average, or whether exposure to a hypothesized risk factor will cause disease, on the average, then the model of studies being random is the appropriate one. When the question concerns whether treatment has produced an effect, on the average, or whether exposure has caused disease, on the average, in the studies in hand, then the model of studies being fixed is the appropriate one.' (31)

And in summary (32): 'The choice between these fixed effect methods would rarely materially affect the conclusions being drawn.' (31)

Pladevall-Vila, in an investigation of conflicting meta-analyses, investigating the relationship between oral contraceptive use and rheumatoid arthritis, conclude by saying:

> 'Consensus is needed on how to conduct meta-analyses of observational studies, the methods to be used in the presence of heterogeneity, and when conclusions should be considered reliable.' (33)

Greenland:

> 'In situations in which addition of a random effect to the model yields materially important changes in inferences, the degree of heterogeneity present will often (if not usually) be so large as to nullify the value of the summary estimates (with or without the random effect). Such a situation is indicative of the need to further explore sources of conflict among the study results.' (34)

Greenland:

> 'I maintain that the primary value of a meta-analysis is in the search for predictors of between-study heterogeneity. If use of random effects makes a difference, the analysis is incomplete: the analyst should carefully search for the source of the discrepancy between the fixed- and random-effects interval estimates. The random-effects summary is merely a last resort, to be used only if one cannot identify the predictors or causes of the between-study heterogeneity.' (26)

The 1992 National Research Council (35) report on statistical issues in combining information favours random effects models for meta-analysis.

Pocock:

> 'A sensible overall conclusion is that neither the fixed effect nor the random effects model can be trusted to give a wholly informative summary of the data when heterogeneity is present. Perhaps the presentation of both approaches reveals the inevitable uncertainty inherent in an overview with heterogeneity. Indeed any strong claims by proponents of one method over the other are liable to be counterproductive in that polarized statistical disputes may discourage the medical profession from accepting overviews' (36)

also: 'the difference between the two models is sometimes over-emphasized.' (36)

Further research

A study to compare the methods of Hardy and Thompson (8), and Biggerstaff and Tweedie (9,10) must be pertinent (the two new methods for random effects incorporating more uncertainty). This would continue the work of Smith *et al.* (37), who compared the results obtained by many of the previous meta-analytic models.

Investigation into how robust are random effects to departures in normality? Should likelihood methods (11) be employed more often?

More work is required on combining sibpair linkage studies.

Summary

At this present time, it would seem neither fixed nor random effect models could be considered the ideal analysis, beyond any dispute, for a given situation. Indeed, it has been illustrated that both methods have their shortcomings. As the point estimates of effect size given by both methods are usually very similar, the only time the choice of model will be critical is if its significance is marginal using a fixed effect model. Here there is a chance that the more conservative CI given by the random effects approach would consider the effect to be non significant. It is interesting to note that Peto, one of the strongest opponents of random effects models, takes 3 standard deviations rather than 2 (1% not 5%) as his critical value when considering the significance of an (fixed) effect in an overview, considering 2 standard deviations to be not stringent enough for the magnitude of the implications of an overview. ['.........we are messing around if we take two standard deviations, two-and-a-half standard deviations, as serious evidence. We get so much nonsense mixed up in with the sense that it is just irresponsible. I think we've got to get better standards of evidence than we normally have, and this means in the individual trials and in overviews. I think you need to go to at least three standard deviations.' (27)]. The point in mentioning this is that one of the world leaders in the field, although conceptually at poles with the advocators for random effects, through this more stringent cut point is actually making an adjustment with practical implications very similar to those inherent by the use of a random effects model. While it would appear that the conceptual debate over the correct model is some way off a conclusion, a practical line to take may be to say: use whichever strategy (single analysis or several) you yourself feel is most appropriate for the situation. However, if there is evidence of heterogeneity (significant or not) and a fixed effect analysis is the sole analysis carried out and the result is only marginally significant (5% level), then extreme caution is needed when reporting and interpreting the results. Another key point to consider here relates to the clinical significance rather than the statistical significance of the pooled estimate obtained. One should be concerned about estimates and their SEs, rather than *p*-values. It should be pointed out that other models do exist for meta-analysis, chapter 12 covers mixed models, and chapter 13 Bayesian models. It is interesting that the National Research Council (35) take the approach of calling random (and fixed) effects models a special case within a hierarchical model framework, of which other models [such as mixed and cross-design synthesis (chapter 26)] are simply extensions. Another point worthy of note is when using a Bayesian approach, one does not necessarily have to choose between the two models (fixed and random), but rather we can average across models using BFs (see chapter 13).

References

1. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet* 1991;**338**:1127–30.
2. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177–88.
3. Shadish WR, Haddock CK, Cooper H, Hedges LV, editors. Combining estimates of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 261–84.
4. Fleiss JL. The statistical basis of meta-analysis (review). *Stat Methods Med Res* 1993;**2**:121–45.
5. Hasselblad VIC, Mccrory DC. Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making* 1995;**15**:81–96.
6. Hedges L. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Stat* 1981;**6**:107–28.
7. Sankey SS, Weissfeld LA, Fine MJ, Kapoor W. An assessment of the use of the continuity correction for sparse data in metaanalysis. *Commun Statist Simulation Computation* 1996;**25**:1031–56.
8. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med* 1996;**15**:619–29.
9. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med* 1997;**16**:753–68.
10. Biggerstaff BJ. Confidence intervals in the one-way random effects model for meta-analytic applications. University of Colorado: Technical Report, 1996.
11. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Stat Med* 1993;**12**:2273–84.
12. Louis TA, Zelterman D, Cooper H, Hedges LV, editors. Bayesian approaches to research synthesis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 411–22.
13. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 1977;**39**:1–38.
14. Schmid JE, Koch GG, LaVange LM. An overview of statistical issues and methods of meta-analysis. *J Biopharm Stat* 1991;**1**:103–20.
15. Emerson JD, Hoaglin DC, Mosteller F. A modified random-effect procedure for combining risk difference in sets of 2x2 tables from clinical trials. *J Ital Statist Soc* 1993;**2**:269–90.
16. Emerson JD, Hoaglin DC, Mosteller F. Simple robust procedures for combining risk differences in sets of 2x2 tables. *Stat Med* 1996;**15**:1465–88.
17. LI ZH, Rao DC. Random effects model for metaanalysis of multiple quantitative sibpair linkage studies. *Genet Epidemiol* 1996;**13**:377–83.

18. Raudenbush SW, Bryk AS. Examining correlates of diversity. *J Educ Statist* 1987;**12**:241–69.

19. Seltzer M. The use of data augmentation in fitting hierarchical models to education data. University of Chicago: unpublished doctoral dissertation, 1999.

20. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. *Epidemiology* 1993;**4**:218–28.

21. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. *Stat Med* 1991;**10**:1665–77.

22. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Stat Med* 1989;**8**:141–51.

23. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science (review). *Epidemiol Rev* 1992;**14**:154–76.

24. Mengersen KL, Tweedie RL, Biggerstaff BJ. The impact of method choice in meta-analysis. *Aust J Stats* 1995;**37**:19–44.

25. Raudenbush SW, Cooper H, Hedges LV, editors. Random effects models. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 301–22.

26. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;**140**:290–6.

27. Peto R. Why do we need systematic overviews of randomised trials? *Stat Med* 1987;**6**:233–40.

28. Meier P. Proceedings of methodologic issues in overviews of randomized clinical-trials – commentary. *Stat Med* 1987;**6**:329–31.

29. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction (review). *Stat Methods Med Res* 1993;**2**:173–92.

30. Spector TD, Thompson SG. Research methods in epidemiology. 5. The potential and limitations of meta-analysis. *J Epidemiol Comm Health* 1991;**45**:89–92.

31. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127–39.

32. Bailey KR. Inter-study differences – how should they influence the interpretation and analysis of results. *Stat Med* 1987;**6**:351–60.

33. PladevallVila M, Delclos GL, Varas C, Guyer H, BruguesTarradellas J, AngladaArisa A. Controversy of oral contraceptives and risk of rheumatoid arthritis: meta-analysis of conflicting studies and review of conflicting meta-analyses with special emphasis on analysis of heterogeneity. *Am J Epidemiol* 1996;**144**:1–14.

34. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.

35. National Research Council. Combining information: statistical issues and opportunities for research. Washington DC: National Academy Press, 1992.

36. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med* 1990;**9**:657–71.

37. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* 1995;**14**:2685–99.

38. Hedges LV, Olkin I. Statistical methods for meta-analysis. London: Academic Press, 1985.

39. Rao PS, Kaplan J, Cochran WG. Estimators for the one-way random effects model with unequal variances. *J Am Stat Assoc* 1981;**76**:89–97.

40. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc* 1977;**72**:320–38.

41. Senn S. Meta-analysis with Mathcad. *ISCB News* 1996;**20**:4–5.

Chapter 11

Meta-regression

Introduction

Occasionally, the studies whose effect estimates are to be combined may all be very similar. This may be the case, for instance, if results are being combined from multi-centre trials, all using the same protocol. It is more common though, for there to be substantial differences between the studies. Examples of ways studies may differ include; treatment dose magnitude; age of study population; study conduct; and study maturity (1) (see pages 41–3 for a detailed account of how studies may vary). These differences may contribute to heterogeneity of the results between studies. Chapter 8 addressed the issue of heterogeneity and explained how to investigate and deal with it if it is present. When heterogeneity is present it does need investigating, but does not have to be necessarily seen as a burden. Discovering why study results differ can be revealing.

It has also been pointed out that due to the large numbers of patients often involved in a meta-analysis, the difficulties of detecting therapeutic effects within subsets of patients observed with limited data from single studies may be overcome (1). In doing this, treatments could be individualised, so the treatment best for each patient could be identified (1). Thus exploratory analysis investigating associations between study or patient characteristics and the outcome measure (particularly useful in observational studies – see chapter 19), can be seen as one of the advantages of performing a meta analysis (2). So, as well as reducing the heterogeneity, this analysis may produce findings of clinical importance.

It needs to be stressed that this is an exploratory analysis and it is very possible for associations between characteristics and the outcome to occur purely by chance (this problem is not unique to meta-analysis and occurs whenever associations between variables are being investigated). Also, spurious associations may appear due to confounding factors this is explained fully on pages 149–52.

A statistical technique capable of carrying out the sort of analysis described above is regression. Two different underling models are presented for this analysis. One is described in this section and is based on combining studies using a fixed effect model (chapter 9) and has come to be called meta-regression. The second model described in the next chapter (chapter 12) uses the random effects model of chapter 10 as its basis. To distinguish this model from the first it is referred to as a mixed model due to it including random and fixed effects (though it is still a regression type model).

The fixed-effect methods of this chapter include no random variation term and are thus appropriate only when all variation between study outcomes can be considered fixed, predictable and accountable. A mixed model is appropriate when the predictive variables only explain part of the variation/heterogeneity. The random term thus is included to take account for this extra unexplained variation. However, one will not know which model is most appropriate until the amount of variation explained by the predictor variables has been established. For this reason, it is customary to start with a meta-regression model with no random effect term, and include one only if considered necessary, i.e. after the best model is found substantial residual variation remains [this could be tested formally using the *Q* statistic (pages 39–41)].

Modelling using regression models is not a trivial task. It is beyond the scope of this chapter to give a comprehensive beginners guide to regression techniques. For the reader who wants to know more about regression modelling many introductory statistical texts cover the basics, additionally see (3) for further details on modelling binary outcomes.

Model notation

The following account is adapted from Hedges (4); it is the most general meta-regression model, and can include continuous and discrete predictor variables. If only a limited number of categorical predictor variables are being investigated, an ANOVA approach can be taken. Hedges (4) clearly describes this approach; however, it is omitted here due to space limitations, and because it can be regarded as a special case of the more general model below.

Suppose there are k independent effect size estimates $T_1, \ldots, T_k$ with estimated sampling variances $v_1, \ldots, v_k$ [this is the same notation as used for fixed effect model (pages 55–6)].

The corresponding underlying effect size parameters are $\theta_1, \ldots, \theta_k$, for each of the k studies. Suppose also that there are p known predictor variables $X_1, \ldots, X_p$ which are believed to be related to the effects via a linear model of the form:

$$\theta_i = \beta_0 + \beta_1 \mathrm{x}_{i1} + \ldots + \beta_p x_{ip} \qquad (11.1)$$

where $x_{i1}, \ldots, x_{ip}$ are the values of the predictor variables $X_1, \ldots, X_p$ for the ith study and $\beta_0, \beta_1, \ldots, \beta_p$ are the unknown regression coefficients, to be estimated, indicating the relationship between its associated predictor variable and the outcome.

Recall in the fixed effects model of chapter 9, $\theta_1, \ldots, \theta_k$, were all set equal, say to θ. Here they are allowed to vary (as in the random effects analysis: chapter 10). However, unlike the random effects model, here it is the covariate predictor variables that are responsible for the variation not a random effect, hence the variation is predictable not random.[1]

An alternative to the above derivation, when one has binary outcomes, is to use logistic regression. An example of its use is given by Thompson (5), and is very similar to the above model to implement. This example is particularly noteworthy as it looks at cholesterol trials (though a different set to those considered in this report) and the effect of covariates such as the extent and duration of cholesterol reduction (see also pages 44–8, 157–61).

The application of the above model

The coefficients in the above model are easily calculated via weighted least squares algorithms. (Unweighted regression cannot be used because this would make the assumption that the variances from each study could be considered equal.) Any standard statistical package that performs weighted (multiple) regression can be used.

As Hedges states (4), the regression should be run with the effect estimates as the dependent variable and the predictor variables as independent variables with weights defined by the reciprocal of the sampling variances. That is, the weight for T_i is

$$w_i = 1/v_i$$

The predictor variables are created/defined by the researcher: these can take several forms, including 1) binary indicators, e.g. indicating whether the study adjusted for smoking, study population was European etc., 2) categorical, e.g. variable could indicate the type of study design and 3) continuous, e.g. level of exposure (in epidemiological studies) or mean age of the patients recruited.

It is important to note that the SEs of the estimates for the coefficients, produced by standard software packages are based on a slightly different model than the above used for fixed effect meta-regression. This means that the weighting is ignored in the calculation of the SE. Due to this an adjustment needs to be calculated by hand:

$$S_j = \mathrm{SE}_j \Big/ \sqrt{\mathrm{MS}_{ERROR}} \qquad (11.2)$$

where S_j is the corrected SE, SE_j is the SE of coefficient b_j (the obtained estimate for $_j$) as given by the computer programme and MS_{ERROR} is the ‘error’ or ‘residual’ mean square from the analysis of variance for the regression as given by the computer programme.

Each of the regression coefficient estimates (the b_js) are normally distributed about their respective parameter (β_j) values with standard deviations given be the SEs (the S_js). Hence a (100 – α)% CI for each β_j is calculated by:

$$b_j - Z_{\alpha/2}(S_j) \le b_j \le b_j + Z_{\alpha/2}(S_j) \qquad (11.3)$$

where $Z_{\alpha/2}$ is the is the two-tailed critical value of the standard normal distribution. The corresponding two-sided significance test is H_0: $\beta_j = 0$, and is rejected if the above CI contains one. If it is retained, one concludes there is no, or insufficient evidence of a relationship between the jth predictor variable and outcome.

In this way, decisions can be made on which, if any of the predictor variables explain the variation between studies and hence appear to be good predictors. Selecting which of the different predictors are entered and removed from the model and deciding on the ‘best’ model can be a long and complex process, with much being published on

[1] This model assumes approximate normality of the dependent (outcome) variable (7). In situations when the outcomes are in the form of 2 × 2 tables, Greenland reports (7) that simulation studies indicate that such a criterion will be adequately met if the expectations of the counts contributing to the rates or ratios is four or greater.

the topic. Factors that contribute to this decision are: 1) the amount of variation that is explained and 2) the simplicity and ease of interpretation of the model.

As with all modelling exercises, testing of assumptions and considering the adequacy of model fit is an important aspect of the analysis that should not be overlooked.

Advanced model fitting issues

The sections below outline other issues that are pertinent while carrying out a meta-regression.

Testing blocks of variables simultaneously

Hedges (4) outlines a method for testing hypotheses about groups or blocks of regression coefficients (as opposed to individually). Hedges suggests there are situations where it may be desirable to enter a block of variables reflecting methodological characteristics. Then entering another block, say reflecting treatment characteristics to see if the second block of variables explained any of the variation in effect size not accounted for by the first block of variables [see (4), p. 296] for computational details).

Colinearity

This is a problem that can occur in any multiple regression analysis, not just meta-analysis. It basically means that two or more predictor variables are explaining the same variation and are thus correlated. Hedges warns: 'Colinearity may degrade the quality of estimates of regression coefficients, wildly influencing their values and increasing their standard errors.' (4). The reader is again referred to a standard regression textbook for procedures used to safeguard against this.

The application of meta-regression

Situations where the use of meta-regression is applicable

Meta-regression can and has been used in a wide diversity of situations. It can be used both for the synthesis of RCTs and observational studies. Several meta-regression techniques specific to observational studies exist, such as dose–response analysis, these are covered on pages 157–61. If IPD are available, a more highly structured model may be more appropriate (6); regression using patient level (as opposed to study level) covariates is possible, and this is covered in chapter 24. Meta-regression can be used to incorporate study quality (e.g. via a quality score covariate). How study quality is measured is a complex issue; chapter 6 is dedicated to this issue. Meta-regression can be employed as a sensitivity analysis; the sensitivity of inferences to variations in or violations of certain assumptions can be investigated (7). Greenland illustrates this with the following example:

> 'One may have externally controlled for cigarette smoking in all studies that failed to control for smoking by subtracting a bias correction from the unadjusted coefficients in those studies. The sensitivity of inferences to the assumptions about the bias produced by failure to control for smoking can be checked by repeating the meta-analysis using other plausible values of the bias, or by varying the correction across studies.' (7)

Variables that can be included in a meta-regression

Chapter 8 highlighted many ways in which studies can differ. All these factors (and any others the researcher can identify) can be explored using meta-regression. Dickersin and Berlin (2) in their 1992 review of meta-analysis included several examples where meta-regression had been used to explain heterogeneity and find treatments that effected subsets of patients differently.

Problems with meta-regression

A couple of problems inherent when carrying out meta-regression of epidemiological studies have been pointed out by Greenland (7). The first of these he calls aggregation bias or ecological bias. This bias will exist if the relation between group rates or means do not resemble the relation between individual values of exposure outcome (7). Secondly, he notes that further bias can arise from regressing adjusted study results on unadjusted average values for covariates (7). He notes that such bias will, however, be small unless the covariates under study are strongly associated with the adjustment factors.

Another potential problem is that some of the studies may not have the same covariate information as the rest. If this is the case, possible solutions are either to contact the original authors of the reports to try and obtain the necessary variables, or to carry out a subset analysis to see if the variable seems important in the studies that do measure it. Problems also exist due to data missing at the patient level, as this will affect aggregated study level variables; see chapter 17 for further details on missing data.

New developments

Modelling duration of trial

In an investigation of the effect of a reduction in cholesterol levels on overall mortality, Thompson (5) used a non-standard method to investigate the effect of duration of the studies. Due to the suspicion of a non linear relationship, i.e. longer follow-up does not necessarily mean larger treatment effect, a standard dose–response type model (see pages 157–61) could not be used. Instead data in the time intervals 0–2.0, 2.1–5.0, 5.1–8, 8.1–12.0 years was obtained from the original investigators for most of the studies. This sort of data can be viewed as between that of overall study estimate and individual patient level (see chapter 23) and since IPD was unavailable, was the best that could be obtained.

Further research

Whilst the use of meta-regression can be a powerful tool to the meta-analyst and should be recommended there are a number of issues that are wanting of further work:

1. Checking of modelling assumptions, including the use of residuals.
2. Dealing with (and accounting for) missing data, both at the study level and patient-level. Indeed, this is a recurring issue throughout this report.
3. Measurement error – this is particularly true when considering study-level covariates. For example, a common study level covariate is age, but unless there are details on, for example, the number of patients within a trial for whom age was not recorded, the use of average age may lead to biased results.
4. The modelling of data when there are some studies with only summary statistics available (i.e. study level covariates) and other studies for which patient level data is available.
5. Model comparison; as with other modelling scenarios, a choice is often made between competing models. How this choice is made can sometimes have a profound effect on the overall conclusions. The implications of the effect of different model selection strategies is an important area which deserves more attention.

Summary

This chapter has extended the methods of chapter 9 (fixed effects) to take account of the fact that there are often covariates at either the study level or patient level available, and that these can be important in helping to explain any heterogeneity present. Such an analysis should be seen as a fundamental component of any meta-analysis, but as with any modelling exercise, due care and attention should be paid to the verification of any assumption the models make. One of the potential advantages of this approach is that estimates of the relative benefits of treatments for patients with different combinations of covariates can be derived, or more information on the relative effect of different forms of delivering the intervention. This is the sort of data that is very relevant to clinical practice, where overall average effects may be too general to be useful for particular situations.

References

1. Gelber RD, Goldhirsch A. The evaluation of subsets in meta-analysis. *Stat Med* 1987;**6**:371–88.
2. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science (review). *Epidemiol Rev* 1992;**14**:154–76.
3. Collett D. Modelling binary data. London: Chapman & Hall, 1991.
4. Hedges LV, Cooper H, Hedges LV, editors. Fixed effects models. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 285–300.
5. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction (review). *Stat Methods Med Res* 1993;**2**:173–92.
6. Boissel JP, Blanchard J, Panak E, Peyrieux JC, Sacks H (1989) Considerations for the meta-analysis of randomized clinical trials: summary of a panel discussion. *Controlled Clin Trials* year?;**10**:254–81.
7. Greenland S (1987) Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* year?;**9**:1–30.

Chapter 12

Mixed models (random effects regression)

Introduction

Hedges in 1987 mentioned, at a conference on meta-analysis, the possibility of mixed effect models as a compromise between fixed and random effects (1). He commented that they have been used with success in the social sciences:

> 'Mixed effect models are often very close to fixed effect models in the sense that there is often only a very small component of random variation between studies, but it may be a persistent and very real source of variation that must be modelled.' (1)

Chapter 11 outlined fixed effect regression analysis where all heterogeneity between studies is considered to be explained by covariates included in the model. If the covariates do not explain all the variation, to obtain a more realistic CI around the point estimate, a random effect term needs including to take into account the variation unaccounted for. It is worth reiterating that, although the aim of including covariates is to reduce variation between studies – the covariates themselves should not be considered nuisance factors and indeed may shed light on the generalisability of the treatment under investigation and suggest possible subsets of patients for whom the treatment is more or less effective, in a way not possible in the analysis of subsets of a single study. Indeed such an analysis may also inform the direction of further research.

Having fitted a regression model, if the residual heterogeneity is still significant, a random effects term should clearly be added. If the residual heterogeneity is not significant many researchers still consider it good practice to always include a random term to account for any variation not accounted for.

Rubin conceptually expanded the ideas of the types of models covered in this chapter. His method for extrapolating response surfaces (2) is covered on pages 214–15.

Mixed effect model

Notation

The derivation below is taken (but modified) from Raudenbush (3). Raudenbush also uses an ANOVA model for mixed models; however, to use this one would need a balanced design, which would be very rare in health technology research and thus is omitted (3).

As a starting point, take the random effects model outlined on pages 70–2, i.e.

$$T_i = \theta_i + e_i \qquad (12.1)$$

where T_i is the estimated effect size of the true effect size θ_i for each of the k studies, $i = 1, \ldots, k$; it is also assumed that the e_i are statistically independent, each with a mean of zero and estimation variance v_i.

The variance for these estimates of treatment effect can be expressed as:

$$\text{Var}(T_i) = v_i^* = \sigma_\theta^2 + v_i \qquad (12.2)$$

where σ_θ^2 is the between-study, or random effects variance and v_i is the within-study variance.

Now we extend this to formulate a prediction model for the true effects as depending on a set of study characteristics plus error:

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + u_i \qquad (12.3)$$

where β_0 is the model intercept; $X_{i1}, \ldots, X_{ip}$ are coded characteristics of studies hypothesised to predict the study effect size; $\beta_1, \ldots, \beta_p$ are regression coefficients capturing the association between study characteristics and effect sizes; u_i is the random effect of study i, that is, the deviation of study i's true effect size from the value predicted on the basis of the model. Each random effect, u_i, is assumed independent, with a mean of zero and variance σ_θ^2.

Under the fixed effects specification, the study characteristics $X_{i1}, \ldots, X_{ip}$ are presumed to account completely for variation in the true effect sizes. In contrast, the random effects specification assumes that part of the variability in these true effects is unexplainable by the model.[1]

It is interesting to note that this model is a consistent extension of the models presented in previous chapters. If the model has no predictors, i.e. $\beta_1 = \ldots = \beta_p = 0$, then it reduces to that of the random effects of chapter 10. If the random effects variance is null i.e. $\sigma^2_\theta = 0$, then the results will be identical to that of the fixed effects meta-regression model of chapter 11.

Estimating the parameters

Substituting (12.1) into (12.3) gives:

$$T_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + u_i + e_i \quad (12.4)$$

This equation has two components in its error term $u_i + e_i$, so the variance of T_i, controlling for the *X*s, is

$$v_i^* = \mathrm{Var}(u_i + e_i) = \sigma^2_\theta + v_i \quad (12.5)$$

Ordinary least squares regression assumes that every residual has the same variance (homoscedasticity). This assumption will be violated if the v_i^* vary across studies (which they undoubtedly will). A weighted least squares approach is needed instead (as was used in chapter 11), optimal weights are given by the inverse of each study's variance:

$$w_i^* = 1/v_i^* = 1/(\sigma^2_\theta + v_i) \quad (12.6)$$

We can estimate the v_is from the data (see chapter 9). We also need an estimate of σ^2_θ, which is generally unknown and must be estimated from the data. In fact, an estimate of the regression coefficients (the βs) is required in order to obtain estimates of σ^2_θ and hence w_i^*. So unfortunately, a dilemma exists: estimation of the βs is dependent on knowing σ^2_θ, and estimation of σ^2_θ depends on knowing the βs.

Solutions to the model

Two different approaches to the problem outlined above have been put forward:

The method of moments

Raudenbush reports:

> 'Using the method of moments, the researcher computes provisional estimates of the β's in equation (12.4). Based on these estimates, an estimate σ^2_θ can be obtained and, therefore, the weights, w_i^*. These weights are then employed in a weighted least squares regression to obtain new and (final) estimates of the β's.' (3)[2]

These provisional estimates can be got from ordinary regression or weighted regression as in chapter 11. For explicit details of this procedure see (3), p. 310.

The method of ML

To implement this approach, a further assumption that each T_i is normally distributed is required. Raudenbush (3) reports that MLEs have certain desirable properties: in large samples they are efficient and normally distributed with known SEs, facilitating statistical inference.[3,4]

Obtaining estimates

The authors of this report are not aware of any investigations into the superiority of either method, and hence cannot make recommendations about which of the two methods to use.

Regardless of which method is used, however, the techniques available for any regression analysis such as: assessing fit, comparing models, adding/removing terms can be applied. The reader is referred to the previous chapter and to the regression techniques literature for further details. It is also possible to test whether, $\sigma^2_\theta = 0$, by fitting a model with and without the random variation term, details are given in (3), p. 315.

The majority of the time a normal assumption is made for calculating CIs of the model parameters. However, Larholt suggests the use of the *t* distribution for small samples (4).

Extensions/alternatives to the model

On pages 85–6, a basic mixed effects model was described. Several extensions/alternatives to this model have been derived, and applied in meta-analyses. This section presents a summary of these models. In addition to those presented below,

[1] Huque and Dubey (5) note that if the linear structure of this model is not acceptable, then an appropriate non-linear structure may be considered.

[2] Raudenbush (3) provides a computer program to implement this method.

[3] Raudenbush (3) provides a computer program to implement this method.

[4] For more details on ML based solutions, full details are given in (5).

Huque and Dubey (5) provide a formulation and estimate parameters via the Fisher information matrix.

An alternative random-effects regression model

Berkey *et al.* (6) derive an iterative random effects regression model specifically for the synthesis of 2 × 2 tables. The solution to the model is based on an iterative scheme which alternates between estimating the regression coefficients via weighted least squares, where the weights incorporate the current estimate of the between study variance, and estimating the between-study variance. The authors comment that this model is compatible with that of DerSimonian and Laird (7) (chapter 10) model but as the DerSimonian and Laird model becomes more difficult to evaluate, when considering a continuous covariate or 2 or more categorical covariates simultaneously then this model is an 'efficacious' alternative. A Statistical Analysis System (SAS) program for implementing this procedure is given in the paper (6).

The authors went on to apply this model to evaluate the efficacy of the BCG vaccine for preventing tuberculosis. One of the variables that reduced heterogeneity was the number of miles from the equator the site of the study was. They comment that small biases were present in the estimates of the regression coefficients and the between study variance, and that there is the potential to eliminate these using an alternative estimator for σ_i^2.

An additional noteworthy point is that they use a smoothed estimator of the within-study variances, which produced less bias in the estimated regression coefficients. The authors comment that Emmerson *et al.* (8) demonstrated that because each study's estimate of risk difference and the corresponding estimated variance (s_i^2) are not independent the DerSimonian and Laird random effects approach (see chapter 10) may produce a biased estimate of overall treatment efficacy. (Note: this is nothing to do specifically with mixed effect regression models; however, this model gets round the problem.) Due to this, there exists a correlation between $\log_e(\text{RR}_i)$ and $\hat{\text{var}}[\log_e(\text{RR}_i)]$, which leads to slight bias towards the null. Therefore an alternative estimator of the variance of $\log_e(\text{RR}_i)$ is given. This smoothed estimator reduces the correlation:

$$\hat{\text{var}}[\log_e(\text{RR}_i)] = \left[\sum_{i=1}^{k}(b_i/a_i)\right]\Big/kn_{i+} + \left[\sum_{i=1}^{k}(d_i/c_i)\right]\Big/kn_{i-} \quad (12.7)$$

where a_i, b_i, c_i and d_i are the values in the cells of the 2 × 2 table for the ith study, and $n_{i+} = a_i + b_i$ and $n_{i-} = c_i + d_i$. In the same vein, an adjusted variance for the $\log_e(\text{OR}_i)$ is also given:

$$\begin{aligned} \hat{\text{var}}[\log_e(\text{OR}_i)] = {} & \left[(a_i + c_i)\left(\sum_{i=1}^{k}\left(a_i/(a_i + c_i)\right)\right)\Big/k\right]^{-1} \\ & + \left[(a_i + c_i)\left(1 - \left(\sum_{i=1}^{k}\left(a_i/(a_i + c_i)\right)\right)\Big/k\right)\right]^{-1} \\ & + \left[(b_i + d_i)\left(\sum_{i=1}^{k}\left(b_i/(b_i + d_i)\right)\right)\Big/k\right]^{-1} \\ & + \left[(b_i + d_i)\left(1 - \left(\sum_{i=1}^{k}\left(b_i/(b_i + d_i)\right)\right)\Big/k\right)\right]^{-1} \end{aligned} \quad (12.8)$$

Model for adjusting bias when a covariate is an aggregate measurement of the treated population

McIntosh (9) discusses cases in which including the observed control group event rate appears to reduce heterogeneity. The author warns in these circumstances that the association, or some part of it, may simply arise as a consequence of measurement error (sometimes known as regression to the mean). A model is presented that corrects for the correlated measurement error peculiar to this application. The model is hierarchical in structure and both Bayesian (see chapter 13) and ML solutions are given. The author concludes that this method is appropriate whenever a covariate of interest is an aggregate measurement of the treated population. See pages 46–8 for more details on this topic.

General model form

Recently, Stram (10) presented a very general mixed-effects regression model framework. He developed a model from which most of the previous models can be viewed as special cases. So, this model incorporates the random effects model (11) (chapter 10), the mixed model (pages 85–6), the model of Begg and Polite (12) (see pages 201–3) and the model of Tori *et al.* (see page 214) (13). After presenting the general form of the model, the author goes on to describe its relationship to these models.

Model form:

$$Y_i = X_{i\alpha} + Z_i\beta_i + \zeta_i + e_i \quad (12.9)$$

where $i = 1,2, \ldots, K$ independent studies. Y_i is an $(n_i \times 1)$ vector of one or more related estimates of treatments or treatment comparisons of interest;

X_i is an $(n_i \times p)$ matrix of known covariates related to the p vector of unknown fixed effect parameters, α; and Z_i is an $(n_i \times q)$ vector of known covariates related to a $(q \times 1)$ vector of unobserved random effects, β_i, for each study. The two remaining $n_i \times 1$ unobserved random vectors, ζ_i and e_i, specify two types of error in Y_i. The ζ_i specify the sampling errors in Y_i and e_i specifies other sources of error or heterogeneity between studies and between arms of the same study.

In this model, it is assumed that β_i, u_i and e_i are each independent multivariate normal random vectors. On of the new extensions offered by this model is the possibility for random effect covariates. It is worth noting that the procedures for implementing the above model have recently been incorporated into the MS DOS-based clinical trials and epidemiology package, Epilog (14).

Using multi-level models for meta-analysis

Lambert and Abrams (15) present a method for carrying out a meta-analysis using multi-level models. They illustrate their method using a dataset of cholesterol lowering trials, very similar to the one used in chapters 8, 9 and 10. Using the software package ML3 (16), they implement a random effects model very similar to that of DerSimonian and Laird (7) (chapter 10). This is then extended into a mixed model to include a study-level covariate for baseline risk. The authors comment that it is in the mixed model scenario where this method can be used to great advantage because mixed models such as this, and more complicated situations can be modelled with relative ease. For a Bayesian formulation of multi-level models see chapter 13 and also page 200 on cross-design synthesis.

Problems/advantages with methods

Advantages

Generally, mixed models in general can be viewed as the best of both worlds. One can explain as much variation as reasonable and in the process possibly, create clinically important hypotheses for further investigation. The random effects term then accounts for whatever residual variation remains.

Disadvantages

There are drawbacks to this method. Firstly the limitations of a random effects analysis (chapter 10) exist in this method as well, notably: 1) the uncertainty from estimating σ^2_θ from the data is not incorporated in the model, 2) the need to assume that the random effects are normally distributed with constant variance. This is difficult to assess when the number of studies is small (though a t-distribution can be used if preferred). Compounded on these are all the pitfalls of fitting meta-regression models (discussed in chapter 11). Aside from these technical drawbacks, there are some practical ones: Raudenbush (3) notes that as with meta-regression models, the mixed-effect method is most useful when the number of studies is large, and indeed cannot sensibly be attempted when very small numbers of studies are being combined. It is worth noting that the methods outlined here cannot deal with dose–response regression analysis, for these see pages 157–61.

Further research

Similar further work issues to those covered in chapter 11 are relevant here, but there is an additional complication when using a multi-level approach as the general model (12.9). This is the question of how to choose which covariates are included with random coefficients and which are not, i.e. how do you decide whether there is sufficient heterogeneity between the studies identified by the levels of a factor to allow for a separate variance term to be included into the model. Whilst a comparison of deviances between the various models can be performed, such a method might not necessarily be appropriate, and further work is needed in this neglected area of mixed-effect modelling, certainly with respect to meta-analysis.

As with meta-regression, there is the issue of distributional assumptions, not only of the data, i.e. T_i dist $N[-, -]$, but also of the random effects. Previous work has often made choices on the grounds of computational convenience.

It should be noted that practical applications of these models seemed a bit thin in the meta-analysis literature.

More specifically, in considering their model, Huque and Dubey report:

> 'More theoretical and computational work is needed to assure the robustness of the estimates derived, or to derive other robust estimates and examine distributional aspects of the parameter estimates in the model.' (5)

Similarly, Berkey *et al.* (6) make suggestions for further work needed on their model: 1) the development of an alternative estimator of the σ_i^2 (the within-trial variances) because the two estimators they considered provide biases on the results in opposite directions, for both the no-covariate and single-covariate model; 2) because each new situation needs a new simulation study to determine the number of degrees of freedom of the *t*-distribution are necessary to get nominal coverages close to the 90% and 95% levels, further work defining a general rule would be desirable.

Summary

This chapter has extended the methods of meta-regression in chapter 11, to allow for the existence of between study heterogeneity that cannot be adequately modelled by fixed covariates in a meta-regression model. The simplest models simply allow for a single random effect term, whilst more complicated models can allow for different levels of between-study heterogeneity associated with differing levels of a factor using a hierarchical modelling framework.

References

1. Hedges LV. Commentary. *Stat Med* 1987;**6**:381–5.
2. Rubin D. Wachter KW, Straf ML, editors. A new perspective. In: The future of meta-analysis. New York: Russell Sage Foundation, 1992, p. 155–65.
3. Raudenbush SW, Cooper H, Hedges LV, editors. Random effects models. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 301–22.
4. Larholt KM. Statistical methods and heterogeneity in meta-analysis. Boston: Harvard School of Public Health, 1989.
5. Huque MF, Dubey SD. A metaanalysis methodology for utilizing study-level covariate-information from clinical-trials. *Commun Statist Theory Methods* 1994;**23**:377–94.
6. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med* 1995;**14**:395–411.
7. Dersimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177–88.
8. Emerson JD, Hoaglin DC, Mosteller F. A modified random-effect procedure for combining risk difference in sets of 2x2 tables from clinical trials. *J Ital Statist Soc* 1993;**2**:269–90.
9. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med* 1996;**15**:1713–28.
10. Stram DO. Meta-analysis of published data using a linear mixed-effects model. *Biometrics* 1996;**52**:536–44.
11. Peto R. Why do we need systematic overviews of randomised trials? *Stat Med* 1987;**6**:233–40.
12. Begg CB, Pilote L. A model for incorporating historical controls into a meta-analysis. *Biometrics* 1991;**47**:899–906.
13. Tori V, Simon R, Russek-Cohen E, Midthune D, Friedman M. Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. *J Natl Cancer Inst* 1992;**84**:407–14.
14. Epilog Plus. Statistics Package for Epidemiology and Clinical Trials [computer program]. Pasadena, California: Epicenter Sofware, 1994.
15. Lambert PC, Abrams KR. Meta-analysis using multilevel models. *Multilevel Modelling Newsletter* 1996;**7**:17–19.
16. Rasbash J,Woodhouse G. MLn Command Reference Multilevel Models Project (computer program). University of London: Institute of Education, 1995.

Chapter 13

Bayesian methods in meta-analysis

Introduction

This chapter reviews the use of Bayesian (both full and empirical) methods that have been used in the synthesis of studies (meta-analysis). Bayesian methods have become more frequently used in a number of areas of healthcare research, including meta-analysis, over the last few years (1–3). Though much of this increase in their use has been directly as a result of advances in computational methods, it has also been partly due to their more appealing nature, and also specifically the fact that they overcome some of the difficulties encountered by other methods traditionally used.

Unlike other chapters in this report a section providing background material is given. This includes a brief overview of what Bayesian methods are, and perhaps more importantly the underlying principles utilised. Whilst one of the reasons for the lack of use of such methods has been, that they required an understanding of the underlying principles, it is also due partly to the fact that until recently, there was little or no software available. Hence, non-statisticians found the use of such methods daunting, The first section of this chapter attempts to give a non-technical introduction to them.

General introduction to Bayesian methods

Non-technical introduction

Bayesian methods can be considered as an alternative to the classical approach to statistical analysis. The name, originates from the Reverend Thomas Bayes (1702–1761), who in papers published posthumously (4), outlined a different system for making statements regarding probabilities and random phenomena. At the heart of this alternative system was an equation which forms the basis of all modern Bayesian theory. This is now commonly referred to as Bayes' theorem.

Though this chapter is primarily concerned with meta-analysis, it is perhaps instructive to consider how Bayes' ideas relate to a single study before generalising it to the case when we have a number of studies. Consider again the motivating example outlined in chapter 5 on the relationship between cholesterol reduction and all-cause mortality. The first study, using cholesterol reduction as a primary intervention in that meta-analysis was carried out in 1969, now, in a classical statistical framework the analysis of that randomised trial would make use of only the data contained in the trial, it would certainly not take into account any laboratory, animal or non-randomised evidence. A Bayesian analysis would proceed, at least initially, by first summarising what the evidence of a relationship was prior to the RCT being conducted, this might be in terms of the OR or some similar measure of relative effect. Obviously, this in itself raises a number of issues; for example when extrapolating from animal studies to humans different people will hold different beliefs about how animal results will carry over to humans. They might also hold differing beliefs about how reliable the evidence was from say a number, of perhaps small, observational studies. The key aspect here is that different people will interpret the evidence prior to the RCT being conducted differently. This is a key element of the Bayesian approach, namely that different individuals have their own view of the world, and this introduces the idea of **subjective probability** (5). Traditionally, probabilities attached to specific events, say that a dice rolled will land with a six facing up, have been objective, and whilst this seems sensible for events such getting a six from a dice, when we consider human phenomena such interpretations have less meaning. Returning to the trial example, assuming that an individual has been able to summarise quantitatively their beliefs prior to the RCT being performed, then the key question which the Bayesian approach addresses is how do these beliefs change in the light of the evidence generated by the trial? The answer to such a question is that the prior beliefs of the individual are combined with the evidence generated by the trial using Bayes' theorem. The resulting beliefs *a posteriori* to the trial are then the beliefs the individual would hold if they updated their prior beliefs in the light of the trial evidence in a rational and coherent manner. A number of points should be noted. First, the posterior beliefs obtained by the application of Bayes' theorem may not indeed be the posterior beliefs held by an individual, since that individual may not be rational and coherent in their probabilistic reasoning.

Secondly, it has been the explicit beliefs of an individual that had been used, it could well be that a group of individuals have collectively expressed their prior beliefs regarding the possible relationship between cholesterol reduction and subsequent mortality. The issue of whose prior beliefs to use is an important one, and is an aspect of the Bayesian approach that has led to considerable criticism; it is discussed briefly below. The actual numerical application of the Bayesian approach to this issue of the first RCT in the cholesterol meta-analysis is considered again on pages 92–5.

Despite Bayes laying the foundations many years ago, it is only very recently that this approach has been adopted, not least due to the computational difficulties this method often poses. However, with the increase in computing power, which has facilitated specialist software to be written, a complete Bayesian analysis is now possible in many different research fields, including meta-analysis. Although increasing in popularity, this approach still has some way to go before it is accepted as common place in the science (medical) literature. It is also responsible for causing strong polar reactions among statisticians, who are either strong advocators or opponents of the general fundamental approach.

An issue raised above, and one that deserves more discussion in this section is the use of specific prior beliefs. Indeed, the specification of prior beliefs quantitatively is a difficult area, and one that to date has been neglected in the statistical literature, as the ability to consider realistically complex problems has been hampered by computational difficulties.

To add confusion, there is another group of methods which are termed **empirical Bayes** methods, and which are discussed from a technical perspective on page 95. The use of term empirical Bayes methods is unfortunate since some methods classified as such are not Bayesian at all. Generally, empirical Bayes methods proceed just as fully Bayesian methods, except that they do not incorporate subjective beliefs into the analysis, but rather estimate the prior from the data. Thus, they help to add structure to a problem, but remain ‘objective’ in terms of interpretation.

An important point to note is that the use of Bayes’ theorem, the basis of Bayesian methods, is also used in a diagnostic setting where manipulation of conditional probabilities is required. This application has aroused no controversy, and is not considered in this report.

General advantages/disadvantages of Bayesian methods

Whilst there are specific advantages to adopting a Bayesian approach, there are also a number of disadvantages. Below is a brief, and certainly not an exhaustive, list of some of the main advantages and disadvantages.

Advantages

- Allows probability statements to be made directly regarding quantities of interest, e.g. the probability that patients receiving drug A have better survival than drug B.
- Enables all evidence regarding a specific problem to be taken into account rather than just the current study, and thus allows a summary of the current state of knowledge.
- Enables predictive statements to be made easily, conditional on the current state of knowledge.
- Elicitation of prior beliefs requires investigators to think carefully as to what they really do expect. Combined with the elicitation of **demands**, i.e. the magnitude of difference that would be considered clinically significant, this allows for an investigation into the initiation, monitoring and stopping of studies.
- Similar units of analysis, i.e. in meta-analysis studies, to **borrow strength** from other studies in estimating say an individual study effect.
- Bayesian methods lead naturally into a decision theoretic framework which can also take into account utilities when making health care or policy decisions.

Disadvantages

- The use of prior beliefs destroys any element of objectivity.
- Eliciting prior beliefs is a non-trivial exercise, and at present there are few guidelines to help the Bayesian analyst. Though when also adopting a decision theoretic framework much work has been done in the elicitation of utilities.
- There is no automatic measure of statistical significance such as a *p*-value.
- They can be computationally complex to implement, and thus time consuming to perform.
- At present, there are software limitations, though this is changing rapidly.

Technical background

As described above, the Bayesian approach can be summarised as follows: opinions are expressed in probabilities, data are collected, and these data change the prior probabilities, through the operation of Bayes’ theorem, to yield posterior probabilities (6). Opinions are expressed in probabilities which implies that the subjective beliefs of the researchers (or possibly experts from the field/panel consensus), prior to conducting the analysis, form the starting point for the analysis. These prior beliefs are than combined with the data, in the form of a **likelihood function**, to produce a posterior distribution, which takes both subjective and objective evidence into account. It is in the incorporation of subjective beliefs that the Bayesian approach differs greatest from the classical viewpoint, which only considers objective evidence. However, another key difference between the Bayesian and Classical approach is the role that the likelihood function plays. In the classical approach, the likelihood function defines the support for various values of the parameter of interest, conditional upon the observed data. In the Bayesian approach since both the data **and** model parameters are considered random, the conditioning may be reversed, and thus the Bayesian considers the likelihood function to measure the plausibility of the observed data condition upon the parameters of the model (7). Although as part of the following, a brief example an illustration of the Bayesian approach to healthcare research is outlined, the interested reader is referred to any of the following for a more detailed account of the Bayesian approach generally (8–14). For details regarding Bayesian methods in randomised controlled trials, the following can be consulted (15–26).

The use of a Bayesian approach to inference in a single RCT is first considered. Considering the example of the first trial in the cholesterol example (see chapter 5), reported in 1969, more technical details of how the Bayesian approach proceeds in practice is given below. As mentioned above, the key component in a Bayesian analysis is the way in which *a priori* beliefs are updated in the light of new data via Bayes' theorem to yield posterior beliefs about relevant quantities of interest. Assuming that the quantity of interest is denoted by θ, the posterior density for θ, $P(\theta|\text{Data})$, is given by

$$P(\theta|\text{Data}) \propto P(\theta)\ P(\text{Data}|\theta). \tag{13.1}$$

In order that $P(\theta|\text{Data})$ is a proper density a constant of proportionality, *k*, is required so that $P(\theta|\text{Data})$ integrates in the continuous case or sums in the discrete case to one. Thus

$$P(\theta|\text{Data}) = k\,P(\theta)\ P(\text{Data}|\theta), \tag{13.2}$$

where $k = \int P(\theta)\ P(\theta|\text{Data})\ \partial\theta$, and is the integrating constant. All inference regarding θ then proceeds via the posterior density $P(\theta|\text{Data})$. Various summary measures of location such as the posterior mean, median and mode can be calculated for $P(\theta|\text{Data})$ together measures of dispersion such as the variance. Analogous to the calculation of CIs, **credibility intervals** may also be calculated. Such intervals have a direct probability interpretation, i.e. they are intervals in which θ lies with a certain probability. An important extension to the use of credibility intervals is the notion of **highest posterior density intervals**, which again have a direct probability interpretation, but are also unique intervals such that any specific value of θ outside the interval has lower point probability than points within the interval. Other summary measures such as the probability that θ is greater than a certain value of that θ lies in a certain interval may all be calculated directly from $P(\theta|\text{Data})$. Finally, one further aspect of a Bayesian analysis which is often required is the ability to predict future observations, conditional upon the data so far and *a priori* beliefs. In order to make such predictive statements the **predictive density** is required, and is given by

$$P(x|\text{Data}) = \int P(x|\theta)\ P(\theta|\text{Data})\ \partial\theta \tag{13.3}$$

where *x* is the future observation.

So far, the assumption that there is only parameter of interest, θ, has been made, but often in many healthcare research settings there are a number of parameters which though not necessarily of direct interest have to be considered in the analysis. For example, in regression we may only be interested in the slope of the regression line but we also have to estimate the intercept. Such parameters, which are of secondary interest, are termed **nuisance parameters**. The Bayesian methods outlined above follow when there is more than one parameter, but extra complications arise.
Thus, (13.2) becomes

$$P(\theta|\text{Data}) = k\,P(\theta)\ P(\text{Data}|\theta), \tag{13.4}$$

where θ is a vector of parameters and $k = \int_\theta P(\theta)\ P(\text{Data}|\theta)\ \partial\theta$. Although in theory Bayes' theorem can still be used quite straightforwardly even in the multi-parameter setting, it is often of interest to obtain a summary of posterior beliefs regarding a single parameter of interest or a function of the parameters of interest. This can be achieved by obtaining the **marginal posterior density**. Thus, if the first element of θ, θ_1, was of interest but the other elements of θ were not, these being denoted by θ_{-1}, then the marginal posterior density for θ_1, $P(\theta_1|\text{Data})$, is obtained by integrating out the nuisance parameters, θ_{-1}, from the joint posterior density, $P(\theta|\text{Data})$. Thus

$$P(\theta_1|\text{Data}) = \int_{\theta-1} P(\theta|\text{Data})\ \partial\theta_{-1} \qquad (13.5)$$

The marginal posterior density for θ_1 can now be used in exactly the same way as the posterior density in the single parameter case.

As can be seen by (13.2), (13.4) and (13.5), the routine use of Bayesian methods requires a number of possibility high dimensional integrals to be evaluated. It is this that has hampered the practical application of such methods for a considerable period of time. Essentially, three possible methods are available for their evaluation; asymptotic approximation methods, quadrature (numerical integration) methods and simulation (27). Though all three methods have been used, in practice the first two methods are only practicable when there are a relatively small number of parameters involved. Recently much work has been carried out in developing simulation based methods, and in particular on a group of methods broadly classified as **Markov Chain Monte Carlo** (MCMC) methods (28). Within this broad range of Monte Carlo simulation methods, one method, **Gibbs sampling**, has been increasingly used in applied Bayesian analyses within a healthcare research setting (2,28,29). The appeal of this method is that in wanting to summarise a posterior density, and in particular a marginal posterior density, simulating from often a high dimensional joint posterior density is often difficult, but the posterior conditional distributions, i.e. $P(\theta_1|\theta_{-1},\text{Data})$, are often much easier to sample from. Gibbs sampling uses this fact, together with ergodic theory which says that if the conditional densities are sampled from for a sufficiently long period of time, then the realisations will approximate the marginal posterior densities (28,30,31). Though one advantage of Gibbs sampling is its simplicity, and it can be performed in any programming environment, the development of a specific package, **BUGS** (32), has greatly increased its appeal and use.

An alternative method for implementing Bayesian analyses in practice is to use a specific prior distribution, which when combined with certain likelihood functions yields a posterior distribution from the same family as the prior distribution. In addition, if the family of distributions is relatively standard then this will enable summary statements to be made more easily (33). For example, if dealing with continuous data, and an assumption of normality can be reasonably made, and assuming that the mean is the parameter of interest, then by using a prior distribution for the mean, which is also a Normal distribution, the resulting posterior distribution is also a normal distribution. Thus, in this one parameter case making inferences about the posterior beliefs of the mean only involves summarising a Normal distribution. Such models are termed **conjugate models**, and other examples include the beta-binomial model, i.e. beta prior distribution, binomial likelihood, beta posterior distribution, and the gamma-Poisson model, i.e. gamma prior, Poisson likelihood and gamma posterior. Though all three models are single parameter models, they have the advantage of being fully tractable and often serve as an initial analysis.

As mentioned on pages 91–2, a key aspect of a Bayesian analysis is the role that the prior distribution plays, and indeed one of the criticisms of the Bayesian approach is its dependence on such prior distributions. The specification/elicitation of prior beliefs, especially in a multi-parameter setting, is also a non-trivial task. Therefore, a number of approaches have been developed in which **vague prior distributions** have been used, so that the data effectively dominate the prior distribution. One possibility is for $\mathbf{P}(\theta)$ or $\mathbf{P}(\theta)$ to be simply a constant, in which case the posterior density is in fact the **standardised likelihood**. Unfortunately, the use of vague prior distributions such as this means that they are not always invariant to transformations, and thus an alternative is the so called **Jeffreys' prior** (34). The key message is that the use of prior distributions is an important area and in any Bayesian analysis a **sensitivity analysis** in which a variety of prior distributions are used is a crucial aspect of any analysis.

Obviously, the details that have been discussed so far are somewhat abstract; on pages 95–100, Bayesian methods are specifically applied to the problem of meta-analysis. However, below, a Bayesian analysis of the first trial of the cholesterol meta-analysis is presented as a worked example using a normal–normal conjugate model.

Example

As previously mentioned, one has to express their prior beliefs in terms of a parametric distribution. For the sake of simplicity say the data are viewed as a random sample of size n from a normal distribution with unknown mean θ and known standard deviation σ, and the goal is to assess one's uncertainty about θ in light both of the data and of prior information.

Assuming that the summary statistic for the data, x_n, can be assumed to be normally distributed then

$$x_n \sim N[\theta,\sigma^2/n].$$

Assuming further that the *a priori* beliefs regarding θ can be expressed as a normal distribution with mean θ_0 and variance σ^2/n_0,

$$\theta \sim \mathrm{N}[\theta_0, \sigma^2/n_0]$$

and the resulting posterior distribution is given by

$$\theta \mid x_n \sim \mathrm{N}[(n_0\,\theta_0 + n\,x_n)/(n + n_0),\ \sigma^2/(n + n_0)]$$

where x_n is the log(OR) $\sigma^2 = 4$ (25).

In terms of the first primary study in the cholesterol meta-analysis, the observed data were 174 deaths out of 424 patients on the treatment arm, and 178 deaths out of 422 patients on the control arm. Thus, the log(OR) is –0.05, and n = 352. Assuming that our *a priori* beliefs were consistent with a log(OR) of zero, i.e. no effect, but assuming uncertainty associated with this belief being represented by a hypothetical trial in which 100 events were observed, then $\theta_0 = 0$ and $n_0 = 100$. Thus, the posterior distribution is

$$\begin{aligned}\theta \mid x_n &\sim N[(100\,0 + 352\ {-0.05})/(352 + 100), 4/(352 + 100)]\\ &\sim N[-0.04, 0.092]\end{aligned}$$

Thus, we can see that the posterior mean for θ has been shifted slightly towards zero as a result of the prior beliefs, but that the amount by which is has been modified is in proportion to the ratio of the *a priori* and observed variances. The other point to notice is that the posterior variance, 0.008, is smaller than the observed variance, 0.01, reflecting the fact that there has been an increase in the amount of evidence on which the analysis has been based.

Empirical Bayes

A group of methods termed **empirical Bayes** have become increasingly used in healthcare research, though there is also a considerable body of literature on these methods generally (35–41). Such methods have acquired the term empirical Bayes because they make use of some of the methods of the Bayesian approach, but the key aspect of subjective probability and inclusion of subjective beliefs do not carry over. Such methods are termed Bayes because they use the idea of a prior distribution to impose some sort of structure on a problem, but they do not use subjective *a priori* beliefs to derive/elicit actual numerical values for the hyper-parameters of the prior distributions. Instead they estimate the most plausible values of the hyper-parameters from the data. The key issue is that empirical Bayes methods only use the actual observed data, though some element of subjective judgement does have a role to play in the choice of the form of the prior distribution, as using different prior distributions may change the results of an analysis by imposing different structures on the problem.

For example, consider data $x_1,\ldots,x_n$ assumed to be derived from a normal distribution with mean μ and variance σ^2, and that σ^2 is assumed known but that μ is unknown. Suppose a prior distribution is to be assumed for μ, such that this to is a normal distribution with mean η and variance τ. In a fully Bayesian analysis, the hyper-parameters η and τ would be completely specified by an individual, but in an EB analysis would estimate the most likely values of η and τ given $x_1,\ldots,x_n$, the data. Thus

$$P(\mu|x) = k\,P(\mu|\eta,\tau)\ P(x|\mu) \qquad (13.6)$$

where η and τ are such that

$$m(x_i) = \int P(x_i|\mu)\ P(\mu|\eta,\tau)\ \partial\mu \qquad (13.7)$$

and where the marginal for $x_1,\ldots,x_n$, is given by $m(x) = \prod_{i=1} m(x_i)$. η and τ are then chosen such that $m(x)$ is maximised. Obviously, evaluation of (13.7) requires integration, though by making a number of assumptions, such as normality, analytically tractable solutions exists for a number of special cases.

Applications of Bayesian methods in meta-analysis

Having established the idea behind a Bayesian analysis in the previous section, here we explore how it can be applied in the context of meta-analysis.

Bayesian meta-analysis of normally distributed data

Many of the authors who have considered a Bayesian approach to meta-analysis have indeed extended the normal theory model outlined on pages 4–5 to a hierarchical setting (42–48). In other areas of statistical science such Bayesian hierarchical models have been used for a considerable time (49,50). Before considering specific approaches taken, a basic hierarchical model similar to the random effects model of chapter 10 is outlined.

Assume that the *i*th study in a meta-analysis can be summarised by an outcome measure y_i, for example, a log(OR) or difference in means, and that associated with the outcome measure is a within-study variance $\sigma^2_{\rm i}$, and the size of the study n_i, then the first level of the model relates the observed outcome measure y_i to the underlying effect in the ith study θ_i. At the second level of the model the θ_is are related to the overall effect in the assumed population from which all the studies are assumed to have been sampled η, and τ^2 is the between-study variance or the variance of the effects in a population. So far the derivation of such a model is exactly analogous to the random effects model of chapter 10. However, from a Bayesian perspective, a number of unknown parameters exist which are to be estimated, $\sigma^2_{\rm i}$, μ and τ^2 and therefore require prior distributions in a Bayesian setting. Thus, denoting an arbitrary prior distribution by [-,-] the model has the following form

$$\begin{array}{lll} y_i \sim N[\theta_i,\sigma^2_{\rm i}/n_i] & \sigma^2_i \sim [\text{-,-}] & i = 1,\ldots,k \\ \theta_i \sim N[\mu,\tau^2] & & (13.8) \\ \mu \sim [\text{-,-}] & \tau^2 \sim [\text{-,-}] & \end{array}$$

Having specified the three required prior distributions in terms of the relevant hyperparameters, estimation can then proceed using a number of computational approaches as outlined on pages 4–5. However, the assumption of normality that has been made here, combined with the fact that there are often a reasonable number of studies in any specific meta-analysis make such models particularly suited to MCMC methods.

However, the specification of the prior distributions is not a trivial task and the choice of which prior distribution to choose has received considerable attention recently. (51,52)

Inference regarding $\theta_{\rm i}$

Obviously sometimes interest focuses upon the individual study effects, the θ_is, and conditionally upon μ and τ^2 the θ_is have analytically tractable expressions for the mean and variance, which are

$$\begin{aligned} E[\theta_{\rm i}|y, \mu, \tau^2] &= B\mu + (1 - B)y_i \\ &= \mu + (1 - B)(y_i - \mu) \end{aligned} \qquad (13.9)$$

and the posterior variance is

$$V[\theta_i|y,\mu,\tau^2] = (1 - B)\,\frac{\sigma^2_i}{n_i} \qquad (13.10)$$

where

$$B_i = \frac{\sigma^2_i/n_i}{\sigma^2_i/n_i + \tau^2} \qquad (13.11)$$

These expression are analogous to those in chapter 10 for the classical random effects model, but they are conditional upon both μ and τ^2 being known. They show that the effect in the ith study is shrunk towards the overall population mean by a B, and thus from (13.11) it can be seen that for studies which have a larger within-study variance there will be more shrinkage than for less heterogeneous studies.

Obviously, estimates for both μ and τ^2 are required. As the model stands, it can be thought of as an EB approach, with μ and τ^2 defining the prior distribution, i.e. the second level of the model, for each of the θ_is. They could be estimated from the data using either a method of moments or restricted ML. However, expressions (13.9) and (13.10) for the mean and variance of the θ_is at present do not account for the fact that they have been estimated. Carlin (53) has shown that under an assumption of non-informative locally uniform prior distributions for both μ and τ^2 expressions (13.9) and (13.10) may be re-written to take into account the fact that μ has been estimated, but are still conditional upon τ^2. Thus

$$E[\theta_i|\mathbf{y}, \tau^2] = \mu + (1 - B)(y_i - \mu) \qquad (13.12)$$

$$V[\theta_i|\mathbf{y}, \tau^2] = w_i\,\sigma^2_{\rm i} + (1 - w_i)^2\,\tau^2/\Sigma_i\, w_i \qquad (13.13)$$

where $w_i = (1 + \sigma^2_{\rm i}/\tau^2)^{-1}$ and the second term in (13.13) estimates the posterior covariance between two study effects. In order to obtain estimates which are totally unconditional numerical methods have to be employed, since the joint posterior density for θ_is and τ^2 has to be integrated with respect to τ^2. In essence, a fully Bayesian analysis is required and unfortunately no analytically tractable solutions exist.

Inference regarding μ

Often, the main focus of interest is μ, the overall population effect. As with inferences regarding the θ_is it is only possible to obtain simple expressions for the mean and variance of μ conditional upon τ^2 when vague non-informative prior distributions are assumed for both. Thus

$$E[\mu|\mathbf{y}, \tau^2] = \Sigma_i\, w_i\, y_i\,/\Sigma_i\, w_i \qquad (13.14)$$

$$V[\mu|\mathbf{y}, \tau^2] = \tau^2/\Sigma_i\, w_i. \qquad (13.15)$$

As with inferences regarding the individual effects above, in order to obtain posterior mean and variance unconditional upon τ^2 numerical methods have to be employed.

***Inference regarding* τ^2**

From a classical perspective, chapter 10 demonstrated how τ^2 can be estimated as effectively negative, i.e. the within-study variability is greater than the between-study variability. From a fully Bayesian perspective, such a situation is not possible under any plausible prior distribution (see below).

Choice of prior distributions for σ_i^2, μ and τ^2

Before discussing particular technical details, the issue of the choice of σ_i^2 deserves mention. Some authors have claimed that whether the σ_i^2s are assumed known and replaced with the observed within-study variances or whether they are assumed random and therefore have a prior distribution specified makes little practical difference a part from when there are a number of small studies (Carlin, 1992). If the σ_i^2s are considered random and therefore a prior distribution required then a number of possibilities exist. The most frequently used prior distribution is $P(\sigma_i^2) \propto 1/\sigma_i^2$ which corresponds to a Jeffreys' prior. Although appealing for theoretical reasons, such a prior distribution is not always feasible in practice, and a commonly used alternative distributional-based prior is an inverse gamma distribution. This distribution is particularly flexible, and can accommodate a number of possible scenarios, it also has the benefit of only being defined on the positive real line.

In terms of a prior distribution for μ it is common practice to either assume a particularly vague proper prior distribution or to use a uniform distribution over the whole real line, reflecting the fact that we often wish to remain relatively objective about inferences regarding the pooled overall effect. Frequently, though a suitably vague normal distribution is used as a prior distribution for μ since this can aid estimation of the parameters. Obviously the use of any prior distribution should be subjected to a sensitivity analysis.

Bayesian meta-analysis of binary data

All the model derivation on pages 95–7 has assumed that the outcome measure for each study can be assumed to be normally distributed. Whilst making such an assumption facilitates estimation, this might not be tenable from a practical point of view.

Two possible model formulations exist, and have been considered to date. Consonni and Veronese (54) consider the modelling of binary outcome data in meta-analyses directly in a hierarchical model, with the observed responses in a single arm of the trial being modelled using binomial distributions, with conjugate Beta distributions at the further levels of the model. Though such an approach is computationally attractive, due to the conjugate nature of the model, it is of limited value in comparative experiments.

An alternative model formulation, which has been adopted by a number of authors (55–61), is briefly described in a general form below. In this approach, although the observed responses in each arm of a trial are assumed to follow a binomial distribution, a suitable transformation is then applied, frequently logit in nature, to the rates parameters. Following such a transformation there model formulation proceeds as on pages 95–7, though parameter estimation requires some form of numerical, simulation, or approximation method, to be employed.

Consider a two-arm study in which r_1 and r_2 are the observed number of responses out of n_1 and n_2, respectively. Then the first level of the model is

$$r_1 \sim \mathrm{Bin}[\pi_1, n_1] \qquad r_2 \sim \mathrm{Bin}[\pi_2, n_2] \qquad (13.16)$$

where π_1 and π_2 are the two unknown rate parameters for the two arms of the study. Consider now the logit transformation of each of the two rate parameters such that

$$\begin{aligned} \log(\pi_1/1 - \pi_1) &= \mu_i - \delta_i/2 \qquad (13.17) \\ \log(\pi_2/1 - \pi_2) &= \mu_i + \delta_i/2 \end{aligned}$$

δ_i is now the parameter of interest, being the log(OR). This is often then assumed to be approximately Normally distributed and the second level of the model becomes

$$\delta_i \sim N[\phi, \tau^2] \qquad (13.18)$$

where ϕ represents the overall pooled effect, on a log(OR) scale, and τ^2 is a measure of the between study heterogeneity. As on pages 95–7, a fully Bayesian analysis prior distributions have to be specified for both ϕ and τ^2. Thus, as before the final level of the model is

$$\phi \sim [\text{-},\text{-}] \qquad \tau^2 \sim [\text{-},\text{-}] \qquad (13.19)$$

The key difference between this model and (13.8) is the assumption that at the lowest level of the model the responses in each study are modelled directly. In (13.8) calculation of the log(OR) when there are zero or complete responses in any studies requires various assumptions to be made, usually by the addition of 'small' constants to the responses frequencies. It is this assumption of normality of the log(OR) or other transformed measures of binary data in models such as (13.8) that is frequently not validated.

Empirical Bayes methods in meta-analysis

The use of EB approaches has received much attention in the literature since until recently the use of a fully Bayes approach has been hampered by computational difficulties (36,37,59,62–66).

However, the EB methods that have been used have almost exclusively assumed that the 'prior distribution' has been the second level of the Bayesian hierarchical model (13.8), and that the hyper-parameters, in this case μ and τ^2, have then been estimated from the data. Such an approach is analogous to assuming instead a three-level model as (13.8) and assuming uniform prior distributions for both μ and τ^2 as used by Carlin (53), and indeed the parameter estimates obtained using such an approach are given by (13.12)–(13.15). In addition, Smith *et al.* (57) and Biggerstaff *et al.* (67) explain that the random effects model of DerSimonian and Laird (68) (see chapter 10) could be viewed as an EB approach. The main drawback with this, though, is that no allowance is made from the fact that τ^2 has been estimated from the data, using either ML or moment estimation methods. Indeed, Carlin (53) goes on to show that in order to take account of this, some form of quadrature or simulation method is required.

In theory, distributions other than uniform distributions could be assumed and the hyper-parameters of these prior distributions could be estimated from the data, utilising the general concept of EB methodology outlined on page 95. However, such an approach, though appealing in that it retains the objectivity afforded by the empirical approach and allows for the fact that both μ and τ^2 have been estimated from the data, is as computationally complex as a full Bayesian approach, and it is no doubt for this reason that such a method has not been utilised in practice.

Advantages/disadvantages of Bayesian methods in meta-analysis

Advantages

Unified modelling approach

By using a Bayesian modelling approach for combining studies, the debate over the appropriateness of fixed and random effect models (see pages 104–5) is overcome, whilst at the same time including the possibility of regression models (57).

Borrowing strength

Borrowing strength can be seen as a by-product of a fully Bayesian meta-analysis model. When study estimates are combined, the model updates estimates of the individual studies, taking into consideration the results from all the other studies in the analysis. Thus, narrower CIs will be obtained for each individual study, by **borrowing strength** from all other studies. As well as reducing the width of the CI, the point estimates of the individual studies will also be affected, moving them closer together towards the overall pooled estimate. Gaver *et al.* (69) report that a variety of statistical ideas and terms are used to describe this concept, including **shrinkage**, **empirical Bayes** and **hierarchical Bayesian** modelling.

These concepts are particularly useful if one is interested, not in some overall, 'average', of the study results, but instead about making inferences about any particular treatment effect, then results from the other studies can be used to, 'improve', this estimate. This leads to better point estimates and shorter interval estimates of any particular effect. Gaver *et al.* (69) note that approaches to borrowing strength, with applications to medicine and health, are much rarer than the meta-analytic approach that focuses on estimating population parameters. However, he does point out that DuMouchel and Harris (44) elaborate such a method to improve estimation of cancer effects in humans, by borrowing strength from experimental data on laboratory animals in experiments using the same carcinogens, and Raudenbush and Bryk (63) propose estimation by EB and Stein-type methods; though Morris (65) points out Stein's estimator, used in borrowing strength, can only be used when the variances of the studies are the same (i.e. almost never). In addition Laird and Louis (70) and Carlin and Gelfand (1990) give parametric and bootstrap methods for constructing EB CIs which may be applied to obtain individual study estimation. Indeed, recently interest in the concept of borrowing strength has increased with respect to institutional comparisons, see Spiegelhalter and Goldstein (71).

Allowing for all parameter uncertainty

EM (simple models) and classical approaches do not allow for the fact that both μ (mean), and τ^2 (the between study variance) have both been estimated from the data.

Allowing for other sources of evidence

Often meta-analyses are conducted in substantive areas in which evidence is available from sources other than RCTs, the main source of evidence, i.e. when the majority of evidence is from RCTs, but other evidence exists in the form of observational studies (see chapter 26 for more information).

Ability to make direct probability statements
Tweedie *et al.* (72) can give a probability that the effect is above (or below) one (i.e. direct answer to question of interest).

Prediction
Using a Bayesian analysis [specifically equation (13.3)], it is possible to incorporate evidence from previous trials into the design of a new one (73). Using this approach, one can take the result of a meta-analysis and calculate the probability that the current/planned study (fixed sample size) will produce conclusive results.

Disadvantages

Specification of prior distributions
When carrying out a Bayesian analysis, one needs to specify distributions for the population effect size (μ) and the between-studies standard deviation (τ) in true, underlying effect sizes. Louis and Zelterman report:

> 'Generally, this elicitation is done by asking the respondent for a 'best guess' for the mean or median. Using this measure of centre to anchor the distribution, the respondent is asked for additional percentiles of the distribution. An individual who has 'no idea' of values for these parameters would specify that A (the prior variance of μ) is extremely large and that the distribution for τ also has a large variance. Such priors are called 'non-informative.' Generally, reasonable ranges can be specified for parameters, even if one cannot produce much detail of relative probabilities within the range.' (74)

They go on to comment that:

> 'Although we may 'all think like Bayesians', it can be extremely difficult to evaluate and communicate prior opinions, and considerable research continues on this aspect of Bayesian analysis.' (74)

Sensitivity to prior distributions
Importantly, a meta-analysis is not conducted to inform a single individual, but to communicate the current state of information to a broad community of consumers. If the prior distributions differ substantially for different consumers, then the related Bayesian analyses can produce qualitatively as well as quantitatively different results. Therefore, it is important to perform a sensitivity analysis over the range of opinions. If conclusions are stable then we have 'findings'. If they are not, the collection of Bayesian analyses underscores the finding that the data are not sufficiently compelling to bring a group of relevant consumers to consensus. This situation should motivate additional primary studies.

Calculation of posterior
Producing the posterior distribution and computing it can be difficult. In the continuous case, one needs to evaluate complicated integrals that replace the summations in the preceding formulae, and only the most basic models are mathematically tractable. Until recently, more complex but still quite basic models were handled by approximating the posterior mean and variance. However, recent advances in computational approaches allow the analyst to produce full posterior distributions for complicated models.

Comparison of classical and Bayesian approaches
A number of authors have explicitly compared classical and Bayesian approaches to meta-analysis; these are briefly reviewed below.

Carlin (53) compares Bayes and EB estimates. He observes the empirical ones are artificially accurate, i.e. the variance of the pooled estimate is too small.

Smith *et al.* (57) compare many methods (and software packages) for carrying out meta-analysis, including fixed, random and full Bayesian models.

Su and Po (75) compare EB, fully Bayesian, and classical fixed-effect methods. They use four data sets including beta-blockers as treatment for myocardial infarction, and case control studies investigating the association between smoking and lung cancer. They concluded that Bayesian methods were more conservative, with the fully Bayesian model producing the widest CIs. They also report that the use of any one method exclusively would not have changed the conclusions, though when the heterogeneity was artificially increased Bayesian methods straddled unity while the other methods did not. Differences did exist though in the point estimates and CIs for specific studies (particularly small ones). The authors report that the importance of these differences needs investigating.

Biggerstaff *et al.* (67) compare classical with Bayesian techniques (includes random effects and EB methods) for case–control studies of passive smoking in the workplace.[1]

[1] This study is also noteworthy in that it compares different methods for investigating individual study estimates also, namely the use of Fisher's exact, Mantel–Haenszel, and logit methods.

Tweedie *et al.* (72) compare classical (random effects) and Bayesian (exact) methods with similar conclusions to those of Biggerstaff *et al.* (67).

Rogatko (56) compares random effects, asymptotic Bayes and exact Bayes methods using the risk difference scale on both simulated and real data.

Finally, Morris (65) compares fixed effects, random effects, likelihood, adjusted likelihood, and Bayes (Gibbs sampling) methods.

Extensions and specific areas of application

Incorporating study quality

The assessment of and inclusion of study quality was considered in chapter 6. Clearly the quantitative assessment of a measure of study quality may be included in a Bayesian analysis in the same way that it might be in a classical analysis, and extensions of the Bayesian meta-analysis models considered earlier to incorporate covariate information are presented on below. Alternatively, the assessment of study quality by one or more experts or expert meta-analysts may be considered either in the form of elicited *a priori* beliefs regarding an underlying quality process, i.e. treated as random/latent variables in a meta-analysis or as, perhaps after suitable transformation, a set of prior distributions directly for one of the model parameters, perhaps the individual study variances that are to be used in some form of weighting of the studies.

These various scenarios raise a number of questions, the key one of which is what is data and what are prior beliefs (74).Whilst all three approaches are feasible in practice relatively little work has been conducted in this field. Smith *et al.* (76) considered the inclusion of quality in relation to the probability of publication of studies, since there is often assumed to be a relationship between the two, whilst Smith *et al.* (51) considered the inclusion of quality in relationship to the credibility of different research designs in a generalised synthesis of evidence approach, in which prior distributions were assumed for the variance parameters so as to reflect varying degrees of credibility.

Covariates

To date many of the applications of Bayesian methods in meta-analysis have been to mirror the random effects models of chapter 10. This has been partly due to the computational difficulties in applying fully Bayesian models, and partly due to the fact that the use of Bayesian methods in meta-analysis has been at the beginning of a learning curve. In theory extension of model (13.8) and (13.18) poses no difficulties, with μ being replaced by $\beta^T x_i$, where β is a vector of regression coefficients and x_i is a vector of study-level covariates. In a Bayesian setting just as a prior distribution was required to be specified for μ, one also needs to be specified for β. In such settings it would appear that the use of MCMC methods (see pages 92–5) is particularly appealing, since the inclusion of extra parameters will almost certainly preclude the use of other numerical methods. The inclusion of covariates in a Bayesian meta-analysis has been considered by Louis and Zetlerman (74).

The use of such covariates raises a number of issues. First, is the problem of when there is data at both the study level and the patient level. In theory such a scenario could be accommodated with a more complex hierarchical model (see page 102). Another issue is that study-level covariates, especially when derived from published studies, may be subject to various **measurement errors**. Measurement error here is used in its broadest sense. For example, assume that studies report the average age of patients included, and that it appears that age is an important factor in explaining between-study heterogeneity. If for some studies age was only in fact obtained on a subset of the total patients in a study, then potential biases could be introduced into any analysis.

Model selection

As with any modelling exercise, the eventual selection of a ‘final model’ is a difficult task, and one which in the meta-analysis literature has received little attention. This is partly as a result of the relative lack of use of regression models generally, both Bayesian and classical. That having been said, one aspect of model selection that has received considerable attention and aroused heated debate is the choice between fixed and random effect models (see chapter 10). From a Bayesian perspective, this is almost a non-sequitur, since exploration of the marginal posterior distribution for τ^2 will yield an assessment of any between-study heterogeneity present. However, Abrams and Sansó (77) have considered the choice between such models within a Bayesian framework using BFs to discriminate between the two models. For an introduction to BFs see (78).

The key idea is that the posterior probabilities of the models are obtained using Bayes’ theorem. Thus, consider two models M_1 and M_2, and the *a priori* probabilities of the models being correct $P(M_1)$ and $P(M_2)$, such that $P(M_1) + P(M_2) = 1$, then the ratio of the posterior probabilities is given by

$$P(M_1|\text{Data})/P(M_2|\text{Data}) = P(\text{Data}|M_1)/P(\text{Data}|M_2) \times P(M_1)/P(M_2) \quad (13.20)$$

where $P(\text{Data}|M_1)/P(\text{Data}|M_2)$ is referred to as the **BF**. Thus, the BF is similar to the likelihood ratio, except $P(\text{Data}|M_i)$ is given by

$$P(\text{Data}|M_i) = \int_{\theta i} P(\text{Data}|M_i, \theta_i)\ P(\theta_i|M_i)\ \partial\theta_i \quad (13.21)$$

where θ_i is the vector of parameters associated with model M_i. Various scales have been advocated for the assessment of BFs (78), but one advantage of their use is the fact that if the actual posterior probabilities for each model is calculated, then these can be used to average across all models considered. This has the attractive advantage of avoiding the choice of any particular model, and also taking into account the fact that there is not only uncertainty associated with each of the models, but also between the models. Such model averaging cannot only be applied to the simple case where only fixed and random effects models are considered, but also when there are various models defined by covariates (both fixed and random). Abrams and Sansó (77) consider model averaging with respect to the cholesterol meta-analysis introduced in chapter 5, in which a comparison and averaging was performed with respect to a fixed, random, fixed regression and mixed effect model.

Missing data

As in any healthcare research setting, data are frequently missing. In meta-analysis missing data may either be at the study level or at the patient level; when patient information is available, it may also be that data is missing either in terms of covariate information or in terms of outcome measures. General issues concerning missing data are considered in chapter 17.

Whilst there has been considerable research into missing data generally in healthcare settings from a Bayesian perspective, this has been concentrated in terms of data within individual RCTs, and in the area of longitudinal data analysis (29,79) in essence using data augmentation techniques (80), there has been relatively little work in a meta-analysis setting [however, see Lambert *et al.* (81)]. This latter paper considered the case when mortality data from five different centres investigating neuroblastoma was pooled, and there was random and systematic missing covariate tumour marker data, and utilised the methods of Best *et al.* (29).

Publication bias

Recently, an area that has received attention from a Bayesian perspective is the issue surrounding publication bias. Essentially, two approaches have been taken. The first attempts to estimate the level of publication bias present and to make allowance for it in the subsequent analysis. The second, though similar approach, attempts to estimate the number of unpublished studies in a particular substantive field. Obviously both approaches have a common component, namely assuming that there exist a number of unobserved studies (or latent data), and that ideally inferences regarding the efficacy or effectiveness of any technology utilises both sources of evidence.

Givens *et al.* (82), Tweedie (83), and LaFleur (84) all consider a data augmentation approach to accommodate the latent data structure, with a weight function expressing the probability that any particular study, observed or not, is published which is dependent upon the study's associated *p*-value. In a similar manner, Larose and Dey (85) also adopt a data augmentation approach though they explicitly adopt a distributional form for the density of the unobserved studies, which is dependent upon a vector of unknown parameters. This formulation can accommodate the possibility that the probability of studies being published may be dependent upon a number of covariates, e.g. quality, size, statistical significance. Both approaches use MCMC methods to estimate the model parameters, and Larose and Dey (85) also go on to consider the use of BFs as a means of selecting between various competing models. In a similar manner to both the previous approaches, Paul (86,87) considers the use of a weight function within a hierarchical selection model, but also considers the situation when the weight function/density is modelled non-parametrically, this approach is not extended to the case when covariates are included in the model.

Gleser and Olkin (88) and Eberly and Casella (89) both consider the estimation of the number of unseen/unpublished studies in a meta-analysis. As opposed to the previous approaches which all had the main aim of estimating an overall effect making allowance for publication bias, these two approaches directly model the number of studies that were not found in a literature search. Eberly and Casella (89) assume that the total number of studies can be modelled via a negative binomial distribution, which depends upon the probability of publication. In turn, through specification of a beta prior distribution for the probability of publication, this will depend upon the probability of obtaining a statistically significant result. In order to obtain the marginal distribution for the total number of studies unconditionally upon the probability of publication, MCMC methods are required to perform the corresponding integration. Gleser and Olkin (88) also consider a negative binomial model, but rather than using MCMC methods to estimate the model parameters use ML/EB methods. For more information on publication bias see chapter 16.

Cumulative meta-analysis

Cumulative meta-analytic techniques introduced in chapter 25 are inherently Bayesian, since the sequential updating of the current state of knowledge concerning a particular technology or technologies mirrors that used in Bayes' theorem. However, to date little work has been done in the application of Bayesian methods to the problem of cumulative meta-analysis with the exception of Lau *et al.* (90) and Schmid *et al.* (91). Indeed, both of these papers, although explicitly acknowledging the strong similarity between the two approaches, do not formally adopt a Bayesian methodology. A further link is that with decision theory, that in using a cumulative or temporally sequential approach, a decision is ultimately made, either formally or informally, that there is sufficient evidence to warrant the adopting of a particular technology in routine healthcare practice. To date, however, there has been little or no work in this area (92).

Hierarchical models

The use of Bayesian methods that mirror the random effects models considered in chapter 10 continue to make the assumption that all studies can be considered on an equal footing, and that any one study can be used to help inform inferences made about any other study, i.e. the notion of exchangeability. This may not be a reasonable assumption for a variety of reasons, and when it is not, a hierarchical modelling framework allows some relaxation of this assumption by grouping studies that can be considered exchangeable. The theory underlying the use of hierarchical models has been studied within the mainstream Bayesian literature by a number of authors (9,49,93), and many of the more salient features of such models have been identified.

For example, in the cholesterol example of chapter 5, it might be postulated that drug studies were in some way similar, and similarly that diet intervention studies were similar, but that any one drug study could not **directly** inform the results of any one diet study. Thus, individual studies would be nested within type of intervention. Algebraically, such a model would be an extension of (13.8) and could be formulated as

$$\begin{array}{lll} y_{ij} \sim N[\theta_{ij},\sigma^2_{ij}/n_{ij}] & \sigma^2_{ij} \sim [-,-] & i = 1,\ldots,I_j\ j = 1,\ldots,J \\ \theta_{ij} \sim N[\nu_j,\omega_j] & \omega_j \sim [-,-] & \\ \nu_j \sim N[\mu,\tau^2] & & \\ \mu \sim [-,-] & \tau^2 \sim [-,-] & \end{array} \qquad (13.22)$$

where subscript i_j refers to the ith study in the jth category, and thus ω_j refers to the overall pooled effect in the jth category, and ω_j is a measure of the between-study heterogeneity in the jth category, whilst μ is the overall population effect and τ^2 is a measure of the heterogeneity in this effect between the J categories.

Use of such models has been relatively limited due to computational complexities, though again as with a number of areas the recent increase in the use of MCMC methodology such methods are becoming feasible in practice. Interestingly from a Classical perspective such models often referred to as **multi-level models** (94) or **random component models** (95) have been increasingly used in a variety of healthcare settings with the advent of relatively user-friendly software.

In adopting a hierarchical approach to meta-analysis, there are strong similarities with other areas of application in the healthcare settings. Two particularly relevant ones are in multicentre clinical trials and repeated measures data. In the former, centres take the role of studies, with patients nested within both, and other levels of the hierarchy correspond to factors which define groupings of centres/studies within which centres/studies are exchangeable. In the situation when IPD is used in a meta-analysis, a further level corresponds to patients nested within studies, which are nested within other factors. Similarly, for repeated measures data measurements are nested within patients, who may then be nested within other factors. Bayesian approaches to both of these settings, i.e. multicentre trials and repeated measures have recently been advocated. Stangl (96) considered the use of Bayesian methods in the analysis of a multicentre trial in which the primary outcome was time to event data, with censoring. The analysis proceeds along decision theoretic lines, with parameters estimated using Gibbs sampling. Similarly, Gray (97) also considered a hierarchical in a multicentre setting, again with a survival end-point, but was particularly interest in modelling the heterogeneity, both within and between centres, and also used Gibbs sampling to estimate model parameters. Veronese (98) considers the case when the assumption of exchangeability might not be reasonable for all studies in a meta-analysis, but that uncertainty exists as to the optimum partition of the studies within a hierarchical model. One possible solution advocated uses a partially exchangeable prior distribution at the second level in place the usual exchangeable one. More recently, Larose and Dey (99) also considered the situation in which groups of studies within a meta-analysis cannot be considered exchangeable, and used what they termed a 'grouped random effects model'. This model was a standard hierarchical model, assuming binomial sampling at level 1, and with level 2 defined by a partition, in their case depending on whether the study has closed or not. Estimation of the model parameters was by means of MCMC methods. Though they do not consider explicitly either the issue of model selection within a Bayesian framework nor the extension of the model to allow for study level covariates, they note that the former could be implemented within a Bayesian approach by means of BFs (see page 100), whilst the latter is addressed in a companion technical report (100). Finally, Higgins and Whitehead (60) consider the use of a Bayesian hierarchical approach to the situation when there are both patient level and study level covariates. As with many of the other expositions the model parameters are estimated using MCMC methods.

Generalised synthesis of evidence

Generalised synthesis of evidence refers to the case when the evidence for the efficacy of effectiveness of a particular technology is derived from a number of sources, not necessarily sharing common biases and designs. Approaches to this problem are considered in detail in chapter 26. Here, three Bayesian approaches are considered further.

One particular approach is to utilise a hierarchical model structure as outlined above, and this has been advocated by Abrams and Jones (101) and considered in detail by Smith *et al.* (51). The advantages of such an approach include the ability to explicitly allow or the fact that evidence from different sources may not be considered exchangeable, and therefore source may be used to define a level within the model. A further advantage is that *a priori* beliefs regarding the credibility of the different sources may be incorporated via the prior distributions specified for the model parameters, in particular the variance parameters, which drive the weighting ability of the model in the overall synthesis. As with all the models considered in this chapter, the role such subjective prior distributions play needs to be carefully explored in a subsequent sensitivity analysis (also see pages 203–5).

A second approach has been advocated by Hasselblad and McCrory (102), using the confidence profile method. The confidence profile method is a very general method for combining virtually any kind of evidence about various parameters, so long as those parameters can be described in the model, and was first described by Eddy (103). In comparison the Bayesian hierarchical method outlined above the confidence profile method utilises specification regarding the weighting of various sources by means of weight and bias functions, relating to the model parameters. It is this derivation in terms of essentially an overall likelihood function that makes the confidence profile method suitable to both ML and Bayesian methods of parameter estimation (also see page 202).

Finally, Wolpert and Warren-Hicks (104) investigate the response of fish in lakes to acid rain. The two sources of evidence are derived from laboratory experiments and field survey data and are synthesised using a hierarchical Bayes model.

Combining diagnostic test results

Two aspects of combining evidence relating to diagnostic testing have been considered. The first uses receiver operating characteristic (ROC) curves to summarise evidence in each study regarding a test, whilst the second approach uses the Bayesian posterior probability referred to on pages 92–5.

Zhou (64) considers an EB approach to the combination of areas under ROC curves (plot of specificity versus sensitivity for all confidence thresholds), used in calculating the performance of diagnostic tests. The model is a two-stage hierarchical model, which also permits extension to allow for study level covariates. Hellmich *et al.* (105) extend the work of Zhou to a fully Bayesian hierarchical model in a similar manner to (13.8). They use Gibbs sampling to estimate the model parameters, and pay particular attention to the issues of convergence and dependence upon assumptions regarding the prior distributions used. The hierarchical nature of this model also allows relatively straight forward extension to the case when there are covariates at the study level.

Velanovich (106) considered the case when several Bayesian probabilities (from different studies) have been determined for a given diagnostic test. The analyses use a fixed effects model to synthesise the probabilities, and apart from the fact that the probabilities have been derived using Bayes' theorem, the approach is not Bayesian in nature.

For more information on meta-analysis of diagnostic test accuracy data, see chapter 21.

Other developments

There have been a number of other developments in the field of meta-analysis utilising Bayesian methodology which do not fall into the above categories. These are briefly outlined below.

Adjustment for baseline risk

McIntosh (107) and Thompson *et al.* (108) have both considered the case when adjusting for the baseline risk in a population may be necessary. Both approaches make use of a hierarchical model and use MCMC methods to obtain parameter estimates. Whilst, Thompson *et al.* (108) consider an extension of the binary hierarchical model outlined on page 97, McIntosh (107) uses a bivariate approach to model the log odds separately in both the control and treatment groups. Both approaches, which are applied to the same data set, produce comparable results and both contrast with the naive methods (see pages 46–8) in which the relationship between baseline risk and efficacy was over-estimated.

Time-series and cross-sectional data

Osiewalski and Steel (109) consider the synthesis of evidence in the from of time series and cross-sectional data in a econometric setting.

Combination of p*-values*

Berger and Mortera (110) in a particularly technical paper consider the problem of combining studies in a meta-analysis when either only an exact *p*-value is reported or when only the level of statistical significance, denoted by 'stars' is reported.

Further research

Many of the issues listed below as areas requiring either greater dissemination or further research are not just relevant to Bayesian approaches to meta-analysis, but apply more generally to healthcare research. Indeed some of the specific points raised below will also arise out of a related NHS HTA funded project on Bayesian methods in health services research (93/50/05).

- More widespread **critical use of Bayesian methods**.
- Key need for dissemination regarding the critical use of Bayesian methods with regards to the use of MCMC, especially sensitivity analyses since these methods should not be considered as a 'black box'.
- Encouragement of **expository papers** in the applied literature on the application of Bayesian methods.
- More research on the use of **elicited prior beliefs**. This has wider implications than just the application of Bayesian methods in meta-analysis. There may be significant benefit to be had from a review of the literature on this subject, and in particular exploration of work performed in the psychology sphere.
- More research into **generalised synthesis of evidence**, especially in areas of application in which is difficult to perform RCTs.
- More research into **extrapolating the results** of a meta-analysis to clinical practice, in the same way that there should be from a single RCT.
- **Missing data** in meta-analyses has been an all too neglected area, both from a classical and Bayesian perspective. This has been partly as a result of regression models not having been widely used (with respect to missing covariate information), partly due to the fact that many meta-analyses are conducted at the study level and there missing individual patient-level data is over looked, and also partly due to the fact that, to date, many meta-analyses have been conducted in clinical areas in which clear objective measures are widely adopted, e.g. all cause–mortality. With the use of meta-analytic techniques in areas such as nursing and other professions allied to medicine, this latter point will be increasingly important.
- More research into the use of meta-analytic techniques in conjunction with **decision analysis methods**, that take into account the uncertainty associated with any meta-analysis findings.

Summary

This chapter has summarised the general use of empirical and fully Bayesian methods with respect to meta-analysis, and in particular a number of specific areas in which there has been considerable research over the last few years, and in which Bayesian methods have a potential role to play. Although currently much research is been put into these methods, so far their use in practice is far from routine. Distinct advantages of the Bayesian approach include the ability to incorporate *a priori* information which would otherwise be excluded in a classical analysis. However, when such *a priori* evidence is based on subjective beliefs the issue of whose prior beliefs to use is raised. Though many of the computational difficulties that have plagued the application of Bayesian methods in practice have been partially solved by recent development in MCMC methods, these should not be seen as 'black box' methods, since they raise issues concerning convergence.

References

1. Breslow NE. Biostatisticians and Bayes (with discussion). *Statist Sci* 1990;**5**:269–98.
2. Gilks WR, Clayton DG, Spiegelhalter D, Best NG, McNeil AJ, Sharples LD, Kirby AJ. Modelling complexity: applications of Gibbs sampling in medicine. *J R Statis Soc B* 1993;**55**:39–52.
3. Berry DA, Stangl DK. Bayesian biostatistics. New York: Marcel Dekker, 1996.
4. Bayes TR. An essay towards solving the doctrine of chances. *Philos Trans R Soc Lond* 1763;**53**:370.
5. O'Hagan A. Probability: methods and measurement. London: Chapman and Hall, 1988.
6. Phillips LD Bayesian statistics for social scientists. London: Nelson, 1973.
7. Clayton DG; Hills M. Statistical models in epidemiology. Oxford: Oxford University Press, 1993.
8. Lindley DV. Introduction to probability and statistics from a Bayesian viewpoint. Part 2 – inference. Cambridge: Cambridge University Press, 1965.
9. Berger JO. Statistical decision theory and Bayesian analysis. 2nd edn. New York: Springer-Verlag, 1985.
10. Lee PM. Bayesian statistics: an introduction. London: Edward Arnold, 1989.
11. Press SJ. Bayesian statistics: principles, models and applications. New York: Wiley, 1989.
12. Bernardo JM, Smith AFM. Bayesian theory. Chichester: John Wiley and Sons, 1993.
13. O'Hagan A. Bayesian inference. London: Edward Arnold, 1994.
14. Berry DA. Statistics: a Bayesian perspective. Belmont: Duxbury Press, 1996.
15. Spiegelhalter DJ, Freedman LS, Bernardo JM, DeGroot MH, Lindley DV, Smith AFM, editors. Bayesian approaches to clinical trials. In: Bayesian statistics. Oxford: Oxford University Press, 1988, p. 453–77.
16. Berry DA. Interim analyses in clinical trials: classical versus Bayesian approaches. *Stat Med* 1985;**4**:521–6.
17. Berry DA. Statistical inference, designing clinical trials and pharmaceutical company decisions. *Statistician* 1987;**36**:181–9.
18. Berry DA. Monitoring accumulating data in a clinical trial. *Biometrics* 1989;**45**:1197–211.
19. Berry DA. Bayesian methods in phase III trials. *Drug Information J* 1991;**25**:345–68.
20. Hughes MD. Practical reporting of Bayesian analysis of clinical trials. *Drugs Information Bull* 1991;**25**:381–93.
21. Fletcher AE, Spiegelhalter DJ, Staessen JA, Thijs L, Bulpitt C. Implications for trials in progress of publication of positive results. *Lancet* 1993;**342**:653–6.
22. Hughes MD. Reporting Bayesian analyses of clinical trials. *Stat Med* 1993;**12**:1651–63.
23. Spiegelhalter DJ, Freedman LS, Parmar MKB. Applying Bayesian thinking in drug development and clinical trials. *Stat Med* 1993;**12**:1501–11.
24. Berry DA, Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. Scientific inference and predictions, multiplicities and convincing stories: a case study in breast cancer research. In: Bayesian statistics 5. Oxford: Oxford University Press, 1994.
25. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian analysis of randomised trials (with discussion). *J R Statist Soc A* 1994;**157**:357–416.
26. Lilford RJ, Thornton JG, Braunholtz D. Clinical trials and rare diseases: a way out of a conundrum. *BMJ* 1995;**311**:1621–5.
27. Thisted RA. Elements of statistical computing – numerical computation. New York: Chapman and Hall, 1988.
28. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in practice. London: Chapman and Hall, 1996.
29. Best NG, Spiegelhalter DJ, Thomas A, Brayne CEG. Bayesian analysis of realistically complex models. *J R Statist Soc A* 1996;**159**:323–42.
30. Gelfand AE, Smith AFM. Sampling based approaches to calculating marginal densities. *J Am Statist Assoc* 1990;**85**:398–409.
31. Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *J R Statist Soc B* 1993;**55**:3–23.
32. Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian models. *Statistician* 1994;**43**:169–78.
33. Smith JQ. Decision analysis – a Bayesian approach. London: Chapman and Hall, 1985.
34. Jeffreys H. Theory of probability. Oxford: Oxford University Press, 1961.

35. Casella G. An introduction to empirical Bayes data analysis. *Am Statist* 1985;**39**:83–7.

36. Vanhouwelingen HC, Stijnen T. Monotone empirical bayes estimators based on more informative samples. *J Am Statist Assoc* 1993;**88**:1438–43.

37. Efron B. Empirical bayes methods for combining likelihoods. *J Am Statist Assoc* 1996;**91**:538–50.

38. Smith AFM. Bayes estimates in one-way and two-way models. *Biometrika* 1973;**60**:319–29.

39. Kass RE, Steffey D. Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J Am Statist Assoc* 1989;**84**:717–26.

40. Louis TA. Estimating a population of parameter values using bayes and empirical bayes methods. *J Am Statist Assoc* 1983;**79**:393–8.

41. Morris CN. Parametric empirical Bayes inference: theory and applications. *J Am Statist Assoc* 1983;**78**:47–65.

42. DuMouchel W. Berry DA, editors. Bayesian Metaanalysis. In: Statistical methodology in the pharmaceutical sciences. New York: Marcel Dekker, 1989, p. 509–29.

43. DuMouchel W. Predictive cross-validation in Bayesian meta-analysis. In: Benardo J *et al.* (eds). Proceedings of Fifth Valencia International Meeting on Bayesian Statistics, Valencia, Spain, 1995.

44. DuMouchel WH, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species (with comment). *J Am Statist Assoc* 1983;**78**:293–308.

45. Combining information across sites with hierarchical Bayesian linear models. San Francisco 1993: Proceedings of the Section on Bayesian Statistics.

46. DuMouchel, W. Hierarchical Bayes linear models for meta-analysis. Research Triangle Park, NC: National Institute of Statistical Sciences, 1994, p. 27.

47. Abrams KR, Sanso B. Approximate Bayesian inference in random effects meta-analysis. *Stat Med* 1997;**17**:201–8.

48. Verdinelli I, Andrews K, Detre K, *et al.* The Bayesian approach to meta-analysis: a case study. Department of Statistics, Carnegie Mellon University, 1995.

49. Box GEP, Tiao GC. Bayesian inference in statistical analysis. Massachusetts: Addison-Wesley, 1973.

50. Raiffa H, Schlaifer R. Applied statistical decision theory. Boston: Harvard Business School, 1961.

51. Smith TC, Abrams KR, Jones DR. Using hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. Department of Epidemiology and Public Health Technical Report, University of Leicester, 1995.

52. Smith TC, Abrams KR, Jones DR. Assessment of prior distributions and model parameterisation in hierarchical models for the generalised synthesis of evidence. Department of Epidemiology and Public Health. University of Leicester, Report 96-01, 1996.

53. Carlin JB. Meta-analysis for 2 x 2 tables: a Bayesian approach. *Stat Med* 1992;**11**:141–58.

54. Consonni G, Veronese P. A Bayesian method for combining results from several binomial experiments. Studi Statistic 40, L. Bocconi University, Milan, Italy, 1999.

55. Skene AM, Wakefield JC. Hierarchical models for multicentre binary response studies. *Stat Med* 1990;**9**:919–29.

56. Rogatko A. Bayesian-approach for metaanalysis of controlled clinical-trials. *Commun Statist Theory Methods* 1992;**21**:1441–62.

57. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* 1995;**14**:2685–99.

58. Smith TC, Spiegelhalter D, Parmar MKB. Bayesian meta-analysis of randomized triles using graphical models and BUGS. In: Bayesian biostatistics, 1995, p. 411–27.

59. Waclawiw MA, Liang KY. Empirical Bayes estimation and inference for the random effects model with binary response. *Stat Med* 1994;**13**:541–51.

60. Higgins JPT, Whitehead A. Borrowing strength from external trials in a metaanalysis. *Stat Med* 1996;**15**:2733–49.

61. Spiegelhalter D, Thomas A, Gilks W. BUGS examples 0.30.1. Cambridge: MRC Biostatistics Unit, 1994.

62. Li Z. A multiplicative random effects model for meta-analysis with application to estimation of admixture component. *Biometrics* 1995;**51**:864–73.

63. Raudenbush SW, Bryk AS. Empirical Bayes meta-analysis. *J Educ Statist* 1985;**10**:75–98.

64. Zhou XH. Empirical Bayes combination of estimated areas under ROC curves using estimating equations. *Med Decis Making* 1996;**16**:24–8.

65. Morris CN. Hierachical models for combining information and for meta-analysis. *Bayesian Statistics* 1992;**4**:321–44.

66. Stijnen T, Van Houwelingen JC. Empirical Bayes methods in clinical trials meta-analysis. *Biomet J* 1990;**32**:335–46.

67. Biggerstaff BJ, Tweedie RL, Mengersen KL. Passive smoking in the workplace: classical and Bayesian meta-analyses. *Int Arch Occup Environ Health* 1994;**66**:269–77.

68. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177–88.

69. National Research Council. Combining information: statistical issues and opportunities for research. Washington DC: National Academy Press, 1992.

70. Laird N, Louis TA. Empirical Bayes CIs for a series of related experiments. *Biometrics* 1989;**45**:481–95.

71. Spiegelhalter DJ, Goldstein H. League tables and their limitations – statistical issues in comparisons of institutional performance. *J R Statist Soc A* 1996;**159**:385–409.

72. Tweedie RL, Scott DJ, Biggerstaff BJ, Mengersen KL. Bayesian meta-analysis, with application to studies of ETS and lung cancer. *Lung Cancer* 1996;**14**:S171–94.

73. DerSimonian R. Meta-analysis in the design and monitoring of clinical trials. *Stat Med* 1996;**15**:1237–52.

74. Louis TA, Zelterman D. Cooper H, Hedges LV, editors. Bayesian approaches to research synthesis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 411–22.

75. Su XY, Po ALW. Combining event rates from clinical trials: comparison of Bayesian and classical methods. *Ann Pharmacother* 1996;**30**:460–5.

76. Smith DD, Givens GH, Tweedie RL. Adjustment for publication and quality bias in Bayesian meta-analysis. Technical Report, University of Colorado, 1997.

77. Abrams KR, Sansó B. Model discrimination in meta-analysis – a Bayesian perspective. Department of Epidemiology and Public Health, University of Leicester, 95-03, 1995.

78. Kass RE, Raftery AE. Bayes factors. *J Am Statist Assoc* 1995;**90**:773–95.

79. Arjas E, Liu L. Non-parametric approach to hazard regression: a case study with a large number of missing covariate values. *Stat Med* 1996;**15**:1771–8.

80. Kong A, Liu JS, Wong WH. Sequential imputations and Bayesian missing data problems. *J Am Statist Assoc* 1994;**89**:278–88.

81. Lambert PC, Abrams KR, Sanso B, *et al.* Synthesis of incomplete data using Bayesian hierarchical models: an illustration based on data describing survival from neuroblastoma. Department of Epidemiology and Public Health, University of Leicester, Tech. Report 97-03, 1997.

82. Givens GH, Smith DD, Tweedie RL. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. Unpublished data, 1999.

83. Tweedie RL. Assessing sensitivity to data problems in epidemiological meta-analyses. Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA, 1997.

84. LaFleur B, Taylor S, Smith DD, Tweedie RL. Bayesian assessment of publication bias in meta-analyses of cervical cancer and oral contraceptives. Unpublished data, 1999.

85. Larose DT, Dey DK. Modelling publication bias using weighted distributions in a Bayesian framework. #95-02, 1995.

86. Paul NL. Non-parametric classes of weight functions to model publication bias. Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA, #622, 1995.

87. Paul NL. Hierarchical selection models with applications in meta-analysis. Department of Statistics, Carnegie Mellon University, #621, 1995.

88. Gleser LJ, Olkin I. Models for estimating the number of unpublished studies. Department of Statistics, Stanford University, #313, 1995.

89. Eberly LE, Casella G. Estimating the number of unseen studies. BUM #1308-MA, 1996.

90. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials: builds evidence for exemplary medical care. *J Clin Epidemiol* 1995;**48**:45–57.

91. Schmid CH, Cappelleri JC, Lau J. Applying bayesian meta-regression to the study of thrombolytic therapy. *Clin Res* 1994;**42**:A290.

92. Petitti DB. Meta-analysis, decision analysis and cost-effectiveness analysis. New York: Oxford University Press, 1994.

93. Lindley DV, Smith AFM. Bayes estimates for the linear model. *J R Statist Soc B* 1972;**34**:1–41.

94. Goldstein H. Multi-level statisatical models. 2nd edn. London: Arnold, 1995.

95. Longford N. Random coefficient models. Oxford: Clarendon Press, 1993.

96. Stangl DK. Prediction and decision making using Bayesian hierarchical models. *Stat Med* 1995;**14**:2173–90.

97. Gray RJ. A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics* 1994;**50**:244–53.

98. Veronese P. Mutually compatible hierarchical priors for combining information. 1994.

99. Larose DT, Dey DK. Grouped random effects models for Bayesian meta-analysis. *Stat Med* 1997;**16**:1817–29.

100. Larose DT, Dey DK. Modeling dependent covariate subclass effects in Bayesian meta-analysis. University of Connecticut, #96-22, 1997.

101. Abrams KR, Jones DR. Meta-analysis and the synthesis of evidence. *IMA J Math Appl Med Biol* 1995;**12**:297–313.

102. Hasselblad VIC, Mccrory DC. Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making* 1995;**15**:81–96.

103. Eddy DM. The confidence profile method: a Bayesian method for assessing health technologies. *Operations Research* 1989;**37**:210–28.

104. Wolpert RL, Warren-Hicks WJ. Bayesian hierarchical logistic models for combining field and laboratory survival data. 1991.

105. Hellmich M, Jones DR, Kopcke W, *et al.* A Bayesian approach to meta-analysis of areas under ROC curves. Department of Epidemiology and Public Health. University of Leicester, Technical Report, 1997.

106. Velanovich V. Meta-analysis for combining Bayesian probabilities. *Med Hypoth* 1991;**35**:192–5.

107. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med* 1996;**15**:1713–28.

108. Thompson SG, Smith TC, Sharp SJ. Investigation underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* 1997;

109. Osiewalski J, Steel MFJ. A Bayesian-analysis of exogeneity in models pooling time-series and cross-sectional data. *J Statist Plan Inf* 1996;**50**:187–206.

110. Berger JO, Mortera J. Interpreting the stars in precise hypothesis-testing. *Int Statist Rev* 1991;**59**:337–53.

Part E:

Results IV – other important issues when combining estimates

Chapter 14

Scales of measurement

Introduction

In the previous five chapters (9 to 13), models of increasing complexity have been presented for analysing the results from different studies. Chapter 9 discussed two common scales used to report outcomes, namely the OR and the standardised mean difference, and showed how these could be combined. Not all studies will report either ORs, or a continuous outcome that can be transformed into a standardised mean difference. For this reason, it is necessary to consider several other scales that outcomes can be measured (and pooled) on. This section introduces, and defines, several other scales and gives all the formulae necessary to combine outcomes on these scales. Binary, ordinal and continuous outcomes are considered in turn. In many instances, once the SE of an estimate is known, the simple inverse variance-weighted method can be used to combine outcomes on the same scale.

Binary outcomes

On pages 56–63, details on ORs and how to combine them were given. Several other binary outcome measures are used (in varying degrees of popularity), definitions and details on how they are combined are given below.

The RR (or rate ratio)

Defining the RR

Consider the 2 × 2 table first presented in *Table 3,* reproduced below. Remember this can be constructed for studies comparing two groups using a binary outcome measure.[1]

TABLE 3

	Failure/ non-diseased	Success/ diseased
Placebo/unexposed	*a*	*b*
New treatment/ exposed	*c*	*d*

The RR of being on the new treatment, or being exposed, as opposed to being on the placebo (or old treatment), or unexposed is simply defined as the probability of an event in the treatment group (exposed group) divided by the probability of an event in the control (unexposed group). This can be calculated by:

$$\text{RR} = (a/a + b)/(c/c + d) \qquad (14.1)$$

This can be seen as a ratio of the risks in the two groups being compared.[2,3] In a clinical trial setting an RR of < 1 would imply that the experimental treatment would give benefit over the old treatment or placebo, while an RR > 1 would suggest the new treatment was inferior. In observational studies, an RR of >1 would imply the exposure under investigation is harmful, while an RR of < 1 would imply a protective effect. So, for example, if one was looking at the effect of a new drug compared to a placebo for the treatment of say, lung cancer. If an RR of 0.5 were found, this would imply the probability of death on the treatment is half that of being given placebo.

Greenland comments on combining RRs:

> 'An advantage of this method is that RRs that have been adjusted for confounding, including those adjusted by multiple regression techniques, may be combined across studies, provided that the variance of the adjusted parameter in each study is available or can be calculated' (1).

RRs are often reported for epidemiological studies; see chapter 19 for more details.

It should be noted that the Cochrane Handbook (2) discusses the relative risk reduction (RRR) = (1 – RR) as another possible scale.

[1] Note: it cannot be constructed from case–control studies.

[2] For rare diseases the RR can be approximated by the OR. This is because as b and d become large compared with a and c, $a + b \approx b$, and $c + d \approx d$, then the above formula (14.1) approximates to, $(a/b)/(c/d)$, which is the formula for the OR given on pages 56–63.

[3] Note: this measure is sometimes called the relative rate.

Combining RRs [adapted from Fleiss (3)]

The inverse-weighted variance method (see pages 55–6) can be applied to RRs. Usually the data are transformed and a combined estimate of the log RR is calculated. This is done for reasons similar to those for the OR, explained on pages 56–63. Again, taking anti-logs of the log RR estimate and its respective CI transforms the estimate back to the RR scale.

The RR is used less frequently than the OR, and no methods specifically aimed at combining RRs have been developed. Whether the above approach is efficient and unbiased in all situations is, so far, unresearched.

When using the inverse-weighted variance method, the estimate of the log of the SE for the RR is given by:

$$\mathrm{SE}(L(rr)) = \left(\frac{1 - p_1}{n_1 p_I} + \frac{1 - p_2}{n_2 p_2} \right)^{1/2} \tag{14.2}$$

where p_1 and p_2 are the observed rates of occurrence of the given event in the treatment and control groups, respectively. So $p_1 = a/a + b$ and $p_2 = c/c + d$ and the estimate of the RR = p_1/p_2. $n_1 = a + b$ and $n_2 = c + d$.

To obtain a $(1 - \alpha)100\%$ CI for the estimate of effect size, substitute the above SE formula into equation (9.3).

Rate differences between proportions

Defining the rate difference between proportions

Referring back to *Table 3* (2 × 2 table), the risk difference is defined by:[4]

$$\text{Risk difference} = (a/a + b) - (c/c + d) \tag{14.3}$$

It can be thought of as (and is sometimes called) the risk difference, as it is the difference between the probabilities of an event in the two groups (rather than the ratio used to calculate the RR.) In a clinical trial setting, a positive rate difference implies a benefit of being on the treatment, while a negative value suggests the old treatment is superior or even that the new treatment is harmful. It is not used as often as the OR or RR.

Combining the rate difference between proportions

If the difference between two proportions is to be combined, then the inverse-weighted variance method can again be used. The rate difference of a study, *i,* has a variance:

$$v_i = \frac{p_{i1}(1 - p_{i1})}{n_{i1}} + \frac{p_{i2}(1 - p_{i2})}{n_{i2}} \tag{14.4}$$

(For notation explanation see pages 109–10).

Using the formula above (14.3), a $(1 - \alpha)100\%$ CI for the estimate of effect size can be calculated with equation (9.3).

The number needed to treat

Defining the number needed to treat

The number needed to treat (NNT) is a measure that is being used increasingly when reporting the results of clinical trials. The motivation for its use is that it is more useful than the OR and the RR for clinical decision making (4). However, its role as a measure of comparative treatment effects is limited (5). The definition of the NNT is simply the reciprocal of the absolute risk reduction, put mathematically:

$$\mathrm{NNT} = \frac{1}{\text{Rate difference}} \tag{14.5}$$

$$= \frac{1}{(a/a + b) - (c/c + d)}$$

Clearly, this can be easily calculated if one has the original 2 × 2 table giving *a, b, c, d.* In a clinical trials setting (which is usually where it is used), its size can be interpreted as the number of patients that need to be treated by the experimental treatment rather than the placebo/old treatment in order to prevent one additional adverse event. For example, in testing a new drug for lung cancer against the best old alternative, a NNT of 20 would mean, on average, that for every 20 patients treated with the new treatment one less adverse event (pre-specified such as relapse of death) would be observed. A good example of the use of this scale in a systematic review is that by Tramer *et al.* (6).

Combining the NNT

Since the NNT is the reciprocal of the rate difference. The meta-analysis could be carried out using rate differences as outlined above (pages 109–10) and the pooled estimate along with its corresponding CI could be transformed to the NNT scale by simply taking the reciprocal (4).

[4] This measure is also sometimes called the absolute risk reduction.

The Phi coefficient

Defining the Phi coefficient

Up until this point all the outcomes introduced have been measures of a difference between two groups. The only other 'type' of outcome used in meta-analysis is a measure of correlation. The Phi coefficient is a special case of Pearson's product moment correlation coefficient, which is described under 'Defining ordinal data' (it may be useful to read this section first). It is used to calculate the correlation between two binary variables *(X, Y)* (which is why it is included first here). It is used most commonly in cross-sectional studies.

Consider *Table 12* below.

TABLE 12 *Observed frequencies in a study cross-classifying subjects on two binary characteristics [adapted from Fleiss (7)]*

	Y		
X	Positive	Negative	Total
Positive	n_{11}	n_{12}	n_1
Negative	n_{21}	n_{22}	n_2
Total	n_1	n_2	n

The Phi coefficient estimate is calculated by:

$$\hat{\varphi} = \frac{(n_{11} \times n_{22}) - (n_{12} \times n_{21})}{\sqrt{n_{1.} \times n_{2.} \times n_{.1} \times n_{.2}}} \qquad (14.6)$$

Combining the Phi coefficient

Estimates from separate studies can be combined using the inverse variance-weighted method, where the large sample SE of the estimate of correlation is:[5]

$$\mathrm{SE} = \frac{1}{\sqrt{n_{..}}}\left(1 - \hat{\varphi}^2 + \hat{\varphi}\left(1 + \frac{\hat{\varphi}^2}{2}\right)\frac{(p_{1\cdot} - p_{2\cdot})(p_{\cdot 1} - p_{\cdot 2})}{\sqrt{p_{1\cdot}p_{\cdot 1.}p_{2\cdot}p_{\cdot 2}}} - \frac{3}{4}\hat{\varphi}^2\left[\frac{(p_{1\cdot} - p_{2\cdot})^2}{p_{1\cdot}p_{2\cdot}} + \frac{(p_{\cdot 1} - p_{\cdot 2})^2}{p_{1\cdot}p_{2\cdot}}\right]\right)^{1/2} \qquad (14.7)$$

where p_1 is the observed rate of *Y* positives when *X* is positive etc.

An alternative to calculating this long formula is to use the jack-knife estimation technique. It is beyond the scope of this report to explain this method; however, details are given by Fleiss (7).

Ordinal outcomes

Defining ordinal data

If the outcome of interest is measured on a categorical scale and ordered in terms of desirability, such that C_1 is worst category to be in and C_m is best, then one can consider the data as ordinal (8). For example, Whitehead and Jones (8) investigated whether concurrent treatment with the synthetic prostaglandin, misoprostol, would affect the degree of gastrointestinal damage without reducing the anti-inflammatory effect of the non-steroidal anti-inflammatory drug. Some of the studies classified the number of lesions on a 1–5 scale with 1 being no visible lesions, 3 being 2–10 haemorrhages or erosions and 5 being > 25 haemorrhages or erosions or an invasive ulcer of any size.

Combining ordinal data

Two situations exist when combining ordinal data from different studies: 1) when the response variable is the same in each study, and 2) when different response variables are used in the studies. A unified framework is presented below that can incorporate both possibilities.

The approach put forward for combining ordinal data, is to reduce the outcome to binary by combining categories to produce two categories (9). Log(OR)s can then be calculated for each study and combined using methods previously described[6] (see pages 56–63). Define the treatment effect for the *i*th study as:

$$\theta_{ji} = \log\left\{\frac{Q_{jCi}(1 - Q_{jTi})}{Q_{jTi}(1 - Q_{jCi})}\right\} \qquad (14.8)$$

where $Q_{jTi} = p_{1Ti} + ... + p_{jTi}$, $Q_{jCi} = p_{1Ci} + ... + p_{jCi}$, $j = 1, \ldots, m - k$ $(k < m)$, and p_{jTi} and pj_{Ci} are the probability of a patient in the *i*th trial being in the *j*th outcome category. So, the outcome is partitioned with 1 to $m - k$ outcomes in one group and the k to m outcomes in the other.

This is taken to be the proportional odds model; it assumes the value for θ_{ji} would stay constant if the partitioning of outcomes had been at some other value (e.g. $m - 2$, $m - 5$ etc.). Thus θ_{ji} can be

[5] Fleiss states [(7), p. 249] Bishop, Fienberg, and Holland ascribe this formula to Yule.

[6] For a comparison of how they perform see (8).

considered as the log odds of success on the experimental treatment relative to control for the ith study irrespective of how the ordered categories might be divided into success or failure (8). This analysis has the following advantages: 1) definitions of each category in each study are not crucial 2) no studies need to be omitted from the meta-analysis because of differences in the scoring system, and 3) the data can be used in their original form.

For computational details, along with an example, see (8).[7]

The proportional odds assumption within each study may be investigated be calculating the estimates of the log-odds-ratios, θ_{ji}, from the various binary splits. These estimates should be similar. It would appear that this method can only be used for fixed effect estimates, more research is required for a random effects method to combine ordinal data.

Continuous data

Combining data in their original metric and combining standardised effect sizes have both been covered in the fixed effect chapter (pages 56–66). An effect used less often is the correlation coefficient, this is discussed below.

The product–moment correlation coefficient

Defining the product–moment correlation coefficient

Correlation coefficients measure the association between two variables. Pearson's product–moment correlation coefficient measures the linear relationship between two metric variables.[8] A correlation coefficient is said to be positive if the value for the one variable increases as the other increases, and is said to be negative if the value for the one variable increases as the other decreases (10). A correlation is defined on the scale of –1 to 1. If the correlation is 1 or –1 the two variables are said to have a perfect relationship, and the value for one variable can be predicted without error from the other. A correlation coefficient of 0 indicates no relationship exists between variables. Correlation coefficients are probably used more commonly as an outcome in psychology and educational studies, but examples do exist in the medical literature. For example, the study by Welten *et al.* (11) investigated the association between calcium intake and bone mass density in young and middle aged males and females using correlation coefficients. The formula for calculating Pearson's product–moment correlation coefficient (often notated as r) is given below:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \qquad (14.9)$$

where n pairs of observations, x and y, are observed; and $\bar{x}$ and $\bar{y}$ are the means values of the observations.

Combining product–moment correlation coefficients [adapted from Fleiss (3)]

It has been recommended to first transform the estimates using Fisher's variance stabilising z-transform.[9] Transforming the correlation estimates from each of i studies to be combined is achieved using the equation below:

$$Z_i = \frac{1}{2}\ln\left[\frac{1 + r_i}{1 - r_i}\right] \qquad (14.10)$$

The corresponding underlying correlation parameter for large samples is approximately normal, with mean

$$\zeta_i = \frac{1}{2}\ln\left[\frac{1 + \rho_i}{1 - \rho_i}\right] \qquad (14.11)$$

The variance of Z_i is given by

$$\mathrm{Var}(Z_i) = \frac{1}{n_i - 3} \qquad (14.12)$$

The weight associated with Z_i is thus simply $n_i - 3$. Once estimates have been combined, the result and CI can be transformed back using

[7] Whitehead and Jones (8) note, one can use PROC LOGISTIC in SAS for this analysis. Covariate adjustment for prognostic factors pertaining to individual patients can be made within studies using this procedure. Also has a test score for the proportional odds assumption.

[8] Note: It has also been used as an index of effect magnitude (13).

[9] This is due to the fact that as the population value of r gets further and further from zero, the distribution of the rs sampled from that population becomes more and more skewed (16).

$$r = \frac{e^{2Z} - 1}{e^{2Z} + 1} \qquad (14.13)$$

Correlation coefficients can also be combined directly. Hunter and Schmidt (12) state that the average *z*-transform is positively biased; so they prefer combining without transforming.[10]

Assuming the underling data are bivariate normal,[11] the variance of each r_i is given by (14.14), and estimates are combined again using the inverse variance-weighted method (chapter 9).

$$v_i = (1 - r_i^2)^2/(n_i - 1) \qquad (14.14)$$

A third approach has been proposed; this is to find a MLE for ρ, the underlying correlation parameter for all studies being combined (i.e. assuming a fixed effect). Two approximations to the MLE are given in (13).

Analysis of specific data types

Below are two types of data for which standard models are not appropriate.

Analysis of counts (rates) [adapted from Hassleblad (14)]

This section is concerned with methods used if the results to be combined are in the form of events per person-year, or some other measure of time (see *Table 13*). Often the measure of interest is the ratio of two of these rates, or some combined measure of these ratios may be desired. Standard methods for 2 × 2 tables are not appropriate for this particular outcome. If the assumption is made that the counts (number of cases) follow a Poisson distribution, an estimate of the log of the effect size is:

$$\log(\bar{T}.) = \log(A) + \log(T) - \log(B) - \log(S)$$

TABLE 13

	Treated	Not treated
Number of cases	A	B
Period of time	S	T

An approximate estimate of the variance is given by:

$$\mathrm{Var}[\log(\bar{T}.)] \approx \frac{1}{A + 1/2} + \frac{1}{B + 1/2} \qquad (26.1)$$

The individual estimates can then be combined using the inverse variance-weighting method (pages 55–6). If a more accurate variance estimate is required more advanced methods are given in (14).

Analysis of rare outcomes

If the outcome in either or both arms of a study is rare, so there are near or zero counts, standard methods may be problematic. Combining ORs and RRs in these instances may lead to spurious results. If the difference in rates is used as the measure, these 0s are informative about the difference being small. Because the normal approximation to likelihood function will not work for very small counts, it is necessary to compute the exact likelihood function for the difference (14). The likelihood function is given in Hasselblad (14) and can be solved via the confidence profile (page 200) or Bayesian approach (see chapter 13).

Further research

Investigation into when to use the different measurement scales, and a study of instances when you get different answers using different scales. (Recently, Deeks *et al.* (15) carried out an investigation into differences observed when using ORs or RRs to combine studies using over 2000 syntheses form the Cochrane Database of Systematic Reviews. In some instances considerable differences between the results of the two analyses were observed.)

- Criteria for selecting between the different scales: a) from a statistical perspective, and b) from a clinical perspective. The Cochrane Handbook comments: 'The choice of which measure to use is not always straightforward, and the measure used in the analysis may not be the best measure for clinical decision-making.' (2)
- A random effects model for combining ordinal data.

[10] However, Shadish and Haddock state:

> 'Few statisticians would advocate the use of untransformed correlations unless sample sizes are very large because standard errors, CIs, and homogeneity tests can be quite different.' (17).

[11] However, these methods should be robust (17).

Summary

This chapter presents other scales commonly used when assessing outcomes in medical research. One needs to be aware that scales other than ORs and standardised mean differences exist and can be used to combine studies. Additionally, it is important to note that since different studies may report outcomes on different scales then it may be necessary to transform a proportion of them before synthesis can proceed. Methods for doing this are presented in the next chapter.

References

1. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.

2. Oxman AD, editor. The Cochrane Collaboration handbook: preparing and maintaining systematic reviews. Second edn. Oxford: Cochrane Collaboration, 1996.

3. Fleiss JL. The statistical basis of meta-analysis (review). *Stat Methods Med Res* 1993;**2**:121–45.

4. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995;**310**:452–4.

5. What form should clinical guidance take? NNTs and Guidelines or CGTs (Clinical Guidance Trees). Workshop 'From Research Evidence to Recommendations Exploring the Methods', February 17, 1997.

6. Tramer M, Moore A, McQuay H. Prevention of vomiting after paediatric strabismus surgery: a systematic review using the numbers-needed-to-treat method. *Br J Anaesth* 1995;**75**:556–61.

7. Fleiss JL, Cooper H, Hedges LV, editors. Measures of effect size for categorical data. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 245–60.

8. Whitehead A, Jones NMB. A meta-analysis of clinical trials involving different classifications of response into ordered categories. *Stat Med* 1994;**13**:2503–15.

9. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. *Stat Med* 1991;**10**:1665–77.

10. Meinert CL. Clinical trials dictionary: terminology and usage recommendations. Baltimore, Maryland: The Johns Hopkins Center for Clinical Trials, 1996.

11. Welten DC, Kemper HC, Post GB, van Staveren WA. A meta-analysis of the effect of calcium intake on bone mass in young and middle aged females and males. *J Nutr* 1995;**125**:2802–13.

12. Hunter JE, Schmit FL. Methods of meta-analysis: correcting error and bias in research findings. SAGE Publications, 1990.

13. Mi J. Notes on the mle of correlation coefficient in meta analysis. *Commun Statis Theory Methods* 1990;**19**:2035–52.

14. Hasselblad VIC, Mccrory DC. Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making* 1995;**15**:81–96.

15. Deeks JJ, Altman DG, Dooley G, Sackett DLS. Choosing an appropriate dichotomous effect measure for meta-analysis: empirical evidence of the appropriateness of the OR and RR. *Controlled Clin Trials* 1997;**18**:84s–5s.

16. Rosenthal R, Cooper H, Hedges LV, editors. Parametric measures of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 231–44.

17. Shadish WR, Haddock CK, Cooper H, Hedges LV, editors. Combining estimates of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 261–84.

Chapter 15

Issues concerning scales of measurement when combining data

Introduction

Chapters 9 and 14 outlined ways that categorical, ordinal and continuous data could be combined. Under each type of data several common measures were highlighted and explained. For example, under categorical data information on how to combine ORs, RRs and risk differences (as well as several other less common measures) was given. It is often the case that primary studies will present their results differently to one another, and different effect measures may well be used. However, presently meta-analysis methodology dictates that it is necessary to combine (binary) outcome data on the same scale. Before these estimates can be combined, it is necessary to transform at least some of them to a common scale (if this is possible). With limited data from only the published report, this may be a non-trivial task. In extreme cases, it may be necessary to change the type of data (e.g. from continuous to categorical) for some of the studies before data is combined. Other factors needing to be taken into account before any decision on which scale to transform the data to is taken, are: 1) will the choice of scale used effect the final result? and 2) if a choice of scales is possible for the analysis which one is most desirable. Intuitively, the answers to these questions should be no, and it does not matter, respectively, but under certain conditions this may not be the case. This chapter aims to cover these issues starting with a critique of the various scales then describing methods for transforming data.

Critique of the different scales

Binary outcomes

Below is a very brief summary of the relative merits and drawbacks of the binary outcome scales commonly used. Two different and often opposing considerations are: 1) whether it is statistically convenient to work with, and 2) whether it conveys the necessary clinically useful information (1). For a lengthier and more detailed discussion see (1,2).

Risk difference – simplicity is perhaps its only virtue. A technical difficulty is that its range of variation is limited by the magnitudes of the reported risks in groups one and two. The possible values for the difference when the risks in groups one and two are close to 0.5 are greater than when the risks in both groups are close to 0 or 1. So heterogeneity may appear solely due to these mathematical constraints imposed [(2), p. 246].

Relative risk – this scale is popular; being used in both randomised and observational studies. Its biggest drawback is that only the finite interval from 0 to 1 is available for indexing a lower risk in population 1, but the interval from 1 to infinity, is theoretically, available for indexing a higher risk in population 1. A lesser problem is in interpretation; it is non-intuitive that 1 not 0 is the value taken when there is no difference between the groups. For these reasons, inferences usually carried out on the logarithms of the RR. Also, the sampling distribution of the logarithm of RR is more nearly normal than the sampling distribution of the RR.

Two other technical problems exist with the RR: 1) if the chances are good in the two groups being compared (experimental treatment versus control or exposed versus not exposed), that a subject will experience the outcome under study, many values of the rate ratio are mathematically impossible. For example if the probability of an event in group 2 is 0.4, then only the values in the interval $0 \leq \text{RR} \leq 2.5$ are possible for the RR. This means that between-study heterogeneity will emerge in the value of the RR only because the studies differ in the probabilities of an event in group 2[1]; 2) the RR is not estimable from data collected in retrospective studies.

Number needed to treat – an advantage is that it is useful in clinical decision making. A disadvantage is that when reporting a NNT no cost information is incorporated into the results, another factor that may have large implications for clinical decision

[1] Note, this constraint does not occur when using the OR scale.

making (although this is also the case for the other measures).

Odds ratio – has several advantages over other scales:

1. Can be estimated from all types of study design.
2. It is the key parameter in a linear logistic regression model.
3. The OR estimated from a retrospective study (rate ratios are not directly estimable) will, in the case of relatively rare events and in the absence of bias, provide an excellent approximation to the rate ratio.[2,3]

Disadvantage: meaning of the OR is not intuitively clear (2) or clinically meaningful (1).

Phi coefficient – several problems exist when using the phi coefficient:

1. If the binary random variables X and Y result from dichotomising one or both of a pair of continuous random variables, the value of phi depends strongly on where the cut points are set.
2. If two studies with populations having different marginal distributions but otherwise identical conditional probability structures may have strongly unequal phi coefficients. This may give the appearance of between study heterogeneity.
3. Measure is invalid when sampling other than cross-sectional is used (2).

Transforming scales, maintaining the same data type

Different studies may have used different scales to present effect size estimates. This is a problem that needs overcoming when pooling. Under 'Binary outcome data', methods for transforming estimates from binary outcome data are discussed. Chapter 14 discussed combining ordinal data combined on different scales, within a unified framework, so this issue has already been dealt with for ordinal data. Similarly, it is common practice to work with the standardised effect size for continuous variables, when scales differ between studies (see chapter 9).

Binary outcome data

Because several different binary measures are common within the medical literature, effect estimates from the different primary studies (within the same meta-analysis) may have been reported on different scales. Thus, the issue of having to combine binary data reported on different scales is often a problem. No meta-analysis techniques are presently available to combine different scales, so to proceed one needs to transform all the estimates to a single scale. This introduces the issue of which scale should be used to combine, and hence which study estimates are to be transformed.

If 2 × 2 tables are available for all trials to be combined, then theoretically any of the measures on pages 109–11 (plus ORs) can be calculated. The choice of scale used to combine, in this instance, should take into account the properties of each scale discussed on pages 115–16; therefore, it would often seem sensible to convert data to the (log) OR scale, because of its statistical advantages. However, contrary to this, Sinclair and Bracken state:

> 'The perceived advantages of the odds ratio estimator are largely statistical ones: its sampling distribution, and its suitability in complex situations for modelling using the logistic transformation. The reciprocal relationship between the odds ratio for death and for survival may be a mathematically attractive feature of the odds ratio although we suspect it is a characteristic rarely used in clinical situations.' (1)

In addition, when considering which is superior, the OR or the RR, they conclude:

> 'It remains to be demonstrated empirically which estimator of treatment effect generally yields the more stable value when the same treatment is tested in populations at different baseline risk.' (1)

Another factor to consider, which was first mentioned on page 44, is that different scales can give different values for the test for homogeneity [for an example of this, see (2), p. 258]. For this reason, an investigation of the results by combining on different scales may be instructive. In this spirit Huque and Dubey suggested:

[2] Other theoretical advantages are given by Fleiss in (2).

[3] Note: The OR serves as an approximation to the RR in case–control studies (see chapter 19). Due to this, confusion arises and authors frequently mislabel and misinterpret ORs as RRs. The difference between the two measures will be greatest when large treatment effects are shown in trials carried out in populations at high baseline risk. If in these instances the OR is interpreted as a RR a physician would substantially overestimate the treatment effect. In addition subgroups within a meta-analysis (see pages 209–10) that have different baseline risks could have different ORs while the RR estimate stays constant [see (1), p. 887 for an example).

> 'The results by various measures for the same meta-analysis data may serve to check the robustness of the results. That is, if all results by using different treatment measures for the same data are similar then one would give more credence to the common treatment effect across studies.' (3)

If this information is not given for all trials, it may be possible to reconstruct them from other data sources within the paper (such as *p*-values), or through making contact with the original investigators. However, if this data cannot be obtained, the transformations which are possible may be restricted, and thus dictate to a greater or lesser degree the scale the trials are combined on.

The permutations for ways in which combinations of scales can be transformed is virtually endless. In addition, with the amount of additional information available varying greatly between reports it is impossible to suggest a standard set of procedures for how data should be transformed. However one of the possible transformations are given below. In addition, chapter 19 deals with observational studies, and outlines ways in which binary outcome estimates can be reconstructed/estimated from other information.

Transforming ORs to RRs

When the raw 2 × 2 table is not known, if the incidence of the event in the control group is known an RR can be calculated from this information, and the OR (1).[4]

$$\text{RR} = \frac{\text{OR}}{1 + I_c(\text{OR} - 1)} \qquad (15.1)$$

where I_c = incidence of the event in the control group.

Transformations of data involving a change of data type

In some situations, primary studies may not only have used different scales, but they have used different data types also. Firstly, a methodology for combining continuous and binary outcomes is presented. This is followed by a (non-parametric) method that can be used to combine many different outcome comparisons.

Combining summaries of binary outcomes with those of continuous outcomes

Whitehead *et al.* (4) present methodology for combining trials, some of which report continuous outcome measures, and others binary outcomes created by a dichotomy of the continuous measurement.

They illustrate their method with the original motivating example, which is, combining results from a series of perinatal trials investigating the effect of prophylactic use of oxytocics on blood loss following childbirth. In some trials the number of women in each treatment group who had a postpartum haemorrhage was known, whereas in other trials the blood loss was summarised as a mean and standard deviation.

This method is based on the work of Suissa (5), who presents a method of estimating the probability of an event when the outcome variable is continuous. It is based on the assumption of the normal distribution[5] using ML theory (and is more efficient than the binary approach applied to dichotomised data).[6]

This paper makes the important comment that when study results are measured on different scales, the ideal solution would be to obtain IPD and dichotomise the continuous outcome at the patient level (see chapter 24), and to investigate through a sensitivity analysis whether the cut-point makes a difference.

The Mann–Whitney statistic

This can be used to combine treatment effects when different scales/measures are used for the primary studies.

The following description is adapted from Colditz *et al.* (6).

For a RCT, the Mann–Whitney statistic can be used to estimate the probability that a randomly selected

[4] A nomogram for doing this conversion is included in (1).

[5] The authors comment in certain circumstances may be worth considering the logistic distribution rather than normal: 'Although the two distributions are similar, the logistic distribution has the proportional odds property, which means that the log-odds-ratio would remain constant across all cut-off values.' (4)

[6] Estimates the ln(OR) for each study by the MLE, based on a logistic regression model, using PROC LOGISTIC in SAS.

patient will perform better given the new treatment, than a randomly selected patient given the standard treatment, this can be notated *P(I>S)*. To estimate *P(I>S)* directly from data for two groups, we count the proportion of all possible comparisons between outcomes on the new treatment and on the standard that favour the new treatment. If *M* patients received the innovation and *N* patients received the standard, there would be $M \times N$ such comparisons. The same computation method can be used to compute *P(I>S)* from a survival graph.

The formula for estimating *P(I>S)* from the proportions of treatment failures is:

$$0.5 + 0.5(p_S - p_I), \qquad (15.2)$$

where p_S and p_I are proportions of treatment failures on the standard and on the new treatment, respectively.

For continuous data, we calculate *d*, the difference between a randomly selected patient's score on the innovation and a randomly selected patient's score on the standard, then *P(I>S)* is an estimate of the probability that *d* is greater than zero. *P(I>S)*, then, is computed by using, as a normal deviate, the standard score which is the difference in average scores for the innovation and the standard divided by the standard deviation of this difference.

The Mann–Whitney statistic can be calculated for many different statistical measures, for example, proportion surviving, mean change in blood pressure, and frequency of side-effects.

Miscellaneous methods

Below are a few other issues regarding scales one may need to be aware of when conducting a meta-analysis.

Interpretation of correlation coefficients

Rosenthal (7) warns of dismissing what may appear very small values for r and r^2. He relates a story of a randomised double blind experiment on the effects of aspirin in reducing heart attacks that was terminated early as it had become clear that aspirin did, in fact, prevent heart attacks. The data from this study would have produced values for r and r^2 of 0.034 and 0.0011, respectively, both very low. He comments that in biomedical research results such as these are not uncommon.

Calculating effect size estimates from *p*-values

In some (hopefully increasingly rare) instances an effect size may not be given in a published report. If the author cannot be located to provide an estimate, if a *p*-value (exact or level of significance) was provided, it may be possible to calculate, or at least provide bounds for a treatment effect estimate. Pages 148–52 suggest possible ways for doing this for binary outcomes; see also (7) for methods for continuous outcomes.

Measurement errors

Measurement error of outcome variables is inevitable. What effect does this have for meta-analysis? Lund reports (8) that it can be shown that a group difference – adjusted or unadjusted – expressed in the observed score metric will not be systematically affected by such errors, whereas a standard deviation of observed scores will increase for increasing error variance. It follows that all the standardised indexes mentioned above will decrease with decreasing reliability of the outcome variable and thus lead to an underestimation of the treatment effect. Lund discusses the case in which a covariate is used to investigate the difference between two treatments. If one treatment has a larger measurement error, a relationship, purely spurious, due to the problem outline above may be found. Lund suggests making an adjustment for this by multiplying the standard deviation of the treatment estimates by the square root of the reliability coefficient, and goes on to say the analysis could be repeated for different metrics as a form of sensitivity analysis. Additionally, it should be noted that measurement error affects all measured variables in a model, not just the outcome. For more information on measurement error in meta-analysis see (9), p. 131.

Further research

Undoubtedly, the issue of which common measurement scale to use when performing a meta-analysis is an important one, but also one which is often neglected in practice. Though this deficiency could almost certainly be partly addressed by greater dissemination regarding the need for sensitivity analyses, there also remains a lack of empirical evidence regarding the precise implication of which measurement scale to use in specific applications. Further empirical research is required to reinforce the importance of this issue, but also hopefully to make some general recommendations in specific situations.

Different scales can give different values for the test for homogeneity. For this reason an investigation of the results by combining on different scales may be instructive.

It would also be useful to study the effect of using different estimators (such as exact and approximate methods) to calculate treatment estimates and their SEs, and the impact this has on meta-analysis.

Summary

This chapter has considered some of the issues that must be considered when deciding which scales of measurement are to be used when combining data. Though there are specific statistical methods that can be employed when the studies in a meta-analysis use a variety of measurement scales, so as to produce a single unified scale of measurement, a number of issues should be considered. Firstly, different scales may lead to different results, both quantitatively and qualitatively. Secondly, the most convenient common scale, statistically, may not be the most appropriately clinically. Finally, where possible sensitivity analyses should be performed to check the inter-dependence between the quantitative result obtained and the measurement scale used.

References

1. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol* 1994;**47**:881–9.
2. Fleiss JL, Cooper H, Hedges LV, editors. Measures of effect size for categorical data. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 245–60.
3. Huque MF, Dubey SD. A metaanalysis methodology for utilizing study-level covariate-information from clinical-trials. *Commun Statist Theory Methods* 1994;**23**:377–94.
4. Whitehead A, Bailey AJ, Elbourne D. Combining summaries of binary outcomes with those of continuous outcomes in a meta-analysis. University of Reading, submitted.
5. Suissa S. Binary methods for continuous outcomes: a parametric alternative. *J Clin Epidemiol* 1991;**44**:241–8.
6. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med* 1989;**8**:441–54.
7. Rosenthal R, Cooper H, Hedges LV, editors. Parametric measures of effect size. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 231–44.
8. Lund T. Some metrical issues with meta-analysis of therapy effects. *Scand J Psychol* 1988;**29**:1–8.
9. Hedges LV, Olkin I. Statistical methods for meta-analysis. London: Academic Press, 1985.

Chapter 16
Publication bias

Introduction

Chapter 4 discussed various strategies for searching for the studies to include in a meta-analysis. It was stressed that in order to avoid a biased result all,[1] or at the very least, the majority of the relevant studies need to be identified. Unfortunately, even comprehensive searches of the literature (including grey material) and the use of other less formal methods such as personal communication may not reveal an unbiased set of studies. It has long been accepted that research with statistically significant results is more likely to be submitted and published than work with null or non-significant results (1), which leads to a preponderance of false-positive results in the literature (2). The implications of this for meta-analysis are that, even if all published studies have been identified, these may be only a subset of the studies actually carried out. Since positive results are more likely to be published than negative ones, combining only the published studies[2] uncritically may lead to an over optimistic conclusion. This is the problem known as publication bias.

Several other biases directly associated with the reporting and publication of results have been highlighted. Rosenthal [discussion of (3)] splits the bias incurred when trying to obtain study results into three distinct mechanisms: 1) publication bias (explained above) 2) retrieval bias (bias left after trying to obtain unpublished results) 3) pipeline effects (effect of waiting (or not) for unpublished studies to become published). Begg (4) and Begg and Berlin (2) comment on the effect of subjective reporting of results as a form of publication bias. They suggest that exaggerated claims based on the 'biased' opinion of the investigator(s) may effect what results are reported, and in extreme cases only significant results are included in a report while the non significant ones are omitted. This phenomenon has been studied (5) and an association demonstrated. Some additional reasons for results not been published were given recently by Givens *et al.* (6), who comment that students who leave the academic arena may not publish their PhD or MS dissertations; or studies are suppressed by those who do not wish to have results appear that are against their own vested interests, political beliefs, or funding source's interests.

Another source of possible bias is due to the duplication of reporting (publishing) results. This may occur because authors want to increase their authorship by essentially submitting the same results to different journals, or because different groups report multicentre trials based on at least part of the same data.

An additional problem is that of a language bias. It is common to limit searching for research reports only published in English. Grégoire *et al.* suggest this may lead to a 'Tower of Babel bias', with the following rationale:

> 'Authors having completed a clinical trial yielding negative results might be less confident about having it published in a large diffusion international journal written in English and would then submit it to a local journal. If these investigators work in a non-English speaking country the paper will be published in their own language in a local journal. Positive results by authors from non-English speaking countries are thus more likely to be published in English, and negative results in the investigators language.' (7)

This issue has been investigated by Grégoire *et al.* (7) and Moher *et al.* (8), who have found evidence of its existence.

Although it is acknowledged that these latter biases may be a real problem, and indeed do need careful consideration, very little has been written on them as separate issues. The rest of this chapter focuses on the general problem of publication bias based on the idea that non-significant results are less likely to be published than positive ones. It is worth being aware that the line between published and

[1] Or in exceptional cases when the number of trials is too large to be manageable, a representative sub-sample of studies.

[2] Of course, unpublished studies may have been found in a initial search – and indeed searching for such studies is one way of dealing with publication bias; however, these studies are much harder to locate (63) so the chance of not being able to identify all of them is large.

unpublished results can get blurred because often trial results presented at conferences are never published (9).

One of the largest sources of information for this chapter was Begg's chapter in *The Handbook of Research Synthesis* (4); this is recommended further reading. In addition a report in this series has recently been commissioned titled, 'Publication and other selection biases in systematic review'.

Evidence of publication bias

Many studies investigating the existence and magnitude of publication have been carried out, firstly in the social science, and later in the medical literature, using various methodologies (4). A full description of all these studies is beyond the scope of this report; for a fuller review of many of them see (2, 3). Also see (9) for a review of the problem of the non-communication of trial results.

The study by Esterbrook *et al.* (1) is noteworthy. They surveyed 487 research projects approved by the Central Oxford Research Ethics Committee between 1984 and 1987. By May 1990, 52% of these had been published. It was concluded that studies with statistically significant results were more likely to be published than those finding no difference between the study groups. They also concluded that observational studies were found to be at especially high risk of publication bias and also that larger studies are more likely to be published.

The study of Dickersin *et al.* (10) also deserves a mention. A total of 318 authors of published trials were asked if they had participated in any unpublished RCTs. Of these, 178 complete unpublished RCTs were identified of these the results of only 14% preferred the new therapy compared with 55% for the published reports by the same investigators.

Simes (11) compared alkylating agent monotherapy with combination chemotherapy in advanced ovarian cancer. Meta-analysis of only the published trials yielded a large and significant survival advantage for combination chemotherapy. This was not substantiated when all studies in the International Cancer Research Databank, an unbiased list of unpublished and published trials, were used.

The general message from these studies, and the many others like it, is that a considerable proportion of studies remain unpublished and those that are, are much more likely to have non significant results.

Non-experimental evidence also exists, Melton (12) states explicitly as editorial policy that the journal gives preference to study reports demonstrating statistical significance. However, contrary to previous thought, there is evidence suggesting author preferences are a more important cause of publication bias than editorial preferences (13,14). That is to say authors do not submit non-significant studies to journals because they believe it to be a waste of time as they will get rejected. A large percentage of the non-significant studies that are submitted do in fact get published, so, in fact, it is the authors beliefs not journal policy that prevents many from being published.

McPherson [discussion of (3)] also comments on publication bias due to editorial policy in the case of studies investigating the association between oral contraception and breast cancer. She observes that both the propensity to submit research and the propensity to accept it have changed dramatically over time. Even worse, these propensities show systematic variation not only between countries but between journals within countries. What editors and researchers do seems to depend on what they believe, and unhappily what they believe depends somewhat on publication bias.

Empirical evidence of language bias

Grégoire *et al.* (7) investigated foreign language bias in meta-analysis. The hypothesis under test was that negative results are more likely to go in smaller journals with low distribution.

They searched for and located meta-analysis with linguistic constraints (36 in all). They then identified relevant foreign papers for these meta-analyses. Of the 36 under consideration, new foreign language papers were found for four of them. When the analyses were re-done, the results changed in one of the four analyses (perversely the treatment became significant, contradictory to the hypothesis under investigation!).

Very recently, Egger *et al.* (15) also investigated language bias.

As an aside, Moher *et al.* (8) compared the completeness of reporting, design characteristics and analytical approaches of RCTs published in English with those published in French, German, Italian, and Spanish. No differences in reporting quality between trials published in English and trials published in other languages was found.

The authors suggest that this strengthens the argument for inclusion of all trial reports, irrespective of the language in which they are published, in systematic reviews.

Empirical evidence of factors associated with publication bias

Dickersin and Min (16) investigated the association between trial characteristics, findings and publication. The publication rate in trials investigated was 93%. Trials with 'significant' results were more likely to be published. No other factor was positively associated with publication. The authors conclude that even when the overall publication rate is high, publication bias remains a problem.

Dickersin *et al.* (17) investigated factors associated with the publication of research findings. They found no association with sample size, presence of a comparison group, or type of study. External funding and multiple data collection was found to be positively associated with publication. They found association with significance of results. An interesting side issue is that only six of 124 studies found, but not published, had actually been rejected for publication.

The seriousness and consequences of publication bias for meta-analysis

As already hinted at in this chapter, Easterbrook states:

> 'the most serious potential consequence of this (publication) bias would be an overestimate of treatment effects or risk-factor associations in published work, leading to inappropriate decisions about patient management or health policy.' (1)

Thus, Dawid and Dickey comment:

> 'Objective data reported in the literature cannot necessarily be accepted at face value.' (18)

When a meta-analysis has a large aggregated sample size, this problem is attenuated because the results may appear to be extremely precise and convincing, even though the observed association is entirely due to bias (2).

However, in contrast to this, Freirnan *et al.* (19) and Angell (20) consider the problem of publication bias to be exaggerated, and studies with negative results tend to be poorer in quality, weakened by small sample size and type II error, or based on tenuous hypothesis.

Simulation studies

Several simulation studies into the effect of publication bias have been carried out, as an attempt to understand better the implications. These are summarised in (3,21). Two distinctive models have been used for this purpose, namely truncated and ranked sampling, both assume an extreme form of sampling (3).

Truncated sampling applies specifically to studies in which the data were analysed by means of a significance test, such as comparative studies or studies in which some kind of association was examined. It is assumed that all studies in which the results are statistically significant are published, while studies with non-significant results remain unpublished. The bias in a single study can be determined by comparing the expected results conditional on a significant result, with the expected result in the absence of this condition. See (22–25) for individual simulations and (26,27) for model extensions. These investigations employed continuous outcomes with normally distributed errors, so they are of limited use when considering binary outcomes often used in the clinical setting.

Ranked sampling can be applied to studies in which descriptive statistics are emphasised. It is assumed that a number of similar studies have been conducted to estimate a measure of interest, and that the published study is the one which exhibits the largest estimate of the measure. The bias can be determined by examining the theoretical distribution of rank order statistics. This has been used as a framework for exploring the potential bias in published uncontrolled clinical trials of new cancer treatments, where it is known that many studies are conducted but remain unpublished (28). It is possible to speculate on how many similar studies might have been conducted, assuming that the published estimate is the largest estimate, and one can calculate the bias by making normal distribution assumptions [see (28) for further details]. An example of its use is given by Begg *et al.* (29), who compare bone marrow transplant with chemotherapy as treatments for acute leukaemia. Publication bias was demonstrated to be small for the transplant series and relatively large for the chemotherapy series, a feature that clarified the likely superiority of transplantation.

A general conclusion drawn from these simulation studies by Begg (4) is that the magnitude of bias is inversely related to sample size and positively associated with the number of concurrent studies. This implies that one should be especially concerned about publication bias in settings in which lots of small studies are being conducted.

Predictors of publication bias (i.e. factors effecting the probability a study will get published)

Beyond the significance of the treatment effect, and the size of the study, other potential factors which may effect the chances of a study being published have been put forward. Dickersin *et al.* (10) found that studies which favour the new therapy are more likely to be published. Clearly the quality of a study (see chapter 6) may effect its chance of being published; however, in a particular field, Smith (30) found the quality of unpublished studies to be superior to those that were published. Hemminki (31) showed that studies in which side effects of the new drugs had been observed were less likely than others to be published. McPherson comments that the importance of the medical question, the fashion and visibility of the treatments under study, may all be influential of chances of publication [discussion of Begg and Berlin (3)]. In addition, Begg and Berlin (3) highlight other potentially distinguishing features as being the presence or absence of randomisation, sample size, exploratory versus confirmatory studies, protocol definition, the nature of the journal, calendar time and source of funding.

Identifying publication bias

Begg reports:

> 'When the component studies have been assembled for the meta-analysis, a preliminary analysis should be undertaken to assess the chances that publication bias could be playing a role in the selection of the studies.' (4)

Various methods to aid this investigation are given below.

Correlations with known risk factors for publication bias

One can correlate the observed effect sizes with important design features of the studies that are risk factors for publication bias (see above). If an association is found, Begg (4) suggests abandoning the meta-analysis as being unreliable or focus on a subset of the studies believed to be unbiased, or least biased.[3,4] An example given (4) is when randomised and non-randomised studies are being combined, and randomisation status may appear to be associated with the effect sizes, one might choose to eliminate all the non-randomised studies from the analysis, on the grounds that they lead to less reliable data (see chapter 6).

Graphical display – the funnel plot [adapted from Begg (4)]

Sample size is the most important factor for identifying publication bias because small studies produce highly variable effect size estimates. 'Therefore, the most aberrant values that occur by chance are much farther from the true mean effect size than the aberrant values for large studies. Therefore, if selective publication causes the more extreme effect sizes to be selected for publication, regardless of the sample size, then the effect sizes from the small studies will be more extreme than those from the larger studies, leading to as induced association. It is also possible that small studies may be less likely to be published because of perceived unreliability, and so authors may feel that statistical significance is necessary to justify publication to a greater extent than for larger studies.' (4)

A plot of sample size versus effect size can be constructed. 'If no bias is present, this plot should be shaped like a funnel, with the spout pointing up – that is, with a broad spread of points for the highly variable small studies at the bottom and decreasing spread as the sample size increases.' (4) If negative studies are less likely to be published, the graph will tend to be skewed, inducing a negative correlation in the graph (32). This graph is commonly referred to as a funnel plot (33).

Further discussion of the funnel plot

The funnel plot method makes the assumption that the true effects in the various studies are unrelated to sample size. Begg and Berlin (3) suggest that this is reasonable under the fixed effect assumption, and compelling under a random effects model. They go on to comment that, although text book theory suggests that the size of future trials will be based on the expected treatment difference, and thus could be influenced by previous ones, in practice this may not usually be the case and sample sizes are determined on more pragmatic grounds. But 'It is not unlikely that the large, well-

[3] It should be noted that associations found may be purely spurious, or due to some factor other than publication bias.

[4] Recently, methods have been proposed for adjusting the analysis for publication bias. Although they are at an experimental stage, they have the potential of removing the need to exclude studies. See pages 126–32 for more details.

planned studies will more often be undertaken when there is such a positive likelihood of a treatment difference. So this may lead to a small positive association between the true effects and the sample sizes.' However, early stopping rules terminating early trials with a large treatment effect may induce a negative association. So the assumption of independence is speculative. However, Begg and Berlin believe the effects of the above will be weak, at least in the context of cancer clinical trials their paper used as an example.

Vote-counting procedures used to examine the extent of publication bias

Chapter 7 of this report discusses the use of vote-counting techniques in meta-analysis. For completeness, a pointer to a method for assessing publication bias using vote-counting methods is given here. Hedges and Olkin (34) outline the procedure for mean differences which is also given in (4), with an extension for correlation coefficients. The rationale behind the procedure is thus:

> 'When both positive and negative significant results are counted, it is possible to dispense with the requirement that the sample available is representative of all studies conducted. Instead, the requirement is that the sample of positive and negative **significant** results is representative of the population of positive and negative **significant** results. If only statistically significant findings tend to be published, this requirement is probably more realistic.' (34)

See original sources (4,34) for more details.

Statistical tests

Begg (4) suggests that the best formal test for publication bias is to use a rank correlation test based on Kendall's tau, after first standardising the effect sizes to stabilise the variances. This test is easy to calculate and is a direct statistical analogue of the funnel plot presented above. It works by examining the correlation between effect estimates and their variances, to exploit the fact that publication bias will tend to induce a correlation between the two factors and constructing the rank-ordered sample on the basis of one of them.

The formulae for the test is given below; for an explanation of the rationale behind it, see (32). Define the standardised effect sizes of the k studies to be combined to be

$$T_i^* = (T_i - \bar{T}.)/(\tilde{v}_i^*)^{1/2} \tag{16.1}$$

where

$$\bar{T}. = \left(\sum_{j=1}^{k} v_i^{-1} T_j \right) \Big/ \sum_{j=1}^{k} v_i^{-1} \tag{16.2}$$

and T_i and v_i are the estimated effect size and sampling variance from the ith study

and also where

$$\tilde{v}_i^* = v_i - \left(\sum_{j=1}^{k} v_j^{-1} \right)^{-1}, \tag{16.3}$$

is the variance of $(T_i - \bar{T}.)$.

It is then necessary to evaluate P, the number of all possible pairings in which one factor is ranked in the same order as the other, and Q, the number in which the ordering is reversed. A normalised test statistic is obtained by calculating

$$Z = (P - Q)/[k(k-1)(2k+5)/18]^{1/2} \tag{16.4}$$
[5,6,7]

This statistic is compared to the standardised normal distribution. Any effect size scale can be used as long as it is assumed distributed asymptotic normal.

Practical considerations when using this test

Begg (4) suggests using a very liberal significance level and notes that any evidence of publication bias should make us cautious about proceeding with the analysis. Begg also comments: 'For a meta-analysis with relatively small numbers of studies we should rely on an informal assessment of the funnel graph as an 'eyeball' test for bias.', indeed this test should always be considered as a 'formal procedure to complement the funnel-graph' (32). Begg and Mazumdar (32) investigated the power of the test and found it to be fairly powerful for meta-analyses with 75 component studies but it only had moderate power for meta-analyses with 25 component studies. Begg also notes:

[5] If there are tied observations, the denominator should be modified. However, the modifications are negligible unless there are substantial groups of tied observations.

[6] This test involves no modelling assumptions (but suffers from a lack of power.) An alternative test suggested (4) is based on Spearmen's ρ statistic.

[7] Begg and Mazumdar (32) gives an extension to this test to calculate over stratified subgroups.

> 'The test based on premise that if publication bias is present, it is induced by a mechanism in which studies with large effect sizes are more likely to be published. A graph of T_i^* against $\nu_i^{1/2}$ may be preferable to funnel plot because one does not have to judge the anticipated increase in spread at the base of the graph.' (4)

As has already been discussed, an alternative premise is that the decision to publish is based primarily on the *p*-value rather than the absolute value of the effect size. If the *p*-value is the only determinant of selective publication, then the test should demonstrate no correlation between effect size and sample size. In reality, the decision to publish is probably due to both, however no methodology available is sensitive to both influences. It is worth noting that the methods on pages 126–32 (adjusting analysis for publication bias) characterise publication bias as a function of the *p*-value.

Methods of estimating the magnitude of publication bias

This section presents methods for estimating the magnitude of publication bias. As well as the methods presented below, funnel graphs (pages 124–5) can also give some magnitude indication. In addition, Altman suggests calculating the pooled estimate after successive elimination of the smallest studies (see chapter 25):

> 'A plot of the pooled estimate and point of truncation may give a good idea of the relation between effect size and sample size and may indicate where the true effect lies. Because large studies dominate the calculations, the exclusion of the smaller studies may not have a large effect on the pooled estimate.' [discussion of (3)]

Assessing random effects within different sample size groupings

Begg and Berlin state:

> 'To estimate the magnitude of the bias, we can perform an analysis which allows us to estimate the distribution of random effects within different sample size groupings.' (3)

Essentially, this is a formal assessment of funnel plot idea. Since the distribution should be unrelated to sample size, by our independence premise, then any observed shifts in the distribution are likely to be due to publication bias. Two problems have been noted with this method: 1) the researcher may only have the *p*-value, but this can usually can be converted to an effect measure, and 2) the direction of a non-significant effect may not be known. To get round this second problem, one can assume the sign is unknown for all published estimates; assuming normality, this gives a 'folded normal' form (fold at 0).

Construct likelihood: data classified into *J* sample size groupings with m_j published estimates in group *j*. These groups have mean effects μ_j (j=1,…,J) and assumed common variance σ^2. The data (y_{ij}) and (v_{ij}) comprise of the observed estimates and the sampling variances (number of failures in survival-type studies), respectively:

$$L \propto \prod_{j=1}^{J} \prod_{i=1}^{m_j} \frac{1}{\sqrt{(\sigma^2 + \nu_{ij})}} \left[\phi\left(\frac{y_{ij} - \mu_j}{\sqrt{(\sigma^2 + \nu_{ij})}} \right) + \phi\left(\frac{-y_{ij} - \mu_j}{\sqrt{(\sigma^2 + \nu_{ij})}} \right) \right] \tag{16.5}$$

MLEs can be obtained numerically. Alternatively, the missing signs can be considered as 'missing data' and MLE obtained using the EM algorithm [details of this are given in (35)].[8]

An example which should make the above clearer is given by Begg and Berlin [(3) p. 436].

Adjusting meta-analysis for publication bias

The methods proposed for adjusting a meta-analysis can be split into two broad categories, namely analytic and sampling methods. Following an outline of the methods, a discussion of their relative merits follows. It should be pointed out that if one suspects publication bias exists, through the detection methods of pages 124–6, then efforts could be made to try and find these studies, before, or instead of, adjusting the analysis. One way of doing this would be to write to interested investigators, although this can be a painstaking process (36), or consulting registry of trials (11). However, if one does include data from unpublished studies, which have not passed peer review, one is at risk of lowering the quality and credibility of the data (3). Opinion seems split among researchers, whether this is a sensible thing to do (see pages 133–4).

[8] Begg and Berlin (3) go on to comments that one can use some studies as bench-mark for reliability of others, and as an approximate tool for calibrating the true strength of evidence from a particular study (using the above method), i.e. the preceding study provides estimate of average bias as a function of sample size. One could adjust for other factors such as presence or absence of randomisation etc., given sufficient data.

Analytical methods

Source augmentation (the 'file-drawer' method)

In essence, this method considers the question: 'how many new studies averaging a null result are required to bring the overall treatment effect to non-significance?' (37). It was developed by Rosenthal (37,38) and has been referred to as the 'file drawer problem', as it could be seen as estimating the number of studies filed away in researchers files without being published.

The method is based on combining the normal (z) scores corresponding to the p-values observed for each study (this is covered in more detail in chapter 7). The overall z-score can be calculated by:

$$Z = \sum_{i=1}^{k} Z_i / \sqrt{k} \qquad (16.6)$$

where k is the number of studies in the meta-analysis. This sum of z-scores is a z-score itself and is significant if $Z > Z_{1-\alpha/2}$. Now we determine the number of unpublished studies with an average observed effect of zero that there would need to be in order that the overall z-score is no longer significant. Define k_0 to be the additional number of studies required such that:

$$\sum_{i=1}^{k} Z_i / \sqrt{k + k_0} < Z_{1-\alpha/2} \qquad (16.7)$$

rearranging the above gives:

$$k_0 > -k + \left(\sum_{i=1}^{k} Z_i \right)^2 \Big/ (Z_{1-\alpha/2})^2 \qquad (16.8)$$

After k_0 is calculated, one can judge if it is realistic to assume that this many studies exist unpublished in the research domain under investigation.[9] If the answer is yes, then one must have doubts about the validity of the meta-analysis.

This test is often used as a sensitivity test once a meta-analysis has been found to give a significant result to examine the robustness of the finding.

Practical considerations when carrying out source augmentation

Begg and Berlin state: 'This method provides, at best, a very crude adjustment for publication bias' (3)

There are clearly shortcomings of the method: firstly the combining of z-scores does not directly account for the sample sizes of the studies. Secondly, the choice of zero for the average effect of the unpublished studies is arbitrary and certainly biased (3). Also, it is guesswork estimating the magnitude of unpublished studies in the area. Fourthly, the method does not adjust or deal with treatment effects. Fifthly, heterogeneities among the studies are ignored (27).[10] Lastly, the method is not influenced by the shape of the funnel graph (4). In its favour, the value k_0 is easy to calculate and easily interpretable (4). However, despite all the drawbacks this method has been used widely as a tool in meta-analysis (39).

Extensions

Several extensions and variations of Rosenthal's 'file-drawer' method described above have been presented. These are briefly outlined below.

Orwin (39) proposed a statistic analogous to Rosenthal's 'file-drawer' applicable to standardised differences between treatments (d) (i.e. continuous effect sizes – see pages 64–6), with no obvious choice of a critical value for d (unlike 0.05 used for p). Orwin comments:

> 'a researcher may have reason to believe that the file drawer studies have a nonzero mean effect size, or he or she may wish to test a range of values around zero.' (39)

Iyengar and Greenhouse (27) proposed a modification to this using a truncated distribution in the same way as for Rosenthal's method (see below). They make the observation that Orwin's scheme is more stable with respect to the choice of mean of the unreported studies than Rosenthal's scheme.

Iyengar and Greenhouse offer a modification to Rosenthal's approach. They argue:

> 'if there were publication bias in favour of studies with statistically significant findings, then the Z values for the unpublished studies would not be a sample from the standard normal distribution. Instead, they would be selected from the part of the population of studies whose significance levels exceed α, and hence, whose Z values are less then z_α.' (27)

[9] Rosenthal's book (27) provides a rough guide to help decide what is an unlikely number of studies in the file draws, but this guide does not seem to be used due to its ad hoc nature, more often ones self knowledge of the field is used instead.

[10] Although Rosenthal and Rubin state that simple heterogeneities can be addressed by stratifying studies and making file drawer computations within strata.

They present a modified formula for k_0 using a truncated normal density (27) (see paper for details), which always gives a smaller estimate for k_0.

However, Rosenthal and Rubin [comment on (27)] pointed that their modification is one-tailed, which assumes only results significant in one direction are published, while Rosenthal's test is two-tailed. They argue that one-tailed is less realistic because: a) early in the history of a research domain results in either direction are important news; b) later in the history of the domain, when the preponderance of the evidence has supported one direction, significant reversals are often more important news than are further replications.

Iyengar and Greenhouse (27) also presented a method for calculating k_0 using Fisher's method of combining *p*-values (see chapter 7).

Klein *et al.* (43) produced a modification of the file drawer method so that the OR scale can be used (instead of the *p*-value). As before, assume *k* published trials, and *m* unpublished trials, which on average show no treatment effect, i.e.

$$\sum_{j=1}^{m} \ln(\widehat{\mathrm{OR}}_j) = 0$$

Suppose that the pooled analysis of the *k* published studies gives a statistically significant result at the 5% level, where the weight for each study is given by $W_i = 1/V_i$. Then, the number of unpublished null trials (of similar weight to the published studies) necessary to reverse this result and render the conclusion statistically insignificant at the 5% level is the smallest integer, *m*, greater than

$$\left(\frac{k\ln(\widehat{\mathrm{OR}})}{1.96}\right)^2 \bar{w} - k \qquad (16.9)$$

where $\bar{w}$ is the average weight of the *k* published studies. Klein (40) notes similar results can be obtained using other choices of weights (see pages 56–63).

Selection models using weighted distribution theory

The purpose of these methods is to model the (publication) bias through the use of weighted distributions. These methods are more complex than the file drawer method given in the previous sections. For a brief history of weighting functions see (27); they were first introduced into meta-analysis by Iyengar and Greenhouse (27). The premise for weighted functions in meta-analysis is that each study is included in the analysis with a probability that is determined by the outcome (in all the below cases this is the observed *p*-value, rather than the effect). These selection probabilities are related to different possible outcomes via a weight function (4). Thus an adjusted effect size estimate, adjusted for the fact that the studies obtained were a bias sample, can be calculated.

General formula for model incorporating the weight function

Presented below is the general framework for the model used (4):

$$g(T;\theta) = \frac{f(T;\theta)\,w(t)}{A(\theta)} \qquad (16.10)$$

where *T* is the observed effect size, θ is the true mean effect size, $f(T;\theta)$ is the probability density function of *T* irrespective of whether or not the study was published, $w(T)$ is the weight function, $g(T;\theta)$ is the probability density of *T* given that the study is published, and where

$$A(\theta) = \int_{-\infty}^{\infty} f(t:\theta)\,w(t)\,\mathrm{d}t \qquad (16.11)$$

Thus, we are interested in the true distribution of *T*; $f(T;\theta)$ and inferences can be made by constructing a likelihood function and solving it numerically (specialist software is required).

Specifying weight functions

The simplest version was first presented by Hedges (26), following work by Lane and Dunlap (22); it deals with continuous standardised effect sizes. The weight function is simply given the value one if the test is significant (at the 5% or any other specified level), and 0 otherwise. Put mathematically:

$$w_i(T_i) = \begin{cases} 1 \text{ if } T_i > \mathrm{C}_\alpha(v_i) \\ 0 \text{ if } T_i < \mathrm{C}_\alpha(v_i) \end{cases} \qquad (16.12)$$

where $\mathrm{C}_\alpha(v_i)$ is the critical value of the α-level test for the *i*th study and v_i is the SE of the *i*th effect size. This model assumes all significant studies are published and all non significant ones are not. Once a likelihood equation is formed using equation (16.10), it can be solved using computational iteration (26).[11,12] For a technical coverage publication selection models see Hedges (41). Begg and Berlin comment:

> (it is an) 'appealing method for assessing the potential magnitude of publication bias, especially when most or all the published studies are significant, since in this case there is not much wasted data, although clearly if most of the studies were non-significant the method would be inappropriate.' (3)

The method is strongly dependent on the nature of the distribution of the *p*-values in the range 0.00–0.05, and accuracy is open to question. As the assumption made is that all significant studies are published; marginally significant ones may not be published so bias would be underestimated. They suggest it would be interesting to investigate study properties further, especially the impact of ignoring the available non-significant publications.

Rosenthal comments [discussion of (27)] this method assumes the non-published results mean effect to be 0, which is probably too simple. There is evidence to suggest it pulls in the direction of the mean of the published studies; the MLE approach of Iyengar and Greenhouse (27) addresses this observation by trying to estimate the mean effect size in the population (see next section).

Weight function of Iyengar and Greenhouse

Iyengar and Greenhouse (27) give two different variations of weight function families. Both consider all studies statistically significant at the 0.05 level will be published, and hence the weight functions will take the value one over these values. For non significant results one weight function considers the reporting probability as constant, but not zero as for the Hedges model (see above). The other suggests that the reporting probability increases (exponentially) as the outcome approaches statistical significance [see (27), p. 113 for details]. The likelihood created when this weight function is combined using equation (16.10) is solved using ML methods. The authors comment that this method is flexible, and one can apply sensitivity by varying the assumptions and examining the log likelihood surface, which shows how informative the data are about the parameters in the model.

Laird *et al.* (27) commented that symmetry in these models, and uniform weight in the tails, may not be completely realistic, and illustrates some results to substantiate their claims. Along a similar line, McPherson [discussion of (3)] comments she would like to impose asymmetry on the publication criteria. Significant results in the expected direction will have a different impact than significant results in the opposite direction (i.e. beneficial effects of new therapies are more likely to be published than ones with deleterious effects). Laird *et al.* also present formula to estimate the number of unpublished studies [see (27), p. 128 for details]. Hedges commented on above weight functions [discussion of (27)], saying that the modelling of significant results with the probability of one is unrealistic. He argued that:

> 'when *p*-values are either very small or very large, the decision whether to report or publish is based primarily on other factors than the *p*-value. When the *p*-value is intermediate, the decision to publish may be greatly influenced by the *p*-value.' (27)

He went on to suggest an s-shaped curve may be better [see (27), comment, p. 118 for formulae].

Hedges also went on to say that since both his model and the more realistic ones of Iyengar and Greenhouse assumed a fixed effect model:

> 'It would be interesting and relatively straightforward to study the effects of publication bias on estimates in random effects models' using models such as the two presented above. He notes that 'Further work to elucidate the effects of selection on estimates of the distribution of treatment effects would be an important contribution.' [comment (27)]

[11] Hedges (26) also proposed simpler solutions in view of the ad hoc nature of the procedure by providing tables which give solutions for individual components of equation (16.12). These can then be combined using a weighted average to obtain a solution.

[12] Alternatively, Hedges also presented a simpler procedure (26). It assumes all the studies have equal sample sizes and use vote counting methodology (see pages 33–4). It treats positive and negative results as independent realisations of a Bernoulli process and the adjusted treatment effect can be estimated using a modification of binomial theory [for computational details see (64) or (26)]. This method is clearly very limited as it requires a fairly large number of studies and equal sample size; because of this Hedges comments:

> 'Because unequal sample sizes are the rule rather than the exception in research synthesis, counting estimators are likely to be most useful for providing quick approximate estimates rather than serving as the analytic tool for final analysis.' (26)

Weight function of Champney

Champney (24) did investigate estimation in a random effects analysis using a simple step weight function and assumed the random effects to be normally distributed. This work suggests that publication bias may have substantial effects on estimation of the between study variance even when the estimate of the men is not strongly affected [(27) Hedges comment, p. 119].

Hedges weight function

Hedges later went on to present a more generalised form of the weight function (21). This model allowed the weight function to take different values in different regions of the *p*-value scale, thus turning it into a step function[13] [details omitted; see (21) for details]. This allows for reasoning along the lines that: a study with a *p*-value of 0.01 is more likely to be published than a study with a *p*-value of 0.05 which in turn is more likely to be published than a study with a *p*-value of 0.10, etc. The discontinuities were decided using information from psychological studies; however, plotting the observed distribution of *p*-values may provide insight about the likely shape of the weight function.[14] The model also allows the inclusion of a random effects term. Also presented are tests for publication bias by assessing if all the weights are equal to one, one based on the χ^2 statistic and two based on likelihood ratio tests. The author appears cautious about this method suggesting it should be used to give a 'broad indication of whether selection is operating'. (21)

The weight functions of Dear and Begg

Dear and Begg (42) suggested that the problem with the above weight functions is their monotonicity; in addition to their lack of flexibility for accommodating different shapes of selection functions. They presented an approach which allows the shape of the weight function to vary in as unconstrained a manner as possible, using a semi-parametric model. This model also has the ability to incorporate random effects for treatment effect. Thus, the main distinction between this approach and that of Hedges outlined above is that Hedges chooses to pre-specify the regions of the *p*-value scale within which the weight function, and are assumed constant (0.05, 0.01, 0.001 etc.). The authors comment, 'In practice this will lead typically to a weight function with fewer 'steps', and as a result Hedges' method is probably more robust but less flexible than (this method). .....Research is clearly needed in assessing and comparing the operating characteristics of these two methods. However our intuition suggests that the Hedges model will be more suitable for meta-analyses with substantial numbers of component studies, while our method will be necessary for small meta-analyses'. [See original paper (42) for formula of step function.] The method is complementary to the traditional funnel graph, but sensitive to publication bias even when the study sample sizes are similar (unlike the funnel plot). It can be employed in the context of either one- or two-sided tests (depending whether or not the weight function can be assumed to be the same for negative and positive effect sizes of similar magnitude) (4). The authors stresses this model should be used as an exploratory, informal tool and suggest that it can be used to correct estimates for bias, preferable would be to focus attention on the causes of bias.

Recent developments in weight functions

Recently, new weight functions have been proposed. Paul (43) has investigated a non-parametric class of weight function within a Bayesian analysis (chapter 13). Motivation for this work is to consider the robustness of results on the choice of weight function, which after all is specified by the user and unknown. This approach involves specifying a weight function and a neighbourhood around it and looking at the range of results over the neighbourhood.

Larose and Dey (44) fit weight functions of Iyengar and Greenhouse (27) and Patil and Taillie (45)[15] from a Bayesian perspective (see chapter 13) using non-informative priors. Several model selection criteria are used in the spirit of exploratory data analysis for the appropriate choice of weight function.

Final remarks on weight functions

Many different approaches to weight functions have been outlined here. Unfortunately, there are no software packages that can do this sort of analysis routinely yet (4). Another point worth noting is that Laird *et al.* [(27), p. 126] have

[13] The likelihood can be solved via the Newton–Raphson method or the EM algorithm.

[14] Hedges suggests searching through registries to find unpublished studies and their corresponding p-values to get and idea of the distribution and apply it to the weight function (27).

15 Due to time limitations, these particular weight models are not discussed in this report, but the interested reader should note their existence. The following references may also be of interest (53–56).

commented on the similarity of these methods to survey sample theory for missing data: 'Since the sample survey literature on handling non-response is extensive; we feel that many of the approaches developed for sample survey can be used with advantage in the meta-analysis setting. (see original paper for more details). This may be a methodology whose potential is not fully realised yet.'

Miscellaneous methods

Estimating an unpublished study

Sugita *et al.* (46) claim to present a method of estimating a pooled OR whilst eliminating publication bias. They make the assumptions that log ORs from each study are distributed normally, and then calculate an estimate and CI employing a moment method for the unpublished studies. Then by solving a set of simultaneous equations, which contain the sample second, third, and fourth sample moments they can estimate the summarised OR and CI in all studies, including not only those published but also those unpublished. This method assumes homogeneity and thus a fixed effect analysis. A further drawback not pointed out in the paper is that they assume only one study is not published and it is an estimate of this which is calculated. The authors comment this method can be used with the hazard ratio. This model was later revised by Sugita *et al.* (47) which allowed the probability density function curve of all studies to be drawn.

Analysing the largest studies

Begg (3) gives some general advice. First, explore the apparent association between the measure of interest and sample size, to identify **evidence** of publication bias. If a strong trend is present, large studies should be less biased than the small ones so it may be advisable to eliminate the small ones. This would seem quite radical, but not without good reasoning, perhaps further investigation is needed here. This method would be useful as a form of sensitivity analysis.

Other recent developments in adjusting the analysis for publication bias

Clearly, this is a very active area of interest. At the time of writing this report, there were several unpublished pieces of work on the subject. A brief overview of these is given below.

Methodology that attempts to estimate the number of missing studies has been developed (48,49), though not formally published. Glesser and Olkin (49) present two general methods. Each model allows one to estimate the number, *N*, of unpublished studies using the *p*-values reported in the published studies. *N* and its confidence bounds can then be evaluated for plausibility by the meta-analyst. The authors comments that Begg and Berlin (3) and Iyengar and Greenhouse (27) have emphasised that Rosenthal's fail-safe *N* approach cannot be universally applied. 'At least some specification of the mechanism that consigns studies to file drawers is necessary to justify the method. (49)[16]

The first model considers the possibility that the *p*-values observed are the *k* smallest *p*-values among the $N + k$ reported and unreported studies.[17] ML and best unbiased point estimators of *N* are presented along with a lower $100(1 - \alpha)\%$ confidence bound for *N* assuming the null hypothesis is true. A more realistic modification is then presented, in which the m smallest *p*-values plus a random sample of $k - m$ of the $N + k - m$ remaining *p*-values are observed.

The second type of model is a true selection model in the form of (3) and (42) presented above.[18] Here, if the function is totally unknown, it is shown that *N* is not definable. If, however, an interval is known for which studies will be reported a method for estimating *N* is presented. See (49) for more details.

Eberly and Casella (48) also present a model estimating the total number of studies carried out, both published and unpublished, dependent on the probability of publication. A selection model is again used where all studies significant at level α are published, while non-significant studies are published with probability *p*. Here Metropolis simulation and Gibbs sampling techniques are used (50) to generate random samples from the distribution of the total number of studies and study how it changes as *p* varies.

Very recently, Givens *et al.* (6) proposed a method to estimate and adjust for publication bias. In this approach, the number of studies missing along with

[16] The paper also comments that the 'fail-safe *N*' has no necessary relation to the actual number, *N*, or reported studies estimated by this method.

[17] Paper notes this is not a selection model, rather, it resembles observational models used in accelerated life-testing, where sampling ceases once *k* lifetimes (which are necessarily the smallest lifetimes) are observed.

[18] i.e. the probability that a study is reported is a function $g(p)$, of the attained *p*-value.

their results are estimated and imputed into the dataset containing the published study results. A Bayesian approach is used (see chapter 13), and estimation is based on a data augmentation principle within a hierarchical model. Any of the publication selection mechanisms described above [such as those of Hedges (21) and Dear and Begg (42)] can be used within this methodology. For an application of this method, see (51).[19] The authors go on to comment that if one has approximately 30 studies, 'one might have some confidence that the imputed studies give a credible representation of the truth. What remains to be developed is a method of handling small collections.' (6)

This methodology has been extended even further (52). If one believes that the quality of a study is a factor determining whether it is published, as well as the significance of the results, then different selection mechanisms can be applied to the studies of differing quality.[20] This is done by adding a hierarchical structure to the model of Givens *et al.* (6). Smith *et al.* (52) point out that this model, in principle, can be extended to include any other covariates one believes have an influence on the probability of publication.

Other work in this area include that of Bayarri and DeGroot (53,54), who explore the behaviour of published results using an indicator function of statistical significance to weight the model's likelihood, and show that significant overall results obtained from published data actually can be strongly supportive of the null hypothesis.

Also see (55), and Fongillo (56), who takes a Bayesian approach and uses two-stage hierarchical models to model variability both within and between studies.

Clearly, this is an area which has seen fast development over the last few years, and since several groups are currently working on the problem it is envisaged that new developments will continue to happen. It remains to be seen how useful and appropriate these methods are in practice, they should certainly be considered as experimental methodology.

Invariant sample frames

It is often desirable to include as much information as possible when carrying out a meta-analysis, as we have seen, including only studies found through literature searches and other means may produce a bias sample of the studies that were actually carried out. Combining this biased sample can produce misleading results. A method has been proposed by Simes (11,57), which removes the possibility of publication bias at the cost of potentially only including a selected proportion of all studies carried out. A brief description of the method is given below.

This method reported in (3) involves limiting the meta-analysis to a subset of studies which satisfy the condition that they represent an exhaustive collection from a sampling frame which is independent of the publication process. Explicitly, the idea proposed by Simes was to restrict attention to studies that prospectively would all have been registered at some international trials registry. The meta-analysist would then following up all studies registered, including (a cohort of) unpublished studies which would have been identified by the international register. It is important to note that studies published, but not on the register, are ignored and not included in the meta-analysis. This method would produce a list of trials which would not be influenced by study results (57).

Simes (11,57) give examples of where the above methodology is used in cancer RCTs. Sampling frames such as the International Cancer Research Bank using the database CLINPROT were used in these instances. It is worth noting that the results obtained by this method differed from those obtained from the usual method of selecting trials through literature searches etc.

A further advantage of this method suggested by Simes (11), is that trials not registered may enhance bias, i.e. registration may eliminate the more poorly designed and less well controlled studies. However, unfortunately for most topics, registers do not exist. In 1988, Begg and Berlin (3) considered their construction to be seen as a policy goal. In 1991, Easterbrook did assemble a compendium of existing registries (58), but the methods are restricted to certain subjects. Begg and Berlin (3) highlighted the need for alternative

[19] This method can be seen as a form of imputation of missing data (see chapter 17).

[20] It is worth making explicitly clear that no adjustment to take into account study quality, such as weighting (see chapter 6), is carried out via this analysis. The variation in quality only effects the number and estimates of the studies imputed.

sampling frames and suggested creating registries from studies approved by hospital ethics committees as a possibility (though they may be hesitant to give out information as it is often given them in confidence). A further alternative is Government lists, though these are confidential except in USA and Spain (9). Another problem is that although one knows of the unpublished studies of interest, this does not mean their results are easily obtainable. The above discussion assumes all evidence to be combined is coming from RCTs, in practice other evidence may also be included (see chapters 19 and 26), these sorts of data are susceptible (and probably mores so than RCTs) to publication bias. For example [(3), discussion], Day comments more and more audit research is done that never goes through registries or ethics committees.

Broader perspective solutions to publication bias

The problem of publication bias is a fundamental one: Weisberg comments:

> 'the meta-analytic approach is severely limited by the conventional form in which research results are presently conveyed. Traditional methods for summarizing data are not well suited to the needs of meta-analytic reviewers. The current paradigm provides little opportunity or incentive to report incomplete, ambiguous or negative findings that may be valuable as part of a larger pattern.' [discussion, (3)]

Begg and Berlin observe that:

> 'This phenomenon is encouraged in the competition for academic promotion and lamented for its role in degrading the quality of published medical research' (2)

Begg and Berlin (3) call for a policy agenda that will lead to the improved quality of published research data and so reduce the impact of bias for the future. Several long-term suggestions have been put forward, these have the potential for greater long-term impact than retrospective efforts to correct the bias technically by statistical modelling (3).

Begg and Berlin (2) suggest blinding reviewers of journals to the results, so a decision on publication would be made solely on the methods, this would encourage people to write reports with negative results. Altman [discussion, (3)] believes the concept of positive and negative studies should be abandoned altogether. Another solution suggested would be to give incentives to both authors and editors to publish negative results (3). Chinn [discussion, (3)] suggests creating a '*Journal of Unbiased Results*' by publishing the acceptance date and expected date of publication of forthcoming papers. If the paper failed to materialise, the reason would be given as a deterrent for not publishing. Alternatively, negative studies could be published in a reduced form and hence take up less space in journals. Begg and Berlin suggest that at the very least the title should be published, rather than not at all (3). A similar alternative is that unpublished results could be given concisely in review articles (59). An anonymous editorial in the *Lancet* observes that in 1991 physicians in Spain and France were already required to register all drug trials with their respective ministries of health, and in Japan it is mandatory to publish the results of every single trial, perhaps this should be the case for every country? Another possibility is using the potential of peer reviewed on-line journals, with no space restrictions this could alleviate publication bias (60).

However, registering and reporting all trials does have potential problems. Blair [discussion, (3)] argues that in general, positive results are more important to readers than negative ones and that publication bias only important to meta-analysis not science in general, and hence the publication of all studies in full is not justified. As for research registers, pharmaceutical companies may fear competitors may get an unfair advantage knowing their latest trials (61).

Including unpublished information

Pages 14–15 discussed identifying grey material and thus outlined ways of identifying unpublished studies. It may seem clear, that if one could identify all the unpublished studies and retrieve the relevant information, publication bias could be alleviated. However, Cook *et al.* (62) report an incident where data was requested from a manuscript (for the purpose of a meta-analysis) sent to the *New England Journal of Medicine,* where it had been presented only in abstract form. This request was forwarded to the editor who replied saying the intention to include unpublished data, 'both surprising and disturbing', and that if the data was released he would not consider the manuscript, commenting 'I imagine that most other editors would do the same'.

This led Cook *et al.* (62) to investigate attitudes towards unpublished data in a meta-analysis. They found that 46 of 150 meta-analyses examined used unpublished results. 46.9% of editors asked felt unpublished data should probably or certainly be included, 30% of editors would not publish an overview that included unpublished data. Of meta-analysts and methodologists, 34.3% would definitely exclude material that had not been published anywhere.

An investigation into whether the inclusion of unpublished studies, which are possibly methodologically weak compromise the validity of a meta-analysis, would be useful.

For a discussion on how study quality affects the results, and ways of incorporating study quality into a meta-analysis see chapter 6.

Further research

Delineate more clearly the important correlates of publication bias, possibly using retrospective data, utilising methods analogous to those used on pages 124–6. Additionally, to establish if correlates vary depending on the subject matter (3).

Issues regarding the impact reporting multiple endpoints has on publication bias, e.g. there may be a strong incentive to publish a cancer trial if any one of the end points shows a positive result, however all four may be published in the same paper. In this situation bias may be negatively correlated between outcomes (3).

Estimating publication bias using the OR as the outcome scale using a random effects model (i.e. in presence of heterogeneity) (46).

Assessing the impact of the pipeline problem (see pages 121–2) (27)

Begg comments:

> 'it seems plausible that in practice selective publication will be influenced by both the magnitude of the effect size and the *p* value, and it would be desirable to develop a test that is sensitive to either of these influences, but at present no such methodology is available.' (4)

Begg (4) compiled the list below of currently unanswered questions regarding publication bias in meta-analysis:

1. What are the most sensitive approaches for detecting bias?
2. What are the relative merits of methods based on the funnel graph versus methods based on weighted distribution theory (question could now incorporate the newer methods also)?
3. What are the chances of failing to detect a bias that would have a profound effect on the meta-analysis?
4. How many component studies must there be before one has reasonable power to detect bias? (The new method of study imputation (6) claims to work when one has only got a small number of studies to combine.)

Establishing whether the inclusion of unpublished studies, which are possibly methodologically weak, compromise the validity of a meta-analysis.

Much work has focused on publication bias for RCTs. The problem is probably even greater for observational studies. Perhaps further investigation into this is required.

Rao (27) comments that the problem of heterogeneity cannot be empirically split from that of publication bias. It would seem that a homogeneity test, taking into account selection bias would be useful.

Methods for assessing and adjusting meta-analyses for publication bias, when only small numbers of studies are being combined.

Summary

In conducting a meta-analysis, researchers should always be aware of the potential for publication bias, and make efforts to assess to what extent publication bias may affect their meta-analysis. In terms of the inclusion of unpublished studies, a sensitivity analysis should be performed to assess the likely impact of including unpublished data.

The intention of above sections was to give the reader a brief but relatively complete overview of the methods proposed to deal with publication bias. It has already been noted that many of the methods are new and exploratory. In 1988, Begg and Berlin in their thorough and excellent review on the subject commented:

> 'It is difficult to conceive of a correction methodology which would be universally credible.' (3)

Since then, these methods have grown more sophisticated but the authors always stress the need to use them as a form of sensitivity analysis. The file drawer method of Rosenthal, again is a form of sensitivity analysis which being older has gained a certain amount of acceptance. Time will tell if these newer, correction methods supersede this simple calculation.

It has been demonstrated that publication bias can be elevated by using a sampling frame. Louis commented of the correction methods:

> 'All of these methods are based on fairly strong assumptions, and the current consensus seems to be that although these methods may be valuable tools, long-term policy measures aimed at reducing publication bias are required' [(3), discussion]

This is echoed by Begg and Berlin:

> 'The only method which is likely to gain widespread acceptance is the use of an invariant sampling frame (if available).' (3)

Hedges states:

> 'It is difficult to dispute that the ideal solution to the problem of publication bias is the development of an unbiased sampling frame via a registry or an ongoing census of studies.' [discussion (3)]

With advances in world-wide communications, via the Internet etc., the feasibility of world-wide registries has increased. With the inception of groups such as the Cochrane Collaboration, it would seem that the first advances have been made.[21]

References

1. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;**337**:867–72.

2. Begg CB, Berlin JA. Publication bias and dissemination of clinical research. *J Natl Cancer Inst* 1989;**81**:107–15.

3. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data (with discussion). *J R Statist Soc A* 1988;**151**:419–63.

4. Begg CB, Cooper H, Hedges LV, editors. Publication bias. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 399–10.

5. Chalmers TC. Informed consent, clinical research and the practice of medicine. *Trans Am Clin Climatol Assoc* 1982;**94**:204–12.

6. Givens GH, Smith DD, Tweedie RL. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statist Sci* 1997;**12**:221–50.

7. Grégoire G, Derderian F, Lelorier J, Le Lorier J. Selecting the language of the publications included in a meta-analysis – is there a Tower-of-Babel bias? *J Clin Epidemiol* 1995;**48**:159–63.

8. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, Le Lorier J, *et al.* Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 1996;**347**:363–6.

9. Boissel JP, Haugh MC. The iceberg phenomenon and publication bias: the editors' fault? *Clinical Trials and Meta-Analysis* 1993;**28**:309–15.

10. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith HJ. Publication bias and clinical trials. *Controlled Clin Trials* 1987;**8**:343–53.

11. Simes RJ. Confronting publication bias: a cohort design for meta-analysis. *Stat Med* 1987;**6**:11–29.

12. Melton A. Editiorial. *J Exp Psychol* 1962;**64**:553–7.

13. Gotzsche PC. Reference bias in reports of drug trials. *BMJ* 1987;**295**:654–6.

14. Coursol A, Wagner EE. Effect of positive findings on submission and acceptance rates: a note on meta-analysis bias. *Professional Psychology* 1986;**17**:136–7.

15. Egger E, ZellwegerZahner T, Schneider M, Junker C, Lengeler C. Language bias in randomised controlled trials published in English and German. *Lancet* 1997;**350**:326–9.

16. Dickersin K, Min YI. NIH clinical trials and publication bias. *Online J Curr Clin Trials,* Doc: No. 50, 1993.

17. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA* 1992;**263**:374–8.

18. Dawid AP, Dickey JM. Likelihood and Bayesian inference from selectively reported data. *J Am Statist Assoc* 1977;**72**:845–50.

[21] Very recently, Eberly and Casella (48) have made the following recommendation: Considering there are three basic methods to deal with publication bias: truncated sampling models, invariant sampling, and source argumentation. They state that truncated sampling and invariant sampling often assume that the researcher has access to each studies effect estimates and perhaps sample variances. However, they state:

> 'Reality forces us to acknowledge, though, that often we cannot acquire the original data from a study, sometimes not even the effect size estimates. Especially with older studies, it is likely that only *p*-values or *t*-values can be gleamed from the publication itself (which renders many of the above methods impossible). However, source argumentation does not require more than *p*- or *t*-values. In spite of this advantage, we believe source augmentation should, whenever possible, be carried out in addition to effect size estimation.' (48)

19. Freirnan JA, Chalmers TC, Smith HJ, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized controlled trial: survey of 71 'negative' trials. *N Engl J Med* 1978;**299**:690–4.

20. Angell M. Negative studies. *N Engl J Med* 1989;**321**:464–6.

21. Hedges LV. Modeling publication selection effects in meta-analysis. *Statist Sci* 1992;**7**:246–55.

22. Lane DM, Dunlap WP. Estimating effect-size bias resulting from significance criterion in editorial decisions. *Br J Math Stat Psyc* 1978;**31**:107–12.

23. Kurosawa K. Discussion on meta-analysis and selective publication bias. *Am Psychol* 1984;**39**:73–4.

24. Champney TF. Adjustments for selection: publication bias in quantitative research synthesis. University of Chicago, 1983.

25. Dawes RM, Landman J, Williams M. Discussion on meta-analysis and selective publication bias. *Am Psychol* 1984;**39**:75–8.

26. Hedges LV. Estimation of effect size under nonrandom sampling: the effects of censoring studies yielding statistically insignificant mean differences. *J Educ Stat* 1984;**9**:61–85.

27. Iyengar S, Greenhouse JB. Selection models and the file drawer problem. *Statist Sci* 1988;**3**:109–35.

28. Begg CB. A measure to aid in the interpretation of published clinical trials. *Statist Med* 1985;**4**:1–9.

29. Begg CB, McGlave PB, Bennett JM, Cassileth PA, Oken MM. A critical comparison of allogeneic bone marrow transplantation and conventional chemotherapy as treatment for acute nonlymphocytic leukemia. *J Clin Oncol* 1994;**2**:369–78.

30. Smith ML. Publication bias and meta-analysis. *Evaluation in Education* 1980;**4**:22–4.

31. Hemminki E. Study of information submitted by drug companies to licensing authorities. *BMJ* 1980;**280**:833–6.

32. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994;**50**:1088–101.

33. Light RJ, Pillemar DB. Summing up: the science of reviewing research. Cambridge, Mass: Harvard University Press, 1984.

34. Hedges LV, Olkin I. Vote-counting methods in research synthesis. *Psychol Bull* 1980;**88**:359–69.

35. Berlin JA, Begg CB, Louis TA. A methods for assessing the magnitude of publication bias in a sample of published clinical trials. Boston: Dana-Farber Cancer Institute, 518Z, 1987.

36. Yusuf S, Peto R, Lewis J, Collins R, Sleight P, *et al.* Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Prog Cardiovasc Dis* 1985;**27**:335–71.

37. Rosenthal R. The file drawer problem and tolerance for null results. *Psychol Bull* 1979;**86**:638–41.

38. Rosenthal R. Combining the results to independent studies. *Profess Psychol* 1978;**17**:136–7.

39. Orwin R. A fail-safe N for effect size in meta-analysis. *J Ed Statist* 1983;**8**:157–9.

40. Klein S, Simes J, Blackburn GL. Total parenteral nutrition and cancer clinical trials. *Cancer* 1986;**58**:1378–86.

41. Hedges LV. Gleser LJ, Perlman MD, Press SJ, Sampson AR, editors. Estimating the normal mean and variance under a publication selection model. In: Contributions to probability and statistics: essays in honor of Ingram Olkin. New York: Springer, 1989, p. 447–58.

42. Dear KBG, Begg CB. An approach for assessing publication bias prior to performing a meta-analysis. *Statist Sci* 1992;**7**:237–45.

43. Paul NL. Non-parametric classes of weight functions to model publication bias. Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA, #622, 1995.

44. Larose DT, Dey DK. Modelling publication bias using weighted distributions in a Bayesian framework. #95-02, 1995. *Comput Statist Data Anal* 1998;**26**:279–302.

45. Patil GP, Taillie C. Dodge Y, editors. Probing encountered data, meta-analysis and weighted distribution methods. In: Statistical data analysis and inference. B.V., North Holland: Elsevier Science Publishers, 1989.

46. Sugita M, Kanamori M, Izuno T, Miyakawa M. Estimating a summarized OR whilst eliminating publication bias in meta-analysis. *Jpn J Clin Oncol* 1992;**22**:354–8.

47. Sugita M, Yamaguchi N, Izuno T, Kanamori M, Kasuga H. Publication probability of a study on OR value circumstantial evidence for publication bias in medical study areas. *Tokai J Exp Clin Med* 1994;**19**:29–37.

48. Eberly LE, Casella G. Estimating the number of unseen studies. BUM, #1308-MA, 1996.

49. Gleser LJ, Olkin I. Models for estimating the number of unpublished studies. Department of Statistics, Stanford University, #313, 1995.

50. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970;**57**:97–109.

51. LaFleur B, Taylor S, Smith DD, Tweedie RL. Bayesian assessment of publication bias in meta-analyses of cervical cancer and oral contraceptives. Technical Report, University of Colorado, 1996.

52. Smith DD, Givens GH, Tweedie RL. Adjustment for publication and quality bias in Bayesian meta-analysis. Technical Report, University of Colorado, 1997.

53. Bayarri MJ, DeGroot M. A Bayesian view of weighted distributions and selection models. Department of Statistics, Carnegie Mellon University, #375, 1986.

54. Bayarri MJ, DeGroot M. The analysis of published significant results. Department of Statistics, Carnegie Mellon University, #91-21, 1991.

55. Cleary RJ. An application of Gibbs sampling to estimation in meta-analysis: accounting for publication bias. *Journal of Education and Behavioral Statistics* 1997;22:141–54

56. Frongillo E. Combining information using hierarchical models (dissertation). Biometrics Unit, Cornell University, Ithaca, NY, 1995.

57. Simes RJ. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* 1986;**4**:1529–41.

58. Easterbrook PJ. Directory of registries of clinical trials. *Stat Med* 1992;**11**:345–423.

59. Bero LA, Glantz SA, Rennie D. Publication bias and public health policy on environmental tobacco smoke. *JAMA* 1994;**272**:133–6.

60. Mossman D, Somoza E. Maximizing diagnostic information from the dexamethasone suppression test: an approach to criterion selection using receiver operating characteristic analysis. *Arch Gen Psychiatry* 1989;**46**:653–60.

61. Editorial. Making clinical trialists register. *Lancet* 1991;**338**:244–5.

62. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, *et al.* Should unpublished data be included in metaanalyses – current convictions and controversies. *JAMA* 1993;**269**:2749–53.

63. Heatherington J, Dickersin K, Chalmers I, Meinert CL. Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. *Pediatrics* 1989;**84**:374–80.

64. Hedges LV, Olkin I. Statistical methods for meta-analysis. London: Academic Press, 1985.

Chapter 17

Missing data

Types of missing data

Data missing in meta-analysis can be split into three broad categories, namely, data can be missing in the below situations:

i) Whole studies – chapter 16 discusses publication bias, a situation where not all the studies carried out in an area have been reported. Ways of dealing with missing studies are dealt with in the latter half of chapter 16 and will not be discussed further here.
ii) Study level – data can be missing at the study level. It may be that a significance level or *p*-value, but no treatment effect size estimate is reported. Study level covariates may also be missing, either completely, or only partially reported (e.g. due to missing data at the patient level), the study may only report the mean age of 80% of the people in a study. This type of missing data can be problematic, and a problem unique to meta-analysis.
iii) The individual patient level – chapter 24 discusses meta-analysis of individual patient data (MAP). Here the original data from the primary studies is collected on all patients, this is then merged into one large dataset. The analyses possible are similar in nature to those carried out in multicentre trials. When data are missing at this level, standard techniques for missing data used in trials and observational studies can be employed in the meta-analysis.

This chapter will concentrate on situation ii), where data are missing at the study level, however, many of the techniques could be used in situation iii), at the individual patient level also. For additional information on estimating effect sizes from *p*-values, see pages 149–52.

Reasons why the data are missing

The below account is adapted from Piggott (1), who highlights different situations when study level data is missing.

Influences of research reporting practices

The amount of information given in a study report may be limited by the nature of the publication. For instance, a dissertation may be considerably longer than a paper and thus include more information. In addition the researcher may be influenced by customary reporting practices in the research area, the background of the authors, and also the subjective view of different authors as to what they perceive is important and the emphasis of the article.

Missing for reasons unrelated to the data

This type of data is considered missing at random. The cases with complete information can be treated as a random sample of the original set of studies. Little and Rubin (2) use the term 'missing completely at random' to make it distinct from the situation given below.

Missing for reasons related to completely observed variables

In this instance, missing values occur because of the value of another completely observed variable, and not because of the value of the missing variable itself. Little and Rubin (2) use the term 'missing at random'. Analysing only complete cases in this situation may not provide generalisable results. Methods described in (2,3) are appropriate here.

Missing for reasons related to the missing values themselves

Observations can be missing because of the value of the variable itself or because of other unobserved variables. This can be caused by censoring mechanisms. An example of this situation are missing effect sizes not reported because they were not statistically significant. In this instance the data is missing non-randomly. This situation poses one of the most difficult problems in dealing with missing data.

Missing data at the study level

Missing study level effect sizes

The effect size estimate may be completely missing for particular studies. If a result is non-significant, often the actual magnitude of the effect is not given, this will have an effect on the meta-analysis similar to publication bias (i.e. since treatment estimates from significant studies are more likely to be given, this will lead to a systematic bias overestimating the pooled effect estimate).

When the direction, but not the magnitude, of the effect size estimate is known, vote counting analysis can be carried out. Pigott (1) comments some researchers fill in conservative estimates, such as zero, for missing effect sizes, and warns that, this could lead to bias results which are compounded when a variance of the effect size is calculated and weighted least squares is used.

Missing study level characteristics/ covariates

It should be noted that standard regression programs drop cases missing any variable in the model. Most quantitative research synthesis may use only studies with complete information on both outcomes and predictors when building models for effect size (1).

Simple analytic methods for dealing with missing data [adapted from (1)]

As Pigott states:

> 'The adequacy of these methods depends on the reasons data are missing' (1)

Analysing only complete cases

Although this method is always possible, if there are too many missing values, analysing only complete cases reduces the data considerably. Also, one has to assume that the remaining complete cases are representative of the original sample of studies.

Single-value imputation

The idea of this method is that the missing values are filled with a reasonable value (i.e. all the missing data get the same value). This method has the advantage that all the cases with one or more missing values are not lost. If one believes the missing values have a small value, often the value zero is imputed. Another alternative is to impute the mean value for that variable. The method makes the assumption that missing values are close to the imputed values. Little and Rubin (2) give exact formulae for the underestimation of the sampling variance that results from imputing single values for missing observations, because this artificially deflates the variability of the variable.[1]

Regression imputation: Buck's method

This method is suggested by Buck (4). It uses regression techniques to estimate missing values, replacing missing observations with the conditional mean. For every pattern of missing data, complete cases are used to calculate regression equations predicting a value for each missing variable using the set of completely observed variables. Little and Rubin (2) again give an adjustment for the underestimation of the sampling variance. This method assumes the missing variables are linearly related to other variables in the data.

Buck's method underestimates the sampling variance of Y_2, (Y_1 and Y_2 are both variables) by

$$\frac{\lambda}{(n-1)}\,\sigma_{22.1}$$

where λ is the number of missing values of Y_2, and $\sigma_{22.1}$ is the residual variance of the regression of Y_2 on Y_1[2]. If the two variables are linearly related, information about one variable can provide some information on the missing values of the second variable.

Which method to use when?

Pigott (1) notes that these methods do not work equally well in all situations. When observations are missing at random the complete case, mean imputation, and Buck's method provide unbiased estimates. However, mean imputation and filling in zero underestimate the SE of the mean.

It should be noted that the adequacy of the methods changes when missing data result from a censoring mechanism rather than a random deletion. For instance, these methods would fall down if all the highest, or lowest values were missing. Grossly under or over-estimates would be obtained.

Pigott (1) also comments there may be a danger in using the estimates of the SE of the mean, since the imputed effect sizes are used twice: once for the effect from an individual study and once for the estimate of the variance of that individual effect size. The inverse of the estimated variance is then used to calculate the weighted effect size (so more weight is given to studies with large sample sizes even if those studies values were imputed.) The weighted mean does not reflect the uncertainty due to missing observations and should be used with caution.

[1] Because imputing values deflates the variance, this has important implications for tests of homogeneity of effect sizes and for estimation of categorical or linear models of effect size.

[2] Little and Rubin (2) give the general form of underestimation for more than two variables.

Advanced methods for dealing with missing data

These methods can be used when missing observations are related to completely observed variables or to the values of the missing observations. For a Bayesian approach to this subject, see chapter 13.

ML models for missing data

This method uses all the data contained in the original sample and does not require any adjustments to the data (unlike the Buck method). Little and Rubin (2) state that the mechanism that leads to missing data can be ignored when it is either unrelated to information in the data or related to observed information. These ML methods follow the usual steps for complete data MLE, but also include a model for the reasons for the missing data in the likelihood of the data. The part of the function that governs the missing values is called the response mechanism. When missing data are deemed able to be ignored, the piece of the likelihood pertaining to the response mechanism can be ignored. Little and Rubin (2) present two models for ignorable data in research synthesis.

Mixed normal and non-normal data

Estimates means and cell frequencies when the data contains both continuous and categorical variables (i.e. categorical covariates) [see (2) for more details].

ML methods for multivariate normal data

Methods for linear models with ignorable missing data.

This method assumes the data is from multivariate normal distributions, though the procedure is robust (1), i.e. assumes effect size and predictors of effect size are jointly distributed as multivariate normal. Since effect-size estimators are not distributed identically, and since the variance of an estimate of effect depends on the sample size employed in the study, this method cannot be used directly [it has been suggested that weighted least squares could be used in this instance (5)]. Adjustments are needed to allow weighted least squares to be used in the estimation procedure [an algorithm is given in (6)].

The method does not estimate single missing values; rather, it estimates the means and covariance matrix by obtaining the expected values of the sufficient statistics (in this case sums, and the sums of the cross products, of the variables in the model) of the likelihood. The likelihood is calculated using the EM algorithm. It is necessary to calculate a series of regressions for each pattern of missing data. Missing data can occur on a number of predictor variables resulting in a series of patterns of missing data. SEs of the estimates can be calculated via the second derivative of the log likelihood (2) (often this is not an easy task!). Meng and Rubin (7) present a new algorithm less difficult to compute than (2), easier still, but less accurate, would be to use a jackknife procedure. The original algorithm is available in the BMDPAM program (but not using weighted models[3]).

If the data is non-ignorable, two methods to deal with it are outlined in (2): 1) requires knowledge about the reason for missing data (e.g. reviewer knows the exact value of the effect size above which no effect sizes are observed) and 2) is used most frequently when only missing values in one variable and the reason for being missing is unknown. Since in meta-analysis one never has exact information, nor are missing values often confined to one variable, their use is limited.

Multiple imputation

This method was described originally by Rubin (3), and in more detail in Little and Rubin (8). The method imputes more than one value for each missing observation and thus obtains a range of possible values for each missing observation. In doing so, it avoids the problem of having to assign only one value to the data. Pigott (1) comments that the theory was derived for large scale sample surveys, so theory only precisely valid for a large number of cases, although may be useful for small data sets, more research required.

Repeated measurement (outcome) missing data

Talwalker (9) deals with the problem of repeated outcome missing data in a novel way. The data is stratified according to the patterns of missing data. Then each persons repeated measurements are reduced to a summary measure. These summary measures are then compared using a distribution free test to investigate the treatment effect. Random effects have been incorporated into this analysis. The example given by Talwalker (9) to illustrate this method does not synthesise different studies, but strata from one study, however, the applicability of the method for meta-analysis is

[3] Theoretically, one could use SAS PROC IML also.

discussed. It should be noted in the stratified example the time periods for the repeated measurements are the same for each person.

Further research

Generally more research is needed in this topic. Specifically, methods presently available are only of use when missing values are confined to one variable, hence methods for the multivariate situation need. Also more research is required to ascertain whether multiple imputation can be used on small datasets.

Summary

Not a lot has been written on the problem of missing data in meta-analysis. Most of the methods discussed here have been adapted from other situations. Many of the advanced methods have not been used extensively in a meta-analysis setting (1). Pigott suggests that the current development of computer programs that implement the procedures described by Little and Rubin (2) should advance the development of sensible methods for handling missing data in research synthesis (1).

Cooper and Hedges state (10) that missing data are 'perhaps the most pervasive practical problem in research synthesis'. The also observe that 'the prevalence of missing data on moderator and mediating variables influences the degree to which the problems investigated by a synthesis can be formulated', and predict that new methods will evolve, and that:

> 'Much of this work will likely be in the form of adapting methods developed in other areas of statistics to the special requirements of research synthesis. These methods will produce more accurate analyses when data are not missing completely at random but are well enough related to observed study characteristics that they can be predicted reasonably well with a model based on data that are observed.' (10)

When covariate information is missing this can be a problem when analysing heterogeneity using meta-regression (see chapter 11) as Pigott explains:

> 'A synthesist may try several different analyses with the data to determine if any of a study's characteristics relate to the effect magnitude of the study. In each of these analyses, only studies with complete information on relevant variables may be included. Each of these analyses may utilise a different set of studies that may not be representative of the sample originally chosen for the synthesis and may not correspond with each other. The results of each analysis may not generalise to the population of studies on a topic nor to any of the other samples of studies used in the analyses.' (1)

It should be noted that the methods presented do not help decide the reasons the data is missing in the first place. Pigott suggests creating a missing data index variable, taking the value one if a variable is observed and 0 if it is missing:

> 'This variable can be correlated with other completely observed variables in the data or used as the outcome variable in a logistic regression modelling response as a function of several completely observed variables. If one finds a correlation between a completely observed variable and the missing data index or a plausible model of the missing values, then some evidence exists that the reasons for missing observations depend on completely observed variables.' (1)

It should be stressed that whatever method is used to deal with missing data, a careful sensitivity analysis of the modelling assumptions on the conclusions should be performed as a final step.

References

1. Pigott TD, Cooper H, Hedges LV, editors. Methods for handling missing data in research synthesis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 163–76.
2. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley, 1987.
3. Rubin DB. Multiple imputation for non response in surveys. New York: Wiley, 1987.
4. Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J R Statist Soc B* 1960;**22**:302–3.
5. Draper N, Smith H. Applied regression analysis. 2nd edn. New York: Wiley, 1981.
6. Pigott TD. The application of maximum likelihood methods to missing data in meta-analysis. University of Chicago, 1992.
7. Meng X, Rubin DB. Using EM to obtain asymptotic variance-covariance matricies: the SEM algorithm. *J Am Statist Assoc* 1991;**86**:899–909.
8. Little RJA, Rubin DB. The analysis of social science data with missing values. *Soc Methods Res* 1989;**18**:292–326.
9. Talwalker S. Analysis of repeated measurements with dropouts among Alzheimer's disease patients using summary measures and meta-analysis. *J Biopharm Stat* 1996;**6**:49–58.
10. Cooper H, Hedges LV, Cooper H, Hedges LV, editors. Potentials and limitations of research synthesis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 521–30.

Chapter 18

Reporting the results of a meta-analysis

Introduction

This section discusses ways in which results of a meta-analysis can be reported, interpreted, and presented. In previous chapters of the report it has been necessary to present results for the illustrative examples, so the reader that has read through this report sequentially will already be familiar with several of the methods discussed here.

Overview and structure of a report

Deeks *et al.* (1) report the CRD guidelines for the format their systematic review reports should take. Although the format of a report for other sources may vary, not least due to space constraints, it is a very good starting reference. A shortened version is reproduced below. For further information, Halvorsen (2) has written a chapter on this subject, which follows a very similar structure and gives many more details.

Abstract or executive summary

Background information

> 'The need for the report should be justified by clearly describing the problem for which evidence of effectiveness is required, and describing the needs of the health care professionals and consumers who are to benefit from the report.'

Hypotheses tested in the review

Review methods

> 'The methods used should be described in sections for search strategy, inclusion criteria, assessments of relevance and validity of primary studies, data extraction, data synthesis, and investigations of differences between studies.'

Details of studies included in the review

> '... details relating to the patient groups included, mode of intervention and the outcomes assessed in each study. Details of study results, study design and other aspects of study quality and validity should also be given. Sufficient information should be provided to allow replication of the analysis.'

Details of studies excluded from the review

Given reason for exclusion.

Results of the review

> 'The estimates of efficacy from each of the studies should be given, together with the pooled effect if this has been calculated. All results should be expressed together with CIs. The table or diagram should indicate the relative weight that each study is given in the analysis. The test for heterogeneity of study results should be given if appropriate and all investigations of the differences between the studies should be reported in full. As well as reporting the results in relative terms the impact of the results in absolute terms [such as absolute risk reduction (ARR) and number needed to treat (NNT) (see pages 109–10)] should be given. This permits the clinical significance and possible impact of the intervention to be assessed.'

Analysis of the robustness of the results

> 'Sensitivity analyses should be performed and documented to investigate the robustness of the results where there is missing data, uncertainty about study inclusion, or where there are large studies which dominate the data synthesis.' (see pages 209–10 for more details on sensitivity analysis).

Discussion

> 'A discussion of the strength of the causal evidence, potential biases in both the primary studies and the review, and the limitations they place on inferences, should be given.'

Implications of the review

> 'The practical implications of the results both for health care and future research should be discussed. This section should take account the needs of the target audience.'

Reference lists

> 'Three lists of studies should be given: the studies included in the review, the studies excluded from the review, and any other literature which is referred to in the report.'

Dissemination and further research

> 'Suggestions of the main messages for dissemination and the important target audiences should be discussed. Implications for further research should be outlined with a discussion of lessons of the review for the research methods that may be useful.'

Graphical displays used for reporting the findings of a meta-analysis

Under the results of the review part outlined above, the results that need reporting for a meta-analysis were given. Graphical displays can aid and enhance these results. Several different graphical plots have been used to display the findings of a meta-analysis. These are discussed below.

Funnel plot

This plot was described on pages 124–5 for assessing the presence of publication bias. The effect estimate for each study is plotted against some measure of the size of the study (usually the variance of the effect estimate). Light *et al.* (3) report, this plot can be used to check if the studies are 'well behaved' (i.e. if all studies are estimating a single underlying population effect size parameter). If they are well behaved then the plot should look roughly like funnel because studies with smaller sample sizes will display more variability in effect size than the investigations with larger sample sizes (3).

Forrest plot

This type of plot does not appear to have a standard name, however it has been referred to as a Forrest plot (it is also known as a Cochrane plot). This type of plot has been used in chapters 9 and 10 of this report to display the results of the various analyses of the cholesterol lowering data. Much information is succinctly conveyed in such a figure (3), with point estimates and 95% CIs for each study, along with the final combined result and CI all being displayed. In addition, the size of the study is represented by the size of the box indicating the estimated treatment effect. This plot has its drawbacks, however; it has been pointed out that one's eye is drawn often to the least significant studies, because these have the widest CIs and are graphically more imposing (4). Another problem is when very large trials have been carried out, the corresponding size of box is necessarily bigger than the very tight CI around it (keeping the scale reasonable for the other studies), making the level of certainty difficult to ascertain. A point that needs considering is which scale to use on the horizontal axis. If the results of trials are presented in the form of ORs, then a log scale may be more appropriate. If the linear OR scale is used then this may distort ones interpretations. Galbraith (4) gives an example where a trial with twice as many people in it as another trial, has a CI of longer 'visual' length (see paper for details). Galbraith (4) also comments that when reporting the combined estimate and CI on the graph next to the plot, the linear scale should be used as this is more interpretable.

Radial plots

These were first described on page 41 as an informal method to assess heterogeneity. As for the Forrest plot, information regarding the point estimate and precision of each trial is displayed, however an overall combined result is not given. This plot can be useful in exploratory analysis. Galbraith reports (5) that when many aspects need to be considered when comparing trials, these plots can provide a useful focus for discussion, enabling differences between subgroups[1] and exceptions to be seen easily.

Methods specific to continuous effect measures

Light *et al.* (3) state that (when a continuous outcome variable is being used) it is important to display the distribution of the effect sizes obtained in research synthesis. Commenting:

> 'Then readers will be able to assess the shape of the distribution and draw their own conclusions about the overall size and variability of the effects being integrated'. They present several methods to 'enhance such displays in substantively interesting and methodologically meaningful ways.' (3)

They take a list of the studies and corresponding effect estimates and comment that the first aid to interpretation is to order the studies by treatment effect. From here, stem-and-leaf plots are created. These are simple plots which are similar to histograms, however the original data can be extracted from the plot [see (6) for more details]. Back to back stem-and-leaf plots can be produced to compare studies split by a covariate (e.g. for comparing studies using different treatment regimes). Another type of plot suggested is the box and whisker plot. The end of each whisker denotes the maximum and minimum effect sizes, the two sides of the box, the upper and lower quartiles, and the middle vertical line the median effect size [see (6) for more details]. Several of these plots can be drawn alongside each other permitting subgroups of the studies to be compared.

Graphs investigating length of follow-up

Light *et al.* (7) present a display used when studies have different follow-up times. Here time is plotted on the horizontal axis and treatment effect on the vertical. Each study estimate is plotted along with vertical 95% CIs. A running mean line is plotted through these points in addition to a horizontal median line. This graph could help determine if follow-up time had an effect on the treatment estimate.

[1] These can be plotted using different shaped points or different colours for each subgroup.

Odd man out

This method was discussed on page 50. It is included here to point out that it utilises a graphic display different from the Forrest plot (page 144).

Obtaining a consensus for reporting results – the example of environmental epidemiological studies

Blair *et al.* (8) attempted to reach a consensus on how results of meta-analyses in environmental epidemiology should be reported. Their discussion considers the presentation, interpretation, and communication of results, a summary of which is given below. To the authors of this report's knowledge, this is the only documentation relating to obtaining a standard reporting format. Since many of the points could relate to meta-analyses of other types of studies (including RCTs), the section is included here, rather than chapter 19, which includes methodology exclusively for epidemiological studies.

Blair *et al.:*

> 'The presentation of the meta-analysis should be similar to the presentation of any of the individual studies that make up the analysis in that there should be a background section, description of methods, results, discussion, and conclusion. The report should reiterate the limitations of the studies that are included in the meta-analysis and reasons for the exclusion of studies. The discussion should clearly identify key assumptions and their rationale, address uncertainties, and offer reasonable alternative assumptions and conclusions.' (8)

Throughout this report a lot of emphasis has been place on methods to calculate summary statistics, (e.g. a combined OR and 95% CI). This is a very common and intuitively appealing procedure. However, Blair *et al.* (8) warns against this being the only result reported from a meta-analysis, by commenting this gives no indication of the amount of heterogeneity present. Greenland goes as far to say, 'a meta-analysis should be treated as a study of studies, rather than as a means for combining study results into a single effect estimate.' (9)

Blair *et al.:*

> '**Stratification**: consensus was not reached on the degree of stratification that should be conducted in a meta-analysis. Where stratification is conducted, an important contribution in any meta-analysis is to array results both in a table and in the text by exposure metric, study design, and health outcome.'

This statement was expanded upon:

> 'Some argued that a high degree of stratification defeats a common purpose of a meta-analysis, which is to summarize or show a central tendency. Nonetheless summarizing defeats another and perhaps more important purpose of meta-analysis, the detection and explanation of differences among study results.' (8)

> '**Sensitivity and influence analysis**: the results of sensitivity and influence analysis should be included in the results of a meta-analysis.'

The authors expand upon this by saying: 'It has been argued that sensitivity analysis should permeate all stages of a meta-analysis (10) including study selection, quality scoring, and to determine the effect of cofounders on outcomes (epidemiological studies).' (8)

> '**Documentation**: all procedures used in a meta-analysis should be documented for the purpose of replicability. Documentation should be enhanced by including information on knowledge gaps and research that should be conducted to fill the gaps.' (8)

> 'The meta-analysis table: summary tables of the studies that are included in the meta-analysis should be presented and should include at least the following information: the name of the study, the author(s), the date conducted, the summary statistic (point estimate and 95% CI), the exposure variable (measurement metric, range, average exposure), and key covariates. Graphic displays of the data can also be extremely helpful. If space is not a problem, enough information should be presented to allow the meta-analysis to be replicated.' (8)

> '**Graphics**: portraying study results through graphics, which can greatly assist the interpretation of large, complex tables of numbers is not a substitute for quantitative tables. If space permits only graphical presentations, the underlying data should be made available, for example through services such as the National Auxiliary Publishing Service.' (8)

Summary

This chapter has given a brief overview of methods used to report a systematic review. It is recommended for researchers to include tables of all studies considered in a review, so possible to see which were excluded. The bottom line on reporting a review is that enough information should be provided so people can replicate, or carry out changes/updates to it.

References

1. Deeks J, Glanville J, Sheldon T. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. Centre for Reviews and Dissemination, York: York Publishing Services, #4, 1996.

2. Halvorsen KT, Cooper H, Hedges LV, editors. The reporting format. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 425–38.

3. Light RJ, Singer JD, Willett JB, Cooper H, Hedges LV, editors. The visual presentation and interpretation of meta-analyses. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 439–54.

4. Galbraith RF. A note on graphical presentation of estimated ORs from several clinical trials. *Stat Med* 1988;**7**:889–94.

5. Galbraith RF. Some applications of radial plots. *J Am Statist Assoc* 1994;**89**:1232–42.

6. Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley, 1997.

7. Oxman AD, Oxman AD, editor. The Cochrane Collaboration handbook: preparing and maintaining systematic reviews. Second edn. Oxford: Cochrane Collaboration, 1996.

8. Blair A, Burg J, Foran J, Gibb H, Greenland S, Morris R, *et al.* Guidelines for application of meta-analysis in environmental epidemiology. ISLI Risk Science Institute. *Regul Toxicol Pharmacol* 1995;**22**:189–97.

9. Greenland S. Invited commentary: a critical look at some popular meta- analytic methods. *Am J Epidemiol* 1994;**140**:290–6.

10. Olkin I. Re: 'A critical look at some popular meta-analytic methods' (comment). *Am J Epidemiol* 1994;**140**:297–9.

Part F:

Results V – applications of meta-analysis in other contexts and using other data types

Chapter 19

Meta-analysis of observational studies

Introduction

Spitzer notes that:

> 'The controversies surrounding meta-analysis of experimental trials are equally relevant to non-experimental studies which are usually epidemiological. But there are additional unanswered questions.' (1)

Several common study designs are used in epidemiology, namely cohort, case–control and cross-sectional surveys. It is often difficult to confirm a relationship between exposure and disease, because of small prevalence or incidences, moderate effect sizes, and long latency periods in epidemiological studies (2). It has been noted that in such situations, meta-analysis could be a powerful, and even an essential, tool for integrating and combining the results of several studies to reach an overall statement (3). Indeed, with increased numbers in a pooled analysis, rare exposures can be more easily studied (4). It is important to note that meta-analysis cannot prove causation or confounding; however it does help an epidemiologist decide if a particular association does or does not exist, and if so provides an indication (but not necessarily a firm estimate) of the quantitative relationship between them (3).

Although the techniques used for meta-analysis of epidemiological studies are often similar to those used for RCTs [guidelines for meta-analysis of observational data should at the minimum follow those for clinical trials (5)], their aims may be different. Anello and Fleiss (6) go as far as making the distinction between analytic and exploratory meta-analysis, suggesting different protocols for each (they believe by keeping these two types of meta-analysis separate, it might help improve the reproducibility of future meta-analysis). However, whatever the aims of the meta-analysis are, the procedures used are usually similar.

It is commonly accepted that observational studies are prone to a greater degree of bias than RCTs, since avoidance of several biases is the prime objective of randomisation. For this reason, Spitzer (1) questions whether meta-analytic techniques can be applied to epidemiological studies, but considers the answer to be a 'guarded yes'.

Fleiss and Gross ask:

> 'Has proper control or adjustment been made for the biases that frequently occur in epidemiological studies, such as sociodemographic or clinical differences between study populations, misclassification of subjects with regard to case–control status and to levels of exposure, factors other than the level of exposure that may affect whether a subject is a case or control (i.e. confounding variables), and the publication bias/file drawer phenomenon wherein studies that fail to show a positive association tend not to be published and are thus not candidates for inclusion in the meta-analysis?' (7)

In addition, Morris (8) makes the observation that publication bias has not been systematically investigated for epidemiological studies. A further worry expressed is that if a study is published, it may only remark on the significant results and not mention non-significant ones tested.

For this reason, caution should be used when combining and reporting a meta-analysis of epidemiological studies. Sensitivity analysis can help tackle these shortcomings (see pages 209–10), and its importance cannot be overstated.

Many of the outcome scales used in epidemiological studies are the same as those used for RCTs, specifically the RR and the OR. For this reason, many of the statistical methods presented in this report can be used for combining epidemiological studies.

Whether epidemiological and RCTs, both looking at the same outcome, can be combined together is controversial. In the description of the cholesterol studies, used in the examples, it was noted that RCTs, cohort, case–control, and geographical studies have all been used in investigating the effect of cholesterol level on mortality. The methodology of combining studies of different designs is dealt with in chapter 26; thus the section below deals with combining epidemiological studies exclusively.

A publication which may be of interest to those reading this chapter is Blair (9), the background to which is described below:

> 'The ILSI Risk Science Institute convened an expert working group on meta-analysis during 1994 in Washington, DC as part of its co-operative agreement with the U.S. Environmental Protection Agency Office of Health and Environmental Assessment. The effort was designed to develop a consensus on: (1) the appropriateness of the use of meta-analysis in environmental health studies; (2) a set of guidelines or desirable attributes of meta-analysis applied to environmental health issues; and (3) when meta-analysis should or should not be used.' (9)

Many of the findings of this working group have been integrated into the present review, but obtaining it in its entirety is recommended. Another key paper on this subject is by Greenland (10), who lays out much of the methodology presented here in detail.

This chapter is structured to take the reader through each step of doing a meta-analysis of epidemiological studies, noting the differences from the standard methodology presented in the rest of the report, and discussing methodology exclusive to epidemiological studies.

Procedural methodology

Jones (11) broadly outlines key steps for a meta-analysis of epidemiological studies:

i) Compilation of as complete a set as possible of reports of relevant epidemiological studies.
ii) Identification of a common set of definitions of outcome, explanatory and confounding variables, which are, as far as possible, compatible with those in each of the primary studies.
iii) Extraction of estimates of outcome measures and of study and subject characteristics in a standardised way, and with due checks on extractor bias.
iv) Analysis of the summary data so extracted by one of the methods considered above.
v) Exploration of the sensitivity of the results of the meta-analysis in iv) to the choices and assumptions made in i)–iv).

Using this as a framework, the proceeding sections discuss the above stages in detail.

Compilation of reports

No publication addressing the methodology of searching and retrieving observational studies distinct from searching for other types of studies has been identified. For details on searching and identifying studies in general see chapter 4.

Specifying study variables

One needs to make a decision on the definitions of the variables used in the analysis. For instance, in the cholesterol example, we could use total mortality, or death from CHD, or a CHD event (fatal or not) as outcome variables. Similarly, the precise definition of exposure needs to be specified along with considerations of possible confounding variables (as well as intermediates and effect modifiers). This is influenced by the question one wants to answer and is also dictated by the data available.

Extraction of estimates

Introduction

Chêne and Thompson (12) deal with the problem of (large) differences between the (style of) presentation of results of epidemiological studies. Their illustrative example is a meta-analysis of nine studies investigating the relation between serum albumin and subsequent mortality. The paper summarised the major differences in the reporting of studies:

> 'Some studies presented crude numbers of deaths, mortality rates, or relative risks in groups defined according to serum albumin concentration. Different studies used between three and six groups, some using equally sized groups and others not. These studies, except two, also expressed the risk relation in another way, either as a logistic regression coefficient (or equivalently an odds ratio for a given increment in serum albumin) or as the mean difference in serum albumin concentrations between those subjects who died and those who survived. The logistic regression coefficients were in fact always adjusted for a number of confounding factors, but inevitably different confounding factors were used in different studies. Furthermore, standard errors for either the logistic regression coefficients or the mean differences were not always available, although sometimes a statement about the p value (e.g. nonsignificant or p<0.001) was given.'

For synthesis to proceed, it is necessary to express the results in a consistent (comparable) manner. The section below outlines many of the shortcomings of published reports, where the necessary information (treatment/exposure effect estimate and its variance) is either missing or disguised, and presents ways of deriving/estimating these values on a RR scale (see pages 109–10 for a definition of the RR), where possible. Unless otherwise stated the source of this information is the seminal paper by Greenland (10). For a more general discussion of data extraction, see page 17.

A note on the scales of measurement used to report and combine observational studies

Chapter 14 introduced several binary outcome measures. In considering these outcomes Greenland makes the following comments:

> 'If the outcome under study is rare in all populations and subgroups under review, one can generally ignore the distinctions among the various measures of relative risk (e.g., odds ratios, rate ratios, and risk ratios). The distinctions can, however, be important when considering common outcomes, especially in case–control design and analysis.' (10)

The majority of the methods in this chapter deal with estimates on the RR scale. See pages 116–17 for information on transforming binary outcome measures to this scale where necessary.

In is important to realise that estimates may come from coefficients of a logistic model as well as simply from 2 × 2 tables previously presented.

A good example of how different kind of estimators need considering is given by Piegorsch and Cox:

> 'Twenty-six of the studies included case–control data, and four were based on cohort study data. The case–control studies estimated the relative risk via the odds ratio, while the cohort studies used more complex risk ratio estimators. Thus the combined analysis represents a mix of different types of estimators for the RR endpoint.' [(13), p. 311)]

Finally, a new effects estimate, the standardised mortality ratio (SMR) may be used as an outcome in some situations. This is not covered in the scales of measurement chapter; however, ln(SMR) is usually assumed to be modellable with Normal least squares approaches.

Extracting when both the estimate of effect size as RR and its estimated SE are given in the report

If the scale used in the report is the same as the one to be used for the meta-analysis and its SE is given, then these can be directly copied and used in the meta-analysis.

Calculating the SE of an effect estimate as a RR from a CI

If a CI is given instead of a SE then, a simple computation is required to calculate the SE. This is explained by Greenland:

> 'For a relative risk estimate RR with a given 95 per cent lower limit of $\underline{RR}$ and upper limit of $\overline{RR}$, log RR is the desired log relative risk estimate. If the confidence limits are proportionally symmetric about the ratio (i.e. if $RR/\underline{RR} = \overline{RR}/RR$), an estimate SE of the standard error is given by SE = $(\log\overline{RR} - \log\underline{RR})/3.92$' (10)[1]

Calculating the SE of an effect estimate from a *p*-value

Greenland describes how to estimate the SE when only a *p*-value is given:

> 'If the *p* value is given accurately enough (to at least two significant digits if *p* is over 0.1 and one digit if *p* is under 0.1), one can compute a 'test-based' standard error estimate from $SE = (\log RR)/Z_p$, where Z_p is the value of a unit-normal test statistic corresponding to the *p* value (e.g. $Z_p = 1.96$ if $p = 0.05$, two-tailed test).'
>
> 'Unfortunately, because many reports use few significant digits in presenting *p* values, this method can be highly unstable for near null results, and it breaks down completely if log RR is zero. For example, given RR = 1.1, $p = 0.9$ (two-sided), one can only infer that RR is between 1.05 and 1.15 and that *p* is between 0.85 and 0.95, implying Z_p between 0.063 and 0.188; consequently, the original data could have yielded a standard error of anywhere from (log 1.05)/0.188 = 0.256 to (log 1.15)/0.063 = 2.22, compared to the test-based estimate of (log 1.1)/0.126 = 0.76. Another problem with this test-based method is that it gives a biased standard error estimate when the effect estimate is far from the null. For odds ratios and logistic coefficients, this bias will be small in most applications; for other measures, however, such as standardized mortality ratios, the bias can be substantial. In any case, the applicability of the test-based method is limited by the fact that most reports do not precisely specify *p*-values unless *p* is between 0.01 and 0.10, and often not even then.' (10)

Specialist methods for transforming/adjusting results from reports

When combining epidemiological studies one has to be aware of the potential for many differences between studies. The above sections dealt with ways of extracting estimates from reports. Unfortunately, different studies may have adjusted for different confounding variables, had different patient inclusion criteria, and so on. In addition to these problems, stratified results may have been presented for different levels of exposure. Biases

[1] If the limits are 90% confidence limits, the divisor in this formula should be 3.29 instead of 3.92.

may be known to exist in some studies but it may be possible to adjust for these also. An outline [derived from Greenland (10)] of how results can be adjusted for differences between the studies and for biases, and how to deal with stratified exposure results is given in the sections below.

Qualitative/categorical exposure variables – adjustments using external estimates of confounding

If a number of the studies selected to be combined had not adjusted for suspected important cofounders (in the original analysis), then it may be possible to estimate the degree of confounding present using other studies of the same outcome that provided data on the effects of the putative confounder. In this way the results of the trials with suspected confounding can be adjusted using external data (10). Details of how to do this are given below:

Factorisation of the relative risk: Step one: 'write the unadjusted (or partially adjusted) relative risk RR_u from the study under review (this procedure is carried out separately for each unadjusted study) as a product of two terms: $\mathrm{RR}_u = \mathrm{RR}_a\,(U)$, where RR_a is what the relative risk would be after full adjustment for the putative confounder or confounders, and U is the bias (in multiplicative terms) produced by having failed to fully control for the factor. Given U, a fully adjusted estimate can be derived from the unadjusted estimate via the equation $\mathrm{RR}_a = \mathrm{RR}_u/U$.

The problem confronting the reviewer is how to get an acceptable estimate of U. Given estimates of RR_u and RR_a from external data, one can estimate U via the equation $U = \mathrm{RR}_u/\mathrm{RR}_a$, but this estimate will be accurate only to the extent that the confounding effect (U) of the covariate in question is similar in both external data and the study under review. The value of U is particularly sensitive to the association of the study factor and the confounder, and to the association of the confounder with the outcome.' (10)

It is now necessary to calculate a SE for this externally adjusted RR estimate. Let V_U be an estimate of the variance of log (U), and SE the estimated SE for the unadjusted estimate (log RR_u) from the study under review. Then an estimate of the SE of the externally adjusted estimate:

$$\mathrm{SE}\left[\log(\mathrm{RR}_u/U)\right] = \sqrt{V_u + SE^2} \qquad (19.1)$$

One needs an estimate of V_u. If RR_c is the crude OR or person-time rate ratio from external data, RR_a is a common odds or rate ratio estimate from the same data (e.g. a Mantel–Haenszel OR), $U = \mathrm{RR}_c/\mathrm{RR}_a$, and V_c and V_a are variance estimates for log RR_c and log RR_a. Then, V_u may be computed as $V_a - V_c$, provided this quantity is positive.

Because of the high correlation of RR_c and RR_a, V_U will usually be small relative to V_a, and so if $\mathrm{SE}^2 \geq V_a$, external adjustment will not greatly increase the SE of the final estimate.[2]

Bounds for the magnitude of confounding

Bross (14) and Yanagawa (15) have derived bounds for the magnitude of confounding in studies involving dichotomous exposure and a dichotomous confounder.

However, Greenland reports:

> 'Unfortunately, the utility of such bounds is limited: first, for small effects (RR < 2), even a small percent distortion can be critical; second, in order to compute the bounds, one must know the (conditional) confounder-exposure and confounder-outcome associations; third, the extent of confounding produced by several variables or a single variable with more than two levels can greatly exceed the bounds computed for a dichotomy.' (10)

Adjusting an unadjusted RR: a method using confounder-exposure information

If a cohort study gives only an unadjusted estimate (RR_u), of exposure effect, but provides the joint distribution of exposure and the putative confounder in the total cohort, then these data can be combined with an external estimate of the confounder's effect on risk within levels of exposure. In this way, one can obtain an externally adjusted estimate of exposure effect that is potentially more accurate than the type given above (10).

For computational details to do this see (10), p. 8; extensions for use in case–control studies and adjusting multiple confounders are also discussed.

Adjusting for selection bias

'Occasionally, if the data are available, one may be able to reduce bias in a study by applying more strict exclusion criteria to the subjects, and then

[2] See paper for an example of how to use the above formula. Also see paper for a translation of formula for adjusting a coefficient in a Cox or logistic regression coefficient.

reanalyzing the study using only the subjects meeting the new criteria .... In most situations, there will be sufficient information to reanalyze only the crude data, but in such cases, one can parallel external adjustment for confounding.' (10)[3] [see (10) for method, also see chapter 24 and pages 156–64 for subsequently reported methods for IPD.]

Greenland also notes that the selection bias correction can be estimated from other studies, 'but if the parameters determining bias vary across studies, external correction could increase bias.' (10). However, these estimates could be used as a starting point for a sensitivity analysis.

Adjusting for misclassification

There are no simple methods to allow estimation of a correction factor for misclassification bias. Corrections for misclassification should be based on reconstruction of the correctly classified data [described in (16–18)]. If this cannot be done, informal sensitivity analysis can be carried out using speculated values of its magnitude.

Calculating exposure coefficients from stratified results

Ordered exposure variables often lead to presentations in terms of exposure-specific rates or ratios, and these ratios are usually computed without taking account of the ordering of exposure levels. An estimate of an exposure coefficient from such presentations can be achieved using a weighted least squares regression model if the SEs or CIs for each stratum estimate are given [see (10), p. 10 for details]. If they are not, but the report gives the size of the denominator for the rate in each exposure group, *ad hoc* approximate SEs can be computed [again, see (10), p. 10 for details].

SMRs derived using an external reference population

A related outcome is the SMR. This is often constructed by computing the expected values based on some external reference population. When these external reference rates are assumed known without error, an estimate of the exposure coefficient in an exponential regression may be obtained by a weighted linear regression of log(SMR) on exposure. If the CIs, or SEs, are reported for each SMR then a simple calculation yields the SE for log(SMR) [see (10), p. 11 for details].

Ratios derived using an internal reference group

Greenland:

> 'When a report presents results in terms of relative risk estimates that are computed by using a single internal exposure group as referent, one can perform a weighted linear regression of the log relative risk on exposure. Since the log relative risk for the reference level is necessarily zero (corresponding to a relative risk of one), the computations employ only the nonreference exposure groups, and the fitted line must be forced to pass through zero when the exposure is at the reference level. Because the numbers in the reference group are subject to statistical error and are employed in all the log relative risk estimates, the estimates will have nonzero covariances.' (10)

Estimation from reports employing only broad exposure categories

Greenland:

> 'Many reports treat continuous exposures in a categorical fashion, computing relative risks for broad categories of exposure....... In such cases it is necessary to assign numeric values to the categories before estimating coefficients. When the categories are broad, results will be sensitive to the method of assignment.'
>
> 'A common method is to assign category midpoints to categories. This has no general justification and gives no answer for open ended categories (e.g. more than 40 cigarettes a day). If, however, no frequency distribution for exposure is available, it may be the only choice, along with arbitrary assignments to open-ended categories.' (10)

If one has the frequency distribution preferably from the data in question, but if not, then from a study population with a similar exposure distribution. One may then assign to each broad category a numeric value corresponding to another measure of the centre of the category (e.g. the mean).

Estimation of coefficients from reports presenting only means

'Many reports in earlier (pre-1980) literature present results for continuous exposures in terms of mean exposure levels among cases and noncases, rather than in terms of relative risk estimates (or functions). If such a report supplies a cross-classification of the data by exposure levels and outcome status, crude relative risk and coefficient estimates can be computed from this cross-classification. If no such cross-classification is reported, but standard errors for the means are

[3] This was written before the idea of collecting IPD was adopted.

given, crude logistic coefficient estimates can be constructed by the linear discriminant function method' [see (10), p. 14 for details].

Summarising the risk associations of quantitative variables in epidemiological studies in a consistent form

Chêne and Thompson (12) present an approach to re-expressing results in a uniform manner. They convert results given in quantile groups or as logistic regression coefficients as a mean difference between those subjects who died and those who survived (along with the standard deviation). The appropriateness of the methods used depends on the approximate normality of the continuous variable. A method for investigating normality is given in the paper.

The outline of this method is given below (12):

Converting results given in quantile groups
Firstly, a method for calculating the mean exposure for each study is presented using weighted regression. A method for estimating mean exposure level for each group is given, even when open-ended groups are present. The mean level of exposure for subjects who died and subjects who survived, and hence the mean difference between the two, is then calculated separately for each study. This is an extension of the weighted regression model used to calculate the overall study value.[4]

Converting results given as a logistic regression coefficient
The mean difference of interest can be derived algebraically from the logistic regression coefficient (see paper). The paper suggests using a similar approach to that of Greenland (10) (summarised on page 151), if the logistic regression coefficient is not published, for estimating it from risks in the quantile groups. Here the quantile group means are estimated instead of the midpoints used by Greenland, with the belief that this approach is less arbitrary. This procedure can also be used if the regression coefficient is given but has been adjusted by confounders that vary from study to study. Once a coefficient is obtained the procedure described in this section can be used.

At this point 'a formal meta-analysis could in principle be pursued by combining the mean differences across studies, weighted inversely by their variances, and reinterpreting the pooled difference as a log odds ratio.' The problem with doing this is that usually different confounders are adjusted in different studies and this has not been accounted for.

The authors comment:

> 'The comparison of unadjusted and adjusted published (or estimated) results for particular studies is useful in providing guidance on the potential importance of confounding factors. In some cases, confounding is thought to be of relatively little consequence, and a formal meta-analysis can be pursued.' 'In other situations progress cannot be made without more detailed information from individual studies.' (12)

For example, they note a problem with matched studies, i.e. matching is broken to calculate difference.

Analysis of summary data

Heterogeneity of epidemiological studies

As for RCTs, heterogeneity can have advantages as well as disadvantages. Dickersin and Berlin comment that one of the particular advantages of meta-analysis of observational studies is that:

> 'it may permit exploratory analyses regarding associations between various study characteristics and study outcome: that is meta-analysis allows us to ask whether the associations between an exposure and a disease (or health state) observed in a single study may depend on the composition of the population under study, the level of exposure in the study population , the definition of disease employed in the study , or any of a number of measures of the methodological quality of the study.' (19)

The homogeneity assumption is less likely to be satisfied with epidemiological studies as it is with RCTs due to the inherent variation between studies. Greenland comments:

> 'One should regard any homogeneity assumption as extremely unlikely to be satisfied, given the differences in covariates, bias, and exposure variables among studies. The question at issue in employing the assumption is whether the existing heterogeneity is small enough relative to other sources of variation to be reasonably ignored.' (10)

[4] A discussion is given on whether it is appropriate to consider the standard deviations for the two groups as being the same. When it is an alternative method is presented using a pooled standard deviation.

More specifically, Morris (8) discusses meta-analysis in cancer epidemiology. He states at the time of writing (1992) that only 32 meta-analyses have been carried out on the topic. He goes on to suggest this is due to problems researchers have encountered. Epidemiological studies use a wide range of study populations and methods with a variety of measures of exposure and outcome (all which will increase heterogeneity), making them more difficult to combine than RCTs.

Tests for heterogeneity

Standard methods of chapter 8 can be used to assess heterogeneity of epidemiological studies. In addition, Greenland (10) describes a further procedure to assess the homogeneity assumption by partitioning the studies along characteristics likely to be associated with heterogeneity. See (10) for further details.

Sources of heterogeneity in epidemiological studies

Many possible sources of heterogeneity in epidemiological studies have been identified. A list of factors to be considered is given below by Blair *et al.:*

> **Study design** – 'Although similar and dissimilar study designs do not guarantee homogeneity and heterogeneity, respectively, differences in study designs can be a source of heterogeneity and should be considered as a possible explanation when study outcomes differ.'
>
> **Outcome definition** – 'If definitions of the outcome differ across studies, then there should be an attempt to obtain data from authors to achieve as much comparability as possible.'
>
> **Population type** – 'Heterogeneity in effect estimates may result from variation in types of populations included in the studies. Populations with different distributions of susceptible subgroups may experience different effects with the same exposures. Where data are available, demographic characteristics of study populations such as race, sex, and ethnicity should be obtained.'
>
> **Exposure level** – 'As part of a heterogeneity analysis, it is important to evaluate the variation of exposure between and within studies. Studies with very different exposure levels or definitions of exposure may be inappropriate for combined analyses. In such cases separate analyses may be appropriate.'
>
> **Surrogate exposures** – (environmental exposure) 'In environmental epidemiology, exposure measures often involve surrogates, while specific exposures are not clearly identified. As in the case of health outcomes discussed above, exposure should be specified as narrowly as possible to translate positive findings into effective risk reduction activities.'
>
> **Duration, intensity, frequency** – 'In conducting a meta-analysis, consideration should be given to whether combining studies that differ in duration, intensity, frequency, or routes of exposure will create heterogeneity. In such cases, conversion to a common framework or creation of nominal gradients (e.g. low, medium, high) may be feasible.'
>
> **Exposure metrics** – 'Differences in exposure metrics (between studies or over time) can be an important source of heterogeneity. Measures of exposure used in studies composing a meta-analysis should be as similar as possible or convertible to a common base and, wherever possible, exposure metrics should be quantitative.' (9)

Dealing with heterogeneity when it is present

Blair *et al.* report the following advice, from the expert working group for the application of meta-analysis in environmental epidemiology, on dealing with heterogeneity:

> 'Heterogeneity should be controlled by stratification or regression. Where unacceptable heterogeneity exists, combining disparate studies is not recommended; rather, the reasons for heterogeneity should be explored and controlled if possible. Full assessment of heterogeneity requires consideration of at least two dimensions of study estimates – the absolute magnitude of the differences among the estimates and the statistical variability of the estimates.' (9)

Blair *et al.* expand this by saying:

> 'The decision as to whether estimated differences are large enough to preclude combination or averaging across studies should depend on the scientific context not just statistical significance. For example, a 25% difference among relative risks may be considered unimportant in a study of a very rare cancer, but important in a study of a more prevalent disease. If substantive important heterogeneity is thought to be present, it should be addressed by a careful analysis of possible explanatory characteristics (covariates), by stratifying and by requesting data on the characteristics, if not available from the study reports. If the source(s) of heterogeneity is related to validity problems (confounding, bias) for one or more studies and adjustment is not feasible, then such studies should not be combined with the others.' (9)

As a side note, Dickersin and Berlin suggest:

> 'Occupational epidemiology studies, plagued by the lack of unexposed internal comparison groups, may sometimes be grouped into broad categories of exposure in order to examine the relation between exposure level and relative risk.' (19)

Statistical considerations: fixed or random effects?

Jones comments:

> 'On the whole, epidemiological studies less frequently follow standard designs than do clinical trials, and in the former potential and actual sources of bias are arguably more extensive than in the latter. Hence the appropriateness of the 'fixed effect' assumption needs to be carefully considered in the epidemiological context.' (11)

It should be remembered that:

> 'While the random-effects model allows for between-study variability, it will not correct for bias, and is less desirable than controlling for the heterogeneity.' [Blair *et al.* (9)]

Weighting of epidemiological studies

> 'The inverse of the standard error need not be the only component of the weight; e.g. discarded studies are studies with zero weight. There can sometimes be good reason for downweighting but not discarding a study, as when the uncertainty of the result is not entirely reflected by the computed standard error estimate. For example, after external adjustment, one could legitimately argue that the weight of the corrected coefficient should be less than that computed from the standard error of the original unadjusted estimate, since the original standard error reflects neither the error in estimating the correction term nor the bias from applying the correction to a new noncomparable study setting.' (10)

However, quantifying the extra uncertainty is difficult. A sensitivity analysis can go some way to alleviate this problem.

Colditz *et al.* comment on random effects weighting, and how it probably puts too much weight on large epidemiological studies (even though it weights them less than the fixed effect method):

> 'For epidemiology we need more empirical work on this point. However, theory does show us some astonishing facts. Suppose we had a study with n = 10,000 observations and the correlations between paired observations was $\rho = 0.01$, all with variance $\sigma^2(1+(n-1)\rho/n$, which on independent samples is $n = n/(1+(n-1)\rho)$, which in our example is 99. Had the original sample size been 100, the equivalent independent sample size would have been reduced only to 50, which is division by 2 instead of by 100. Such considerations suggest that we may be weighting large samples not only too heavily but much too heavily. However, before launching on a new program of estimation, we need more in the way of empirical results.' [(20), p. 376]

Methods for combining estimates of epidemiological studies

The techniques for combining estimates discussed in previous sections of this report can be used, where appropriate, for epidemiological studies. Problems may arise when small studies are combined that do not approximate normality. The studies have to be very small and a large proportion of the studies combined have to be small before it becomes a problem. 'In such cases, one may turn to other variants of large sample regression theory to derive heterogeneity and regression statistics, or choose to simply focus the meta-analysis on tabulations and graphic plots.' (10)

Dyer (21) proposed a *Z* score approach for meta-analysis of a continuous exposure. This method can be suitable for testing for effects across small studies if some suitable normalizing transformation of the exposure can be found may be useful in this instance. However, Greenland (10) points out that it does not provide a measure of exposure effect on risk, and so is unsuitable for quantifying strength of heterogeneity of effects.

Applying meta-regression to epidemiological studies

As with RCTs, meta-regression techniques (see chapter 11) can be applied to epidemiological data. Variables relating to study design can be included (case–control, cohort and other design aspects) to investigate whether different study designs tend to systematically report outcomes of different magnitudes. When subsets of studies are looked at in this manner, interpretation needs to be cautious, in view of the possibilities of inconsistent definitions, incomplete data and confounding, and over-interpretation of results.

Greenland (10) compares Cox regression with logistic regression for use in meta-analysis and discusses how to get estimates for one from the other. 'Ideally, the multiplicative–exponential model should be evaluated against various alternatives' He also comments that if a model does not fit too well, one may be able to use the mid range of values as a reasonable approximation.

Reporting the results of epidemiological studies

Chapter 18 of this report deals with reporting results of meta-analysis generally, this section deals with a few extra issues that are relevant to epidemiological studies only.

For a concise description of how to present results form epidemiological studies see Blair *et al.* (9).

> 'In addition to basic information about the studies (such as the number of cases and noncases), a review should present a table of the results of the study reanalyses, showing at least the point estimate, net correction, and standard error (or confidence interval) from each study.' (10)

'When there are many studies, even clearer is a weighted histogram of the study results'. Greenland [(10), p. 15] discusses a weighted histogram for displaying results (used also for subgroup analyses plotting a separate histogram for each group):

> 'The range of results is divided into intervals, and each study result falling within an interval contributes to that interval's bar height an amount proportional to the study's weight. The width of the intervals should not be too broad; at the very least, the range covered by a bar should be well within the confidence interval of any study contribution to that bar. For the identification of studies contribution to a bar, the bars should be divided along the vertical axis in proportion to the relative contribution of each study; certain shading schemes may also be helpful.' (10)

Blair *et al.* note that on the stratification of studies:

> 'Consensus was not reached on the degree of stratification that should be conducted in a meta-analysis. Where stratification is conducted, an important contribution in any meta-analysis is to array results both in a table and in the text by exposure metric, study design, and health outcome.' (9)

Exploration of sensitivity of the results

Blair *et al.* reported that:

> 'Several meta-analyses conducted in environmental epidemiology have directly incorporated sensitivity analysis or influence analysis, and the use of these techniques has been relatively instructive.' (9)

These studies investigated the effects of, for example, the influence of each study and length of follow up on the robustness of the results.

Greenland describes how a sensitivity analysis can be implemented:

> 'For example, one may have externally controlled for cigarette smoking in all studies that failed to control for smoking by subtracting a bias correction factor from the unadjusted coefficients in those studies. The sensitivity of inferences to the assumptions about the bias produced by failure to control for smoking can be checked by repeating the meta-analysis using other plausible values of the bias, or by varying the correction across studies. If such reanalysis produces little change in the inference, one can be more confident that the inferences appear deceptively precise relative to the variation that can be produced by varying assumptions, and thus choose to base the meta-analysis only on those studies that present results adjusted for smoking.' [(10), p. 23]

He goes on to discuss the use of influence analysis:

> 'In influence analysis, the extent to which inferences depend on a particular study or group of studies is examined; this can be accomplished by varying the weight of that study or group. Thus, in looking at the influence of a study, one could repeat the meta-analysis without the study, or perhaps with half its usual weight. In looking at the influence of a group of studies, say all case–control studies, one can again repeat the meta-analysis without them, or give them a smaller weight. If change in weight of a study produces little change in an inference, inclusion of the study cannot produce a serious problem, even if unquantified biases exist in the study. On the other hand, if an inference hinges on a single study or group of studies, one should refrain from making that inference.' [(10), p. 23]

Sensitivity analysis is covered on pages 209–10; many of the ideas there, including simulation studies, are relevant when considering observational studies.

Study quality considerations for epidemiological studies

Chapter 6 discussed the assessment of study quality and how such information could be incorporated into the analysis. This chapter also made clear that the use of quality assessment was controversial with disagreement between researchers on the appropriateness of the methods. The assessment of the quality of epidemiological studies is even more difficult than that of RCTs. No one list, similar to those used for RCTs, exists for the evaluation of the quality of epidemiological studies for meta-analysis (4), i.e. no one list is capable of assessing the different study designs. This would make weighting the studies incorporating a quality score problematic.

The assessment of study quality of observational studies is discussed on page 24 and on page 26 some checklists available are referenced. In addition, Friedenreich (4) gives references that suggest criteria to be considered when evaluating case–control studies have been published (22,23) and similarly for cohort studies (24), though these are somewhat older. Friedenreich (4) considers that too few epidemiological studies have used quality scores for a full assessment to be made on the usefulness of these scores, and goes on to comment:

> 'It has been argued that, because of the uncertainty in dealing with the real impact of many quality attributes

> on the accuracy and precision of the trial results, only major methodologic aspects of a study should be included in a quality assessment (25). The challenge for epidemiologic studies is to identify the parameters that represent the quality of the study most adequately, recognizing that these parameters may differ across different exposure-disease relations.' (4)

Her paper (4) includes table of characteristics common to all epidemiological studies and those for just case–control and cohort that could be included in a quality assessment.

Friedenreich *et al.* (26) proceeded to investigate the influence of methodologic factors in a meta-analysis using IPD for 13 case–control studies of the association between colorectal cancer and dietary fibre.

> 'The analysis was undertaken to determine whether the heterogeneity in risk estimates could be explained by methodologic and quality differences between studies.' (4)

Each study was given a quality score using a questionnaire (reproduced in the appendix of the paper). Also investigated were more specific methods covariates; the summary of which is reproduced here:

> 'Two factors, whether the diet questionnaire had been validated before use in the case–control study and whether qualitative data on dietary habits and cooking methods had been incorporated into the nutrient estimation, explained some of the heterogeneity found between studies. Risk estimates for dietary fibre and colorectal cancer were closer to the null for the studies that had these two characteristics. Quality score did not explain any between-study heterogeneity.' (26)

Dickersin and Berlin review the evidence for the effect of study quality on outcome. Their findings suggest that quality is sometimes shown to have an association with study outcome, but this is not consistent. They conclude:

> 'The intuition that poorer quality studies, or those thought to be most susceptible to 'bias' (e.g. case–control studies) tend to show larger effects than better studies, is not always supported in the data.' (19)

Fredenreich points out:

> 'Ultimately, as more pooled analyses of epidemiologic studies are performed, the influence of methodologic factors on the findings obtained from pooled analyses will be better understood.' (4)

Blair *et al.* (9) report that the expert working group supported by ILSI Risk Science Institute investigating guidelines for application of meta-analysis in environmental epidemiology failed to reach a consensus on the use of quality scores. They commented that some group members rejected any use of such scores in favour of quality-component analysis, i.e. investigating study characteristics believed to be associated with study quality separately. They suggest:

> 'Sensitivity analysis and influence analysis provide alternatives to quality scoring. A sensitivity analysis should always be conducted to determine which attributes, including quality components, are contributing to heterogeneity.' (9)

Other issues concerning epidemiological studies

Analysing IPD from epidemiological studies

MAP is covered in chapter 23 of this report. The below are a few additional notes on the literature written exclusively for epidemiological studies. Much of the methodology for carrying out an IPD meta-analysis of clinical trials can be directly translated to the epidemiology setting and hence many of the procedures described in chapter 23 can be used. See also pages 161–3 for methodology to combine matched and unmatched IPD.

Friedenreich discusses methodology for pooling epidemiological studies at the patient level. She believes the following question is of central importance, and not yet addressed:

> 'In epidemiological studies do differences in the populations and methods used in the original studies influence the results obtained from the pooled analysis?' (4)

She comments on several drawbacks on current methodology, namely: 1) little or no consideration has been given to examining how study sample, design and data collection characteristics influence the results obtained; 2) pooled analyses of epidemiological studies have not combined qualitative with quantitative assessments; 3) no pooled epidemiological studies (at the individual patient level) have included a sensitivity analysis.

The paper presents eight steps of procedures to follow and methodological issues to consider when conducting such a pooled analysis, including methods for examining heterogeneity, influence of study design and data allocation methods on the pooled results, assessment of study quality and integration of qualitative assessments in the analysis. These are summarised below.

Study selection

Friedenreich observes:

> 'For epidemiologic studies, (however) there is no predominant single characteristic on which to base a decision on whether to include or exclude a study from a pooled analysis.' (4)

For this reason, it is necessary to set specific inclusion criteria for which studies to include.

Merging dataset

Fredenreich suggests (4) that variables describing the study subjects, design and data collection methods should also be included in the data set along with the exposure and outcome variables.

She also comments that all the same possibilities for collaborative work with the original study investigators exists here as for IPD analysis or RCTs:

> 'A collaborative analysis permits the original investigators to work together, discussing causal mechanisms, generating new hypotheses, and planning further co-ordinated investigations to study these hypotheses using common study designs and data collection methods.' (4)

Analysis

Logistic regression can be used to estimate study-specific risk (in particular ORs and risk RR) estimates. Friedenreich (4) comments this method is a fixed effect method suited to pooled analyses of observational epidemiological studies. She observes that to date (1993) analyses have used individually matched data from each centre or have stratified by or controlled for 'study centre' in the analysis.

Fredenreich (4) also advises using a random effects model when heterogeneity exists. This can be achieved by including a study indicator variable as a random covariate in the model.

It should also be pointed out that using multilevel modelling techniques could be used to investigate the impact of study characteristics on results. This is achieved by modelling the mean effect of particular characteristics at the second level.

Advantages of MAP of epidemiological studies

Friedenreich observes:

> 'Confounding and interactions between established and suspected risk factors can be more readily examined, permitting more valid and precise conclusions regarding a particular exposure-disease relation than are possible with a (standard) meta-analysis.'

Also:

> 'Pooled analysis may reveal previously unrecognized errors or inconsistencies in the data and associations or dose-response effects that were either previously unknown or only suggested.' (4)

Note that to do a MAP simply, all the studies being combined need to be of the same type. If both matched and unmatched case–control studies are to be combined more consideration is needed. See Duffy *et al.* (27) (pages 161–3) on methods for combining matched and unmatched binary data.

Combining dose–response data

As Tweedie and Mengersen state (28), there are two reasons for assessing dose–response relationships in epidemiology, namely: 1) establishing such a relationship is one of several standard criteria for developing the case for the agent in question actually being harmful; if increased risk occurs with increased dose, it is a strong step in proving that a causal association exists, and 2) when an association has been established between an agent and a disease, the dose–response relationship is of crucial use in predicting the levels of risk established for individuals at different levels of exposure.

Dickersin and Berlin (19) observe that information about multiple levels of exposure may be available from within studies, as well as among studies. This can be used in meta-analyses to great advantage, particularly if standard categories have developed. For example in investigating the effect of physical activity, it could be categorised into sedentary, moderate and high activity. Evidence of a broadly defined dose–response pattern can be provided by a separate combination of RRs for high activity compared with sedentary groups and high activity compared with moderated activity groups. This type of analysis has been done for example by Berlin and Colditz (29).

If it is possible to quantify exposure levels more precisely, then more sophisticated methods are available; for example this was possible when investigating the association between alcohol consumption and breast cancer (30). Methods available when effect estimates for several precise exposure categories are available from each study, are discussed below.[5]

Data reporting results of a dose–response relationship usually appear in two forms. If the data has been modelled using a continuous variable for exposure then a single regression coefficient (along with its SE) will usually be available and can be combined using standard weighted methods. If, on the other hand, a series of risk estimates derived for corresponding exposure levels have been presented (the more common reporting procedure) (31) alternative methods are required. Methods for both these situations are discussed below. The methodology used when a combination of the above methods of reporting are used in a single meta-analysis is also discussed.

Situation A – single continuous exposure parameter estimate

One coefficient in a regression model, say β, is presented for each study that represents the change in the natural logarithm of the RR (or OR etc.) per unit of exposure. Then, provided its variance or SE is given the standard inverse variance-weighted method (outlined on pages 55–6) can be used to combine the estimates of this coefficient.[6]

Situation B – risk/odds estimate for several exposure levels

This section deals with combining dose–response data when the results are presented as in the table given below *(Table 14)* [reproduced from (31)].

It may be possible to obtain a regression slope from a report by pooling estimates for responses at different levels of exposure (or treatment) (32). However standard methods for pooling estimates assume independence of the estimates, an assumption that is never true because the estimates for separate exposure levels depend on the same reference (unexposed) group (32).

Greenland presents two methods of pooling responses at different levels that take account of the correlation between estimates. The first approach is based on constructing an approximate covariance estimate for the adjusted log ORs from a fitted table that conforms to the adjusted log ORs. A brief summary of this method is given below [reproduced from (31)]:

1. Using the crude $2 \times J$ table of margins (disease (present or absent) on one margin of the table, and J levels of exposure on the other, and that an unexposed or baseline category serves as the common reference group for $J - 1$ dose-specific estimates of the relative risk of disease) and the adjusted relative risk estimates, cell values are fitted to the body of the crude table.
2. The sum of the inverse fitted cell values in the reference exposure category is used as a first approximation to the covariance between all pairs of ln relative risks.
3. The correlation between specific relative risks is calculated using the covariances and standard errors of the ln relative risks from the fitted table.
4. The asymptotic covariance of the adjusted ln relative risk estimates is estimated by multiplying the correlation obtained in step 3 by the estimated standard errors of the adjusted ln relative risk estimates. The inverse of the resulting covariance matrix may then be

TABLE 14 *ORs for breast cancer according to duration of oral contraceptive use*

Duration of contraception use in years*	Cases	Controls	OR	95% CI†
Never (0)	96	156	1.0	
0–3 (1.5)	156	205	1.1	0.8–1.6
4–7 (5.5)	80	93	1.2	0.8–1.9
8–11 (9.5)	51	50	1.4	0.8–2.3
12 (14.4)	39	23	2.2	1.2–4.0

** Number of parenthesis is assigned in a dose–response regression model to control for other covariates*
† The SE required can be worked out from the CI (see page 149) so it does not matter which is reported.

[5] N.B. Page 151 presented a method of obtaining an overall estimate of the RR from stratified results; here these estimates are kept separate.

[6] This as for any meta-analysis is the simplest model. If for reasons such as residual heterogeneity a different model, such as that of random effects, is desired this could be used instead. Berlin *et al.* (31) give a random effects model for dose–response coefficients that is directly analogous to the model of DerSimonian and Laird for difference between effects (see chapter 10).

> used in a meta-analysis to weight the results of each study.[7]

Greenland and Longnecker comment:

> 'The objective of the above method is to approximate the logistic coefficient that would have been obtained had either more complete study data or the estimated logistic coefficient been reported, and to provide a less biased variance than was previously available. ....The primary impact of our correction method on such meta-analyses will be to alter relative weighting of the study-specific coefficients and to produce a more accurate variance estimate for the pooled coefficient estimate.' (32)

The derived estimates from this method can then be combined using the inverse variance-weighted method as described in situation A above.[8,9,10]

A second, more flexible, method is then discussed which involves pooling of study data before trend analysis. It is called the 'pool-first' method. 'The 'pool-first' method is algebraically equivalent to the method of pooling the corrected coefficient estimates from each study. The advantage of the 'pool-first' method is that it is easily extended to fitting and testing non-linear logistic models [for example, a model with a quadratic term see example 3 of (31)] [see Greenland and Longnecker (32) for details].

Greenland and Longnecker comment:

> 'The chief limitation of this method is that it cannot incorporate studies that report only a slope estimate: a study must report dose-specific odds ratios or rate ratios to be included; fortunately, such reporting is standard practice.' (32)

New developments for dose–response meta-analysis

Tweedie and Mengersen (28) present a new technique for dose–response meta-analysis. Their motivating example was a meta-analysis involving 18 epidemiological studies investigating the association between risk of lung cancer and exposure to environmental tobacco smoke (passive smoking).

They discuss three approaches for calculating dose–response estimates for each individual study (assuming results presented as in situation B):[11]

1. A non-parametric test for equality of response across dose levels (sometimes called the Armitage test for equality).
2. Imposition of an exponential model (that is, a linear trend in the logarithms of the response) and test of significance of the regression parameter. [Weighted regression model with zero intercept; essentially the same as Longnecker and Greenland (see pages 158–9).]
3. Imposition of a direct linear trend in rates of occurrence and test of significance of the regression parameter. This can be used for studies which provide numbers of cases and controls in each exposure category, and thus the analysis of actual rates of occurrence of cases is possible [in this situation (1) and (2) are also possible].

The relative merits of these approaches are discussed by the authors. One important word of warning is that under the linear model, care is needed to ensure that rates of the same magnitude are being combined. In case–control studies, if the number of controls for each case varies between studies, then the slope will also vary. Also, the paper warns that, if rate for case–control and cohort studies are being compared then one would typically get totally different orders of magnitude for the slope parameters.

Two additional issues related to dose–response data are discussed:

1. Inclusion of the unexposed group may have an important confounding effect: an observed dose–response relationship may be in fact

[7] It should be noted that, if reports included the covariance matrix estimates along with coefficients from models this method would not be necessary. (31)

[8] It is important to point out that using this method both studies presenting a single coefficient for the dose–response relationship (situation A) and ones giving separate estimates for each exposure level can now be combined.

[9] Heterogeneity tests should be carried out to assess the fit of the slope. Indeed, Greenland and Longnecker indicate that using this correction for correlation may have a greater effect on the heterogeneity analysis than the overall point estimate and SE [(31), p. 223] and present a test of subgroups of studies under the null hypothesis of no difference in slope between subgroups A and B (or the corresponding CI). A more complex model using weighted least squares regression, is given for assessing heterogeneity when a covariate describing a study characteristic is continuous or ordinal, or when several covariates are to be considered simultaneously.

[10] Note also that the exposure categories need not be the same in the different studies. (31)

[11] For algebraic details and a discussion of the relative merits of these approaches see original paper.

simply evidence of overall association but not of increasing (or decreasing) risk with increasing dose. The paper shows an example when calculating the slope including the unexposed group gives an estimate of B = 0.037 (SE = 0.002) and when it is excluded of B = 0 (SE = 0.003). Forcing the regression line to go through the origin ignores the structure of the data. In this situation, the authors suggest examination of: a) explanation of the behaviour (either biologically or through study bias), b) an analysis of response at lower dose levels, and c) identification of the influence of the unexposed group on the methodology in producing such a result.[12]

2. The inconsistency of dose measurement may influence not only within-study regressions but also across-study equivalence of regression parameters.

Smith *et al.* (33) discuss methodology for situation A outlined on page 158, when each study gives a single continuous exposure parameter estimate, but without mention of previous work on this subject. Two issues concerning dose–response models are dealt with, namely: 1) a method of weighting studies that gives greater influence to dose–response slopes that conform to the linear relation of the RR to duration (which can lead to large differences in calculated weights as a function of non-linearity); 2) the nature of the intercept of the slope in the dose–response model for each individual study. They highlight two alternatives; a model with zero intercept, and one estimating the intercept on the basis of the data (variable intercept). Selecting a model with a variable intercept implies that the risk between the two groups may differ before initial dose (10), and a model with a zero intercept implies that the risk among subjects taking very low doses is the same as the risk among untreated subjects. See original paper for formulae and computational details.

Formulae are given for the different weighting schemes used for both fixed and random effects models. The weighting schemes are derived from those used much earlier by Cochran (34) and weights can be calculated from tables in that paper.[13]

Several alternative methods estimating the SEs for mean slope estimates were investigated. These included fixed and random effects approaches as well as a components of variance model and a bootstrapping method.[14]

The methods presented were illustrated with an example of a meta-analysis of eight studies investigating the effect of oestrogen replacement therapy on the risk of breast cancer in women who experienced natural menopause. When a zero intercept fixed effects model was used, the two methods of weighting (taking the linearity of the relationship into account and not) did result in fairly large differences between mean dose–response slopes and homogeneity statistics.

The paper concluded that a random effects model (or equivalent) should be used to take into account heterogeneity (detected or not). The following extract describes the authors feeling for zero and variable slope models and their appropriate use:

> ‘Although a biological explanation would be the best way to determine whether a variable intercept is the appropriate dose–response model, uncertainty about initial risk may dictate statistical testing. On the other hand, statistical testing may not always be possible because the number of data points per study may be small. In addition, our data illustrate that with the variable-intercept model, considerable intercept heterogeneity among studies is possible and studies may be excluded from a meta-analysis because they lack enough degrees of freedom for weights to be estimated. Thus, the variable-intercept model may be more difficult to apply than the zero-intercept model. Consequently, when possible, if there are differences in risk between cases and controls at the onset of exposure, both intercept models should be used.’ (33)

The paper also warns that heterogeneity could be associated with different dose ranges for studies:

> ‘Studies with larger dose intervals will receive proportionally larger weights, given otherwise equal slopes, precision, and linearity. Combining slope values from studies with considerably different ranges of duration should be discouraged.’ (33)

DuMouchel (35) presents a general discussion of whether to fit zero (fixed) or variable slope models to meta-analysis dose–response data. He also goes on to give a brief Bayesian analysis of the data first presented by Smith *et al.* (33); he concludes:

[12] Excluding the unexposed group relaxes the fixed intercept and allows it to vary. See the paper below for another model that do not restrict the intercept.

[13] The paper suggests Cochran’s weights are slightly larger than those of DerSimonian and Laird.

[14] The paper calls for further investigation of bootstrap techniques for application in meta-analysis.

> 'Because so many sources of variation exist in the typical set of dose–response studies, the random-effects models for meta-analysis, which allow for unexplained variation in effects among studies, are to be preferred.' (35)

Applications of dose–response meta-analysis models

Steinberg *et al.* (36) implement an advanced dose–response regression model using epidemiological study designs. Their paper investigates the risk of oestrogen use for breast cancer, using an interesting regression model, which allows for duration of use. The paper took the model of page 158 as its basis. Additionally, however, when the original studies reported more than two durations of oestrogen use, the dose–response slope with an estimated intercept term was calculated. Fixed and random effect models were fitted and a more unusual bootstrap resampling method was also implemented to calculate the mean dose–response slopes and their errors. Interestingly the paper states this last method does not make any assumptions about the distribution of the study results.

Maclure (37) reports a meta-analysis investigating the association between ethanol intake and the risk of myocardial infarction. A very detailed account of a complex meta-analysis, combining case–control and cohort studies is given. The application uses quadratic terms in the dose–response meta-regression model, then goes on to look at incremental RRs and moving line regression (not new in themselves but the first time, the authors of this report believe, that they have been applied to meta-analysis). The paper also comments extensively on a deductive inference approach.

Gould *et al.* (38) investigated the relationship between cholesterol lowering and CRD via meta-analysis. They use a dose–response model with an additional parameter for intervention type to allow effects to vary over different interventions. Thompson (39) models dose–response using (a subset) of the cholesterol dataset presented in this report.

DuMouchel (35,40) carried out a meta-analysis of dose–response data of cancer studies in humans and other animals using a Bayesian approach. This approach is discussed in chapter 13 on Bayesian methods.

A final point to note when carrying out or assessing dose–response meta-analyses is that in the past, an ANOVA model for the ln RRs has been used, this model however, gives incorrect results. See Berlin *et al.* [(31), p. 220] for more details.

A note on exposure categorisation used in dose–response meta-analyses

Berlin *et al.* comment:

> 'Unfortunately, the slope estimated in the meta-analysis may be especially sensitive to the method used to assign exposure values to open-ended categories.' (31)

Occasionally better estimates can be derived using results from similar populations where known (see page 151).

Berlin *et al.* (31) note that scales that cannot be readily translated to a standard scale, such as those for ordinal measures can be analysed using dose–response methods, if the scales used are similar. Note though that the parameters from the dose–response model would not provide RRs per unit of exposure. An example of this (already mentioned on page 151) is the meta-analysis investigating the effect of physical activity on the prevention of heart disease (29). Here, physical activity had been frequently classified into three broad categories of activity: highly active, moderately active, and sedentary.

Combining matched and unmatched data

Duffy *et al.* (27) present a method for combining matched and unmatched data from RCTs, though the same methodology is directly applicable to case–control studies, including the situation of several controls per case. The motivating example for this methodology was to provide an estimate for the effect of photocoagulation on the rate of visual deterioration. Results from RCTs presenting data in either matched or unmatched form were combined.[15] *Table 15* illustrates the data structure and notation used in this methodology.

The methodology is an extension of the Mantel–Haenszel procedure, which was initially used for combining strata within a single unmatched study before being applied to meta-analysis where each strata represents a different study (see page 59). This method is used to combine the results from the *m* unmatched studies. For clarity, the formula is repeated below using the notation from the table

[15] A matched study in this situation is one where each patient has one of their eyes (selected at random) treated while the other one remains untreated.

***TABLE 15** Structure and notation for combining data from matched and unmatched studies*

	Event (deterioration)	No event (no deterioration)	Total
Unmatched studies:* ith *study of* m *studies			
Treated	a_i	b_i	$a_i + b_i$
Not treated	c_i	d_i	$c_i + d_i$
	$a_i + c_i$	$b_i + d_i$	N_i
Matched studies:* jth *matched pair of such pairs in the* kth *of* n *studies			
Treated	e_{kj}	f_{kj}	1
Not treated	g_{kj}	h_{kj}	1
	$e_{kj} + g_{kj}$	$f_{kj} + h_{kj}$	2
Matched studies:* kth *study of* n *studies***		Not treated	
Treated: event (deterioration)	x_k	y_k	
Treated: no event (no deterioration)	z_k	w_k	

* $x_k = \sum_{j=1}^{n_k} e_{kj}g_{kj}\,; y_k = \sum_{j=1}^{n_k} e_{kj}h_{kj}\,; z_k = \sum_{j=1}^{n_k} f_{kj}g_{kj}\,; w_k = \sum_{j=1}^{n_k} f_{kj}h_{kj}\,.$

$$\hat{\phi}_1 = \frac{\sum_{i=1}^{m} \frac{a_i d_i}{N_i}}{\sum_{i=1}^{m} \frac{b_i c_i}{N_i}} \qquad (19.2)$$

To combine the results from the matched studies, each matched pair within a study are treated as a stratum. By doing this, stratification by study is performed automatically. The formula for combining the n matched studies is:[16]

$$\hat{\phi}_2 = \frac{\sum_{k=1}^{n} \sum_{j=1}^{n_k} \frac{e_{kj}h_{kj}}{2}}{\sum_{k=1}^{n} \sum_{j=1}^{n_k} \frac{f_{kj}g_{kj}}{2}} = \frac{\sum_{k=1}^{n} y_k}{\sum_{k=1}^{n} z_k} \qquad (19.3)$$

A pooled estimate of both the matched and the unmatched studies can be obtained using the below formula

$$\hat{\phi}_3 = \frac{\sum_{i=1}^{m} \frac{a_i d_i}{N_i} + \tfrac{1}{2}\sum_{k=1}^{n} y_k}{\sum_{k=1}^{m} \frac{b_i c_i}{N_i} + \tfrac{1}{2}\sum_{k=1}^{n} z_k} \qquad (19.4)$$

A test statistic (based on Mantel–Haenszel's χ^2 squared statistic) is derived in the paper and CIs are also calculated by means of the variance estimate derived by Flanders [see (27) for details]. An alternative method for this CI is also given which is based on Miettinen's test-based method [see (27) for details]. An extension of the χ^2 test of homogeneity (see chapter 8) is also presented.

The paper discusses an alternative approach to combining by using the average on the logarithmic scale of RRs and hazard ratios, weighted by the reciprocals of the variances. This method could deal with differing follow-up times and could provide a 'more elegant estimation of hazard ratios and their variances.' However, IPD is required for this. The paper highlights as further work the establishment of different recommended variance estimates for use in different situations.

Moreno *et al.* (41) describe the use of (logistic) regression methods for combining matched and unmatched case–control studies, using IPD (see pages 156–7 and chapter 24). The paper notes that the 'same methodology can be applied to compare relative risk estimators for the same risk factors studies in different phases of a disease in an attempt to explore factors that may be more important in one phase than in another'. (41)

The logistic regression model proposed 'combines conditional logistic regression likelihood function

[16] For matched pair data this formula is identical to the MLE.

for the matched cases and controls and an unconditional logistic regression likelihood function for the unmatched study.' The likelihood expression that is derived[17] can be implemented in specialist software; however if there is only one case in each matched set (which is common) then the model can be simplified, so that standard logistic regression programmes can be used.[18] The model can compare several risk factors obtained in the matched and unmatched studies. The model assures that each group of cases is compared with its own control population. This method also allows the estimation of the ratio of ORs adjusted by potential confounding factors. LRTs can be performed to assess association, heterogeneity, or trend. The SEs of the coefficients allow the derivation of a Wald test and the calculation of CIs.[19]

Combining epidemiological time-series data

Katsouyanni *et al.* (42) present their intentions for a prospectively planned meta-analysis (see pages 213–14) combining epidemiological time series data (it is currently being carried out). The application for this is to provide quantitative estimates of the short-term health effects of air pollution, using a database from 10 different European countries, which represented various social, environmental and air pollution situations. Each country will firstly analyse their own data, using Poisson regression, allowing for autocorrelation and overdispersion.[20] This will allow calculation of RRs. Hence these can be pooled across sites using the standard meta-analysis techniques outlined in this report.

Although longitudinal data (such as time series) is common in both RCTs and epidemiological studies, there seem to be few examples of its inclusion in meta-analyses. In this instance, the data is to be summarised before pooling, in parallel.

Meta-analysis of cohort studies – where IPD collected, but not merged into one dataset

Dyer comments (21):

> In the study of rare diseases one may need to follow a large cohort for many years to obtain an adequate number of cases for meaningful analysis. An alternative approach is to pool data from several cohort studies. A summary of a study (21) that combined several cohort studies investigating the association between cholesterol level and death from cancer follows, and provides a framework for similar undertakings. This study used unique methodology because they chose not to merge patients into one dataset.

This study did not use standard methods for analysing IPD (described on pages 156–7) as it was felt, for a number of reasons that combining the raw data into a single file was not possible. Dyer notes that it would still be possible to use regression techniques on each study individually and then to combine, using a weighted average, standardised or non-standardised coefficients for cholesterol (i.e. by a similar method to that described above). This in fact was not done because of the 'desire to evaluate the possibility that any observed inverse association between cholesterol and cancer might result from the effect of undetected disease on cholesterol level, rather than from low cholesterol as a risk factor for cancer. This method would not have allowed for this possibility.' (21)

The method derived had to permit examination of the consistency of the association within given time periods, e.g. year-by-year. Also the study group wished to look at the association of cholesterol with risk of death at two site-specific cancers (lung and colon).[21,22] In addition the method of measuring cholesterol level differed between studies and the levels for studies varied considerably. Also, it was felt necessary to control for age. For these reasons actual cholesterol levels were replaced by age-specific *Z*-scores (see paper for details).[23] So the

[17] Estimates are found by maximising this likelihood.

[18] Very detailed practical details on implementing the relevant models are given in the paper together with some GLIM 4 macros for one case matched to a variable number of controls. For variable controls matched to variable cases specialist programs would be needed.

[19] Rich (44) discusses a regression method for meta-analysis for cohort and case–control studies combined using ML in GLIM. This work could only be found in abstract form and hence few details were given.

[20] A detailed description of which variables are to be modelled, and how (e.g. time lags, transformations etc.) is given in the paper.

[21] In some studies there were too few events to fit regression models when looking at these specific cancers.

[22] Also, if the annual numbers of death had been larger, one could have performed a regression analysis within each study year-by-year and then pooled coefficients across studies for each year of follow-up. This would have allowed an evaluation of the consistency of the association over time.

analysis consisted of comparing the mean *Z*-score for all men who died of cancer in a given year of follow-up with the mean *Z*-score for all others who were alive at the beginning of the year and did not die of cancer during that year.[24]

Discussion

When is meta-analysis of epidemiological studies useful?

As stated in the introduction to this chapter, when diseases are rare, have moderate effect sizes and long latency periods meta-analysis can increase numbers and hence the power to detect associations. Blair *et al.* (9) produced more explicit and detailed lists of when a meta-analysis may or may not be particularly useful:

A meta-analysis (of environmental epidemiological studies) may be particularly useful when:

- sources of heterogeneity are to be examined formally
- the relationship between environmental exposures and health effects is not clear
- when there are many studies but no consensus on the exposure/disease relationship
- refinement of the estimate of an effect is important
- there are questions about the generalisability of the results
- it is clear that there is a hazard, but no indication of its magnitude
- the finding from a single study is to be confirmed or refuted
- there is a need to increase statistical power beyond that if individual studies
- information beyond that provided by individual studies or a narrative review is needed.

A meta-analysis may not be useful when:

- the relationship between exposure and disease is obvious without a more formal analysis
- there is insufficient information from available studies related to disease, risk estimate, or exposure classification
- there are only a few studies of the key health outcomes
- there is substantial confounding or other biases which cannot be adjusted for in the analysis.

Problems in application of methods

Greenland (10) warns of the potential of aggregation bias or ecological bias when carrying out a meta-analysis of epidemiological studies. This bias can appear when one regresses study results on study characteristics. This type of regression can be thought of as an ecological regression model. Greenland comments that 'it is well known that ecological regression methods can lead misleading results, in that the relation between group rates or means may not resemble the relation between individual values of exposure and outcome.' He warns that this possibility needs to be considered when interpreting meta-analytic results.

Greenland also comments on another potential source of bias:

> 'Further bias can arise from regressing adjusted study results on unadjusted average values for covariates... Such bias will, however, be small unless the covariates under study are strongly associated with the adjustment factors.' (10)

Unanswered questions in meta-analysis of epidemiological studies

Spitzer (1) has created a list of questions he considers are in need of an answer. Many of the questions posed cannot be answered by 'yes/no' type answers but should be explored using a sensitivity analysis. These are reproduced (slightly abridged) below:

1. Operationally, what are the 'stringent conditions' (Fleiss and Gross' phrase) under which both case–control studies and cohort studies may be included in one single meta-analysis? Should such analyses ever be done without access to the raw data of the component studies?
2. When is it permissible to combine different types of cohort? For instance, for both exposed cohorts and comparison cohorts should one integrate data from a fixed cohort with an open one?
3. Is it permissible to integrate exposed patients sampled from hospitals with those from primary care settings?
4. For reference cohorts, **not exposed** to an intervention or risk factor, other questions arise. For example,

[23] This adjusts for the association between age and cholesterol level and eliminates inter-population differences in the mean levels of risk factors. In a similar manner it would be easy to extend this to include adjustment for other confounders.

[24] Studies of the association over several years could be studies by combining the differences for individual years.

- Is a comparison cohort from Sweden combinable with one from Italy or Japan?
- Are cohorts taken from occupational sampling frames sufficiently similar to those from the corresponding general population (or another geographically-defined one) to put them together?
- How separate in time must the accrual or demarcation of unexposed cohorts become to be ineligible for aggregation? (The question is also pertinent for exposed cohorts.)

For case–control studies:

5. Is it admissible to merge hospital-based with population-based case groups? Or in Miettinen's terms, can two or more case series be combined if they are not representative of the same type of base experience?
6. Conceptually, and in execution, is a nested case–control study similar enough to a conventional case–control study for both to be included in the same meta-analysis?
7. When there are two or more control groups in a case–control study does one merge all the control groups? If not, what criteria must one use to exclude any control group from the meta-analysis. There is no parallel between multiple arms defined by exposure in a randomised controlled trial and multiple reference samples demarcated by outcome in a case–control study.
8. Should control groups assembled by matching be combined with independent samples of referenced populations?
9. What constitutes 'proper control or adjustment for the biases that frequently occur in epidemiological studies?' [Spitzer makes lengthy comment on this]
10. Are data provided by proxy informants similar enough to data from respondents to be considered equivalent?
11. Should one include case–control studies in which data-gatherers were unblinded with blinded studies in one meta-analysis? (Should one do so in cohort research?)
12. How homogeneous must the outcome be? For instance, can one pool data from a study that ascertained 'all cancers of the lung', with one that did so only for 'oat cell Ca', or only 'adenocarcinoma'?
13. How do we interpret values and confidence intervals of single estimates derived with meta-analysis? (He goes on to pose the question; if meta-analysis of five studies has the same result and confidence interval as a single cohort, is the interpretation identical?)

Further research

The below is a list of suggestions for improving meta-analysis of epidemiological studies in the future and of research that is needed in the area.

Jones comments (11) that wider registration of epidemiological studies at their inception would greatly reduce the potential impact of publication bias. In addition, Morris (8) comments publication bias has not been systematically investigated for epidemiological studies. See chapter 16 for a general discussion about publication bias in meta-analysis.

Jones (11) calls for agreement on more uniform reporting of the results of epidemiological studies, perhaps through development of more detailed publication guidelines in this area.

Jones (11) suggests the use of sequential meta-analysis; reference to use of simulation of results of new studies before they are published, to allow meta-analysis to be updated.

Friedenreich comments:

> 'Although concerns about pooling data from unrandomised studies have been raised repeatedly, pooled analyses of epidemiologic studies have not addressed these methodologic issues, nor have they included quality assessments, perhaps because a systematic approach for these evaluations has not existed.' (4)

Friedenreich also comments:

> '.. major concern for pooled analysis (ipd), as with meta-analysis , is how to integrate qualitative assessments of the research studies with quantitative estimates of the results. There is a lack of literature on how to integrate qualitative and quantitative aspects of observational analytic studies.' (4) [Work has recently begun on this issue (43)]

Duffy *et al.* (27) highlight as further work the establishment of different recommended variance estimates for use in different situations, when combining matched and unmatched data.

Colditz *et al.* suggest that we may be weighting large samples not only too heavily, but much too heavily. 'However, before launching on a new program of estimation, we need more in the way of empirical results.' (20)

Summary

Most of the considerations for combining observational studies are the same as those outlined in the rest of the report for RCTs. One new question that needs addressing is ‘Has proper control or adjustment been made for the biases that frequently occur in epidemiological studies, such as sociodemographic or clinical differences between study populations, misclassification of subjects with regard to case–control status and to levels of exposure, factors other than the level of exposure that may affect whether a subject is a case or a control (i.e. confounding variables)’ (7). Key references on this subject are the seminal paper by Greenland (10) and the set of guidelines reported by Blair *et al.* (9). The use of sensitivity analysis to deal with the above problems is emphasised.

References

1. Spitzer WO. Meta-meta-analysis: unanswered questions about aggregating data. *J Clin Epidemiol* 1991;**44**:103–7.
2. Herbold M. Meta-analysis of environmental and occupational epidemiological studies: a method demonstrated using the carcinogenicity of PCBs as an example. *Soz Praventivmed* 1993;**38**:185–9.
3. Doll R. The use of meta-analysis in epidemiology: diet and cancers of the breast and colon. *Nutr Rev* 1994;**52**:233–7.
4. Friedenreich CM. Methods for pooled analyses of epidemiologic studies (review). *Epidemiology* 1993;**4**:295–302.
5. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis (review). *J Clin Epidemiol* 1995;**48**:167–71.
6. Anello C, Fleiss JL. Exploratory or analytic meta-analysis: should we distinguish between them? *J Clin Epidemiol* 1995;**48**:109–16.
7. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127–39.
8. Morris RD. Meta-analysis in cancer epidemiology. *Environ Health Perspect* 1994;**102**:61–6.
9. Blair A, Burg J, Foran J, Gibb H, Greenland S, Morris R, *et al.* Guidelines for application of meta-analysis in environmental epidemiology. ISLI Risk Science Institute. *Regul Toxicol Pharmacol* 1995;**22**:189–97.
10. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.
11. Jones DR. Meta-analysis of observational epidemiological studies: a review. *J R Soc Med* 1992;**85**:165–8.
12. Chêne G, Thompson SG. Methods for summarizing the risk associations of quantitative variables in epidemiologic studies in a consistent form. *Am J Epidemiol* 1996;**144**:610–21.
13. Piegorsch WW, Cox LH. Combining environmental information 2. environmental epidemiology and toxicology. *Environmetrics* 1996;**7**:309–24.
14. Bross IDJ. Pertinency of an extraneous variable. *J Chron Dis* 1967;**20**:487–95.
15. Yanagawa T. Case-control studies: assessing the effect of a confounding factor. *Biometrika* 1984;**71**:191–4.
16. Copeland KT, Checkoway H, McMichael AJ. Bias due to misclassification in the estimaton of relative risk. *Am J Epidemiol* 1977;**105**:488–95.
17. Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics* 1997;**33**:414–18.
18. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. *Int J Epidemiol* 1983;**12**:93–7.
19. Dickersin K, Berlin JA. Meta-analysis: state-of-the-science (review). *Epidemiol Rev* 1992;**14**:154–76.
20. Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: commentary. *Am J Epidemiol* 1995;**142**:371–82.
21. Dyer AR. A method for combining results from several prospective epidemiological studies. *Stat Med* 1986;**5**:307–17.
22. Lichtenstein MJ, Mulrow CD, Elwood PC. Guidelines for reviewing case-control studies. *J Chron Dis* 1987;**40**:893–903.
23. Horwitz RI, Feinstein AR. Methodologic standards and contradictory results in case-control research. *Am J Med* 1979;**66**:550–64.
24. Feinstein AR. Twenty scientific principles for trohoc research. In: Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders, 1985, p. 543–7.
25. Detsky AS, Naylor CD, O’Rourke K, McGeer AJ, L’Abbe KA, O’Rourke K, *et al.* Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol* 1992;**45**:255–65.
26. Friedenreich CM, Brant RF, Riboli E. Influence of methodologic factors in a pooled analysis of 13 case-control studies of colorectal cancer and dietary fiber [published erratum appears in *Epidemiology* 1994;**5**:385] (review). *Epidemiology* 1994;**5**:66–79.

27. Duffy SW, Rohan TE, Altman DG. A method for combining matched and unmatched binary data. Application to randomized, controlled trials of photocoagulation in the treatment of diabetic retinopathy. *Am J Epidemiol* 1989;**130**:371–8.

28. Tweedie RL, Mengersen KL. Meta-analytic approaches to dose-response relationships, with application in studies of lung cancer and exposure to environmental tobacco smoke. *Stat Med* 1995;**14**:545–69.

29. Berlin JA, Colditz GA. A meta-analysis of physical activity in the prevention of coronary heart disease. *Am J Epidemiol* 1990;**132**:612–28.

30. Longnecker MP, Berlin JA, Orza MJ, Chalmers TC. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 1988;**260**:652–6.

31. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. *Epidemiology* 1993;**4**:218–28.

32. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am J Epidemiol* 1992;**135**:1301–9.

33. Smith SJ, Caudill SP, Steinberg KK, Thacker SB. On combining dose-response data from epidemiological studies by meta-analysis. *Stat Med* 1995;**14**:531–44.

34. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;**10**:101–29.

35. DuMouchel W. Meta-analysis for dose-response models (comment). *Stat Med* 1995;**14**:679–85.

36. Steinberg KK, Smith SJ, Thacker SB, Stroup DF. Breast cancer risk and duration of estrogen use: the role of study design in meta-analysis. *Epidemiology* 1994;**5**:415–21.

37. Maclure M. Demonstration of deductive meta-analysis: ethanol intake and risk of myocardial infarction. *Epidemiol Rev* 1993;**15**:328–51.

38. Gould AL, Rossouw JE, Santanello NC, Heyse JF, Furberg CD. Cholesterol reduction yields clinical benefit: a new look at old data. *Circulation* 1995;**91**:2274–82.

39. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction (review). *Stat Methods Med Res* 1993;**2**:173–92.

40. DuMouchel WH, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species (with comment). *J Am Statist Assoc* 1983;**78**:293–308.

41. Moreno V, Martin ML, Bosch FX, De Sanjose S, Torres F, Munoz N, *et al.* Combined analysis of matched and unmatched case-control studies: comparison of risk estimates from different studies. *Am J Epidemiol* 1996;**143**:293–300.

42. Katsouyanni K, Zmirou D, Spix C, Sunyer J, Schouten JP, Ponka A, *et al.* Short-term effects of air pollution on health: a European approach using epidemiological time-series data. The APHEA project: background, objectives, design. *Eur Respir J* 1995;**8**:1030–8.

43. Economic and Social Research Council. Meta-analysis of qualitative and quantitative evidence. Analysis of large and complex datasets, 1997.

44. Risch HA. A unified framework for meta-analysis by maximum likelihood. *Am J Epidemiol* 1988;**128**:906.

Chapter 20

Meta-analysis of survival data

Introduction

In many areas of healthcare research the main outcome of interest is time to an event. For example, in cancer the event of interest is often death, though it could also be recurrence of disease. In a transplant setting, the event could be failure of a graft or organ. In nursing related studies the event is often discharge from hospital. In all of these settings, although the event of interest is the time from entry into a study/treatment/admission to the event in question, for some patients this event may not have been observed at the time of analysis/data collection, it only being known that it may/will occur later at some time point beyond a certain point. Such patients are termed censored. This censoring makes the analysis of this type of data unique and often complex (1). Indeed, if no censoring occurred pooling mortality rates using standard methods (such as Peto's method, chapter 9) would be possible. For an example of when censoring did occur, and the above method incorrectly used see (2). Additionally, Abel and Edler (3) warn of the danger in calculating RRs and combining them if the follow-up duration of the studies is different. They suggest working RRs out for different times follow-up times, only using studies whose follow-up extends beyond the time point being examined. If the RRs within each study are not similar, pooling results in this way would give a bias result.

Regarding the application of statistical techniques to survival studies in general, Peto *et al.* have observed that, 'if the course of the disease is very rapid .... and it is unimportant whether a dying patient lives a few days longer or not, a count of the numbers of deaths and survivors on each treatment is all that is required. However, if an appreciable proportion of the patients do eventually die of the disease, but death may take some considerable time, it is possible to achieve a more sensitive assessment of the value of each treatment by looking not only at how many patients died, but also at how long after entry they died.' (4)

It is beyond the scope of this chapter to describe the methods of survival analysis for a single study. See Collett, and Parmar and Machin (5,6) for very readable texts on the subject.

Several different approaches to combining survival data are reported in this chapter. Which of these methods is most suited to a given situation is largely dictated by the type of data that is available for each of the studies to be combined. It should be noted that obtaining accurate data is often a problem, it is sometimes necessary to extract information from Kaplan–Mier curves presented in papers. These may be inaccurate, difficult to read, or simply small, all of which will reduce the accuracy of the data.

The techniques which are available for different situations include weighting and combining survival-rate differences at a fixed point(s) in time (pages 169–72), calculating and combining a summary parameter describing the survival curves (pages 171–3), combining 'log-rank' ORs (pages 173–4), and combining data on individual patients from different studies (page 174). It is considered that the analysis of IPD the ideal (7). An approach not covered here is a confidence profile survival model described by Eddy (8); the general confidence profile approach is, however, covered in page 202. Additionally see pages 209–10 for details for meta-analysis of surrogate measures of survival.

Inferring/estimating and combining hazard ratios

The simplest way to carry out a meta-analysis of survival data would be to summarise each contributing trial by a single number, along with its SE, and use standard methods of meta-analysis to combine them (9).Whitehead and Whitehead (10) present a method for combining estimates of hazard ratios of an assumed proportional hazards model. Ideally, the hazard and survival functions should be known, along with the number of events and times of those events for both groups of patients in each study to be combined.[1] Unfortunately, published reports seldom report sufficient detail for all these to be determined.

[1] No details of this are given here, as the methodology is part of the general parametric approach presented in the paper which follows a different notation from the notation of this report. See original for derivation.

If the value of the χ^2 statistic for the log-rank test is quoted along with the total number of events in the study, then an approximation to the values required for the synthesis can be derived. Alternatively, if Cox's proportional hazards model has been fitted, and the coefficient corresponding to the treatment quoted, then again an approximate value for the sufficient statistics required for the synthesis can be obtained.[2]

Calculation of the 'log-rank' OR of meta-analysis

This method is reviewed by Messori and Rampazzo (11), the techniques having previously been discussed in part (12–14). The method combines standard two-arm RCTs comparing two treatments, A and B. As the authors report:

> 'The meta-analysis has the purpose of combining the results of the trials to generate an overall index of relative effectiveness, expressed in terms of an OR' (11)

The method is briefly described below:

Split time into j–1 consecutive time intervals, which must be the same across the various clinical trials, but whose duration need not be constant. Then work out O–E (the observed deaths minus the expected number of deaths) for both treatment groups (using standard methods).

The values of O–E and its variance are summed over all k studies separately for each time period. These summed values divided by their standard deviation (the square root of the summed variances of O–E) provide a test statistic compared to the normal distribution. Each test statistic compares the survival of groups A and B for its respective time interval.

As well as describing the above methodology, Messori and Rampazzo (11) also presented a test to explore the heterogeneity between trials. In addition, interaction and trend tests for indirect comparisons of trial subgroups are given. In this example the interaction between the timing of radiotherapy and chemotherapy was examined, and trend tests across age groups were carried out (though the same tests could be used on any subgroups.)

Comparison of efficiency of Mantel–Haenszel with log-rank method

Buyse and Ryan (15) compare the asymptotic relative efficiency (ARE) of the Mantel–Haenszel test with respect to the stratified log-rank test, and computes the ARE in situations which are likely to be of practical interest. The motivation for doing this is to compare the situation when one has IPD with time till death, censoring time and a log-rank test for survival curves can be performed, with that when only summary data is available, and the Mantel–Haenszel method of comparing has to be used. They pose the question; how less efficient is using the summary data only?

Buyse and Ryan report:

> 'Because it assumes that the odds of dying on the two treatments are in a constant ratio across trials, the Mantel–Haenszel test may not be the theoretically optimal way of combining proportions of death.' however 'the Mantel–Haenszel test has high efficiency relative to the stratified log-rank test if the proportions test has high efficiency relative to the log-rank test for each individual trial. The results of the efficiency calculations suggest that in many realistic situations the loss of efficiency incurred by using the proportions test instead of the log-rank test is not excessively large and may be surprisingly small in some cases.'

However, they suggest:

> 'The analysis of several clinical trials should be based on full survival data whenever possible. Not only is this approach most powerful, it also provides insight into the course of the disease and treatment effect over time.' (15)

They go on to say, if one has got individual patient data from some of the trials, the assumption of proportional hazards can be assessed and form results in this paper an assessment of the potential gain in power can be ascertained if the full survival information were obtained from all studies.

Calculation of pooled survival rates

This method, described by Coplen *et al.* (14), simply pools the individual survival rates from the different studies at specified values of time. The equation required is:

[2] Whitehead and Whitehead (10) caution that authors may not use the same terminology and so identification of the appropriate statistics may be difficult. They also comment that further reliance is placed on the accuracy of their calculations.

$$P_t = \frac{\sum_{j=1}^{k} (S_{tj} W_{tj})}{\sum_{j=1}^{k} W_{tj}} \qquad (20.1)$$

where P_t is the pooled survival rate at time t (i.e. the proportion of patients surviving at time t estimated by the meta-analysis), S_{tj} is the survival rate at time t in the jth study (S_{tj} can be either the Kaplan–Meier estimate or an actuarial estimate,[3] W_{tj} = 1/(variance of S_t), and k = the number of trials included in the meta-analysis. The variance of S_t for each of the j trials can be calculated by Greenwood's formula:

$$\mathrm{Var}(S_t) = \sum_{i=1}^{h} \frac{D_t}{N_t(N_t - D_t)} \qquad (20.2)$$

where h is the number of time intervals into which the follow-up from time 0 to time t has been divided in to survival analysis (e.g. intervals may be one year each) D_t = the number of deaths during an individual time interval, and N_t = the number of patients at risk during the same interval.

Iterative generalised least squares for meta-analysis of survival data at multiple times

Dear (9) reports, the motivation for the need for this model over the method of combining hazard ratios (pages 169–70) is two-fold: 1) the difficulty of finding the necessary data of sufficient quality in trial reports to combine hazard ratios and 2) the combining hazard ratios approach cannot include single-arm studies in the analysis. For these reasons modelling the original survival data is preferred.

Dear presents an analysis of survival proportions reported at multiple times (e.g. yearly intervals). Generalised least squares (GLS) is used to fit a linear model including between and within trial covariates.[4] This is an extension of the methodology of Raudenbush *et al.* (16) (see chapter 22); here the multiple outcomes are viewed as the same outcome reported repeatedly. It also incorporates multi-arm studies and non-randomised historical controls, and thus also combines the methodological advances of Begg and Polite (17) (see pages 201–2).

The model allows survival data reported at multiple times during a trial to be analysed together. It uses comparisons between the models to test hypotheses about the effects of the treatments on the various outcomes. An iterative procedure is used to derive correlations between times within studies so unlike the multiple outcomes analyses they do not need estimating beforehand.[5,6]

One limitation of this approach is that it cannot incorporate a random effects baseline term, unlike the Begg and Pilote model (17). Dear comments 'the attractiveness of a random-effects model remains, ..., as a way of extracting information from the overall level of single-treatment trials.' (17)

Dear (9) also comments on the possibility of an alternative formulation of this problem using a logistic model through the use of generalised estimating equations (GEE). It would not be necessary to use moment estimators of the covariance, since the variances are assumed known from the reports contributing to the meta-analysis, and approximate correlation estimates are available as functions of the fitted values.[7]

Application

Two examples are given in the paper by Dear (9). One is a meta-analysis of 14 studies on patients with myelogenous leukaemia. Six of the studies compared two treatments, allogeneic bone marrow transplantation (BMT) and chemotherapy. Two included only BMT, and the remaining six included only chemotherapy. *Table 16* gives the empirical

[3] This can be estimated from Kaplan–Meier curves, if these have been included in the report, or if life tables have been included, from these. If neither of these are available then it is not possible to use this method.

[4] Multivariate normality is assumed for inferences.

[5] Code written in SAS/IML matrix language to carry out this analysis is available from author.

[6] See original paper for derivation and formulae.

[7] The paper discusses the contribution of historical controls to this kind of analysis, when the overall level of their data is not permitted to contribute directly to the estimation of the treatment difference. Here, historical controls contribute to the shape of the survival curve, i.e. the changes in survival between successive years. Hence, this information affects the estimated treatment difference and also the covariances. It is worth noting a difference between this model and the calculation of log-rank OR of meta-analysis. Here, the multiple time points are incorporated into a single model. On page 170, a separate analysis was done at each desired time. This is highlighted in the example on pages 171–2.

***TABLE 16** Per cent disease-free survival (SE in parentheses) by year (1–5) [adapted from Dear (9)]*

	BMT					**Chemotherapy**				
Trial	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	49(12)	46(12)	42(12)	40(12)	40(12)	54(8)	25(8)	23(7)	23(7)	23(7)
2	55(10)	50(10)	36(9)			40(8)	23(7)	23(7)	23(7)	
3	54(10)	47(13)	40(13)	40(13)		54(9)	42(8)	28(8)	28(8)	
4	70(23)	70(23)	70(23)	70(23)		48(17)	48(17)	17(13)		
5	54(4)	46(5)	42(6)			40(5)	21(4)	16(4)	16(4)	
6	54(2)	43(3)	40(3)	39(3)		50(4)	32(4)	24(4)	18(4)	
7	59(8)	49(9)	47(9)	47(9)	47(9)					
8	61(8)	53(8)	53(8)	53(8)	53(8)					
9						60(9)	48(9)	32(9)	32(9)	32(9)
10						44(5)	26(4)	17(5)	16(4)	
11						50(3)	33(3)	26(3)	22(3)	19(3)
12						62(3)	38(3)	29(3)	24(3)	22(3)
13						50(10)	24(8)	16(7)	12(6)	
14						76(7)	53(8)	53(8)	50(8)	50(8)

***TABLE 17** Pooled results of chemotherapy versus BMT example – % disease-free survival (SE in parentheses) [adapted from Dear (9)]*

				Survival difference	
Year	**BMT**	**Chemotherapy**	**Hazard ratio**	**GLS model**	**Begg *et al.***
1	59.0 (2.6)	53.2 (2.1)	1.1	5.7 (3.1)	2 (3)
2	49.6 (2.9)	33.6 (2.0)	2.3	16.0 (3.3)	13 (3)
3	45.9 (3.0)	26.1 (2.0)	3.0	19.8 (3.4)	16 (3)
4	45.1 (3.0)	22.8 (2.0)	7.3	22.3 (3.4)	21 (3)
5	45.1 (3.0)	20.8 (2.1)	∞	24.3 (3.4)	(not analysed)

probabilities of disease-free survival at five 1-year intervals after the start of the treatment (where available).

Each line shows results from one clinical trial on patients with acute myelogenous leukaemia. The first six trials compared BMT with chemotherapy; the other eight trials tested only one of these alternative therapies.

Previously, these data were analysed and results were summarised for the first 4 years of follow up separately. This analysis permits them to be jointly analysed. Two advantages of this are that: a) it is possible to test the hypothesis for an interaction between treatments (incorporating all the studies); and b) the treatment effect can be estimated with greater precision by combining information between years. Indeed in this example an additive year effects were found to significantly improve the fit of the model.

The results obtained for this example are displayed in *Table 17*.

Meta-analysis of failure time data

Hunink and Wong (7) present a method for combining failure time data from various sources, adjusting for differences in case-mix among studies by the use of covariates.[8] The approach uses a proportional-hazards model (which is a key assumption[9]) and the actuarial life-table approach,[10] and is capable of combining results from non-controlled cohort studies as well as controlled studies. Note, that this is a fixed effects procedure, and also to use the method enough data must be available to construct life tables – this may not always be possible from study reports.

A summary of the steps involved in applying the technique is given in *Box 5* below (see original paper for further details and for an example of its application).

BOX 5 Summary of method of Hunink and Wong for failure time data

1. Summarise the available data in the form of life tables (different table for each study).
2. Estimate the hazard-rate ratio for each covariate, and calculate the hazard-rate ratio for each stratum, defined by combinations of the covariates compared with the reference stratum (calculated across all studies).
3. Combine the data to estimate the hazard-rate of the reference stratum for every interval.
4. Calculate the hazard-rate and survival curve for every stratum.

   Additional step: for studies that are not fully stratified with respect to the covariates, the effective sample size in each stratum in each interval must be estimated. Adjusted formulae in step three are then used.

Once this is carried out, a sensitivity analysis is recommended using techniques such as: 1) varying the censoring rates among studies and subgroups, 2) determining the contribution of each study using a jack-knife type of sensitivity analysis, 3) stratifying the studies by level of detail presented, and including studies with decreasing levels of detail at each step, 4) Monte Carlo analysis may be performed to derive empirical estimates of the SEs and CIs.

Identifying prognostic factors using a log(relative risk) measure

Voest *et al.* report a study where the objective was to test a new parameter determined by overall survival: the log(relative risk) (LRR). 'This parameter makes it possible to reduce the whole survival curve to one single figure (as the method on pages 169–70 did). This appears more accurate than the use of a 5-year survival rate which is only a single point in the entire survival curve' (18).

The LRR is based on the assumption of proportional hazards (as used in Cox regression model).

Method

All [-log(survival)] curves of the treatment groups were evaluated (59 in this instance) and plotted. Using these curves an average curve is computed. For each regimen the mean difference between its curve and the mean curve is computed. This yields the LRR index for the given regimen.

Note that the paper uses the method (and derived measure) to detect prognostic factors rather than compare treatments, so the LRR can be viewed as a summary parameter describing the survival curves.

Pooling independent samples of survival data to form an estimator of the common survival function

Srinivasan and Zhou (19) consider how to efficiently combine from different samples (with censored end-points) to form an estimator of the common survival function. The authors discuss two approaches in parametric settings: 1) take an optimal weighted average (inversely proportional to the dispersion matrices of the individual estimators) of the two estimators from the two independent studies, and 2) pool the data, form the joint likelihood function and find the MLE of the data from the joint likelihood. Additionally they report, for a non-parametric set-up one could do the same but the pooled

[8] The authors assume any heterogeneity between studies is caused by case mix. Sensitivity analysis will show whether additional unexplained variation exists.

[9] If the proportion hazard assumption is doubted, a primary dataset could validate its use.

[10] This assumes a constant hazard rate per time interval. The life table approach is taken because the alternative Kaplan–Meier approach needs IPD. This model implies survival follows an exponential distribution within each interval.

estimator is strictly superior.[11] This paper gives no example, it may be difficult to implement in practice if the data required for the dispersion matrices cannot be obtained or derived.

Combining quality-of-life-adjusted survival data

Cole *et al.* (20) present a methodology for meta-analysis to compare treatments in terms of quality-of-life-adjusted survival that does not require individual patient-level data. It allows one to investigate the trade-off between treatment toxicity and improved outcome. The motivating example for this method was to determine whether the benefit of adjuvant chemotherapy treatment outweighs its costs in terms of toxic effects [see (21) for a full report of applying the methodology to that application]. A Q-TWiST (quality-adjusted time without symptoms or toxicity) analysis was carried out on each trial. This measure allows one to make treatment comparisons that incorporate differences in quality of life associated with various health states.[12]

These outcomes were then combined using regression models for recurrence-free survival and overall survival (individually).The model presented adjusts for the differing follow-up interval lengths, while incorporating the estimated covariance of the restricted means for each trial. In addition, the assumed 6-month mean duration of toxicity due to chemotherapy is recorded in the overview results.[13]

It should be noted that no specific value judgements were put on the quality of life associated with the time periods of toxicity due to chemotherapy and time spent with disease recurrence.

> 'Instead, we assigned arbitrary utility coefficients to these periods and expressed the overall treatment comparisons in a threshold utility plot according to all possible values of the coefficients. One can use such plots to assist patients and physicians in making treatment decisions, providing that the clinical trials in the meta-analysis consider similar treatments in similar patient populations.' (20)

Meta-analysis of survival data using IPD

For a general explanation of MAP see chapter 24.

The first example of a meta-analysis being performed for survival outcomes on IPD was probably that by the Early Breast Cancer Trialists' Collaborative Group (13) to investigate the effects of adjuvant tamoxifen, cytotoxic chemotherapy, radiotherapy and ovarian ablation on survival after breast cancer.

All the methods presented in this chapter so far have been developed to combine data aggregated to some level, whether that be a single parameter estimator for the study or periodic summaries, such as those obtained by life tables. However, as mentioned in the introduction, the superior way to carry out a meta-analysis of survival data is to use IPD. Indeed, some believe that this is the only reliable way to carry out a meta-analysis of survival data (22). No methodology (with the exception of a test for balanced follow-up[14]) on this subject, to the authors' knowledge, has been published for survival data; however the methodology would not necessarily need to be very different from survival analysis of a single study.

Further research

Hasselblad:

> 'A meta-analytical technique that adjusts for covariates and provides survival curve estimates for every stratum defined by combinations of covariates would be useful.' (23)

A random/mixed effects model, which in principle would not be difficult for the methods of pages 169–70.

[11] The authors then go on to discuss a simpler case where one has a common life distribution, but possibly different random right censoring patterns. Again, they show that combining study estimates rather than data can cause a substantial loss of information.

[12] See (24) for a thorough account of analysing quality adjusted survival data.

[13] See paper for details of the model.

[14] For survival (or other time dependent outcomes) it is necessary to check follow up is up to date and balanced across treatment arms. This balance in follow up can be checked by selecting all patients outcome-free and using the date of censoring as the event to carry out a 'reverse Kaplan–Meier' analysis producing censoring curves which should be the same for all arms of the trial (25).

It would be desirable to tie all the above methodology into a cohesive form.

Summary

Survival analysis data requires specialist meta-analysis techniques (as well as specialist statistical methods in general) because of data censoring. If this censoring is ignored this may bias the overall estimates. Other than this problem, the various standard approaches for meta analysis are possible. In such instances methods such as finding summary measures for survival data (such as the hazard ratio), and then combining those is possible.

References

1. Abrams KR, Everitt B, Dunn G, editors. In: Regression models for survival data. Recent advances in the statistical analysis of medical data. London: Arnold, Chapter 11, 1998.

2. Adamson GD, Pasta DJ. Surgical treatment of endometriosis-associated infertility: meta-analysis compared with survival analysis. *Am J Obstet Gynecol* 1994;**171**:1488–505.

3. Abel UR, Edler L. A pitfall in the meta-analysis of hazard ratios. *Controlled Clin Trials* 1988;**9**:149–51.

4. Peto H, Pike MC, Armitage P. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. *Br J Cancer* 1977;**35**:1–39.

5. Parmar MKB, Machin D. Survival analysis: a practical approach. Chichester: John Wiley & Sons, 1995.

6. Collett D. Modelling survival data in medical research. London: Chapman & Hall, 1994.

7. Hunink MGM, Wong JB. Meta-analysis of failure-time data with adjustment for covariates. *Med Decis Making* 1994;**14**:59–70.

8. Eddy DM, Hasselblad V, Shachter R. Meta-analysis by the confidence profile method. San Diego: Academic Press, 1992.

9. Dear KBG. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* 1994;**50**:989–1002.

10. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. *Stat Med* 1991;**10**:1665–77.

11. Messori A, Rampazzo R. Metaanalysis of clinical-trials based on censored end-points -simplified theory and implementation of the statistical algorithms on a microcomputer. *Comp Methods Prog Biomed* 1993;**40**:261–7.

12. Pignon JP, Arriagada R, Ihde DC, Johnson DH, Perry MC, Souhami RL, *et al.* A meta-analysis of thoracic radiotherapy for small-cell lung cancer. *N Engl J Med* 1992;**327**:1618–24.

13. Early Breast Cancer Trialists' Collaborative Group. Treatment of early breast cancer. Volume 1: worldwide evidence 1985–1990. Oxford: Oxford University Press, 1990.

14. Coplen SE, Antman EM, Berlin JA, Hewitt P, Chalmers TC. Efficacy and safety of quinidine therapy for maintenance of sinus rhythm after cardioversion: a meta-analysis of randomized control trials. *Circulation* 1990;**82**:1106–16.

15. Buyse M, Ryan LM. Issues of efficiency in combining proportions of deaths from several clinical-trials. *Stat Med* 1987;**6**:565–76.

16. Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. *Psychol Bull* 1988;**103**:111–20.

17. Begg CB, Pilote L. A model for incorporating historical controls into a meta-analysis. *Biometrics* 1991;**47**:899–906.

18. Voest EE, Van Houwelingen JC, Neijt JP. A meta-analysis of prognostic factors in advanced ovarian cancer with median survival and overall survival (measured with the log (relative risk)) as main objectives. *Eur J Cancer Clin Oncol* 1989;**25**:711–20.

19. Srinivasan C, Zhou M. A note on pooling Kaplan–Meier estimators. *Biometrics* 1993;**49**:861–4.

20. Cole BF, Gelber RD, Goldhirsch A. A quality-adjusted survival meta-analysis of adjuvant chemotherapy for premenopausal breast cancer. International Breast Cancer Study Group. *Stat Med* 1995;**14**:1771–84.

21. Gelber RD, Cole BF, Goldhirsch A, Rose C, Fisher B, Osborne CK, *et al.* Adjuvant chemotherapy plus tamoxifen compared with tamoxifen alone for postmenopausal breast cancer: meta-analysis of quality-adjusted survival. *Lancet* 1996;**347**:1066–71.

22. Clarke MJ, Stewart LA. Systematic reviews – obtaining data from randomized controlled trials – how much do we need for reliable and informative meta-analyses. *BMJ* 1994;**309**:1007–10.

23. Hasselblad V, Mosteller F, Littenberg B, Chalmers TC, Hunink MGM, Turner JA, *et al.* A survey of current problems in metaanalysis – discussion from the agency for health-care policy and research inter-port work group on literature-review metaanalysis. *Med Care* 1995;**33**:202–20.

24. Billingham L, Abrams KR, Jones DR. Quality of life assessment and survival data (QOLAS). NHS HTA 93/50/04, 1997.

25. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Cochrane Working Group. Stat Med* 1995;**14**:2057–79.

Chapter 21

Meta-analysis of diagnostic test accuracy

By Martin Hellmich[1]

Introduction

A 'diagnostic test' may most generally be defined as 'any measurement aimed at identifying individuals who could potentially benefit from intervention' [Cochrane Methods Working Group (CMWG) (1)]. Though randomised trials of screening may be used to assess the effectiveness of a test regarding patient outcome, the conduct of such trials is infeasible for reassessment of every new test. Hence the focus of investigation is on the test's accuracy to detect conditions for which randomised trials show effective intervention [CMWG (1)]. Readers unfamiliar with diagnostic test methodology may find Swets and Pickett (2), Hanley (3), Begg (4), Abel (5), Campbell (6) useful as general introductory or reference texts.

Comprehensive review articles on meta-analysis of diagnostic test accuracy were written by Irwig *et al.* (7–9), Hasselblad and Hedges (10) and Shapiro (11); Ohlsson (12) dealt with it in the general systematic review context. Irwig *et al.* (4) is recommended as a good introductory text on the subject. The CMWG on Systematic Review of Screening and Diagnostic Tests is currently compiling recommended methods and key references [CMWG (1)]. The group has summarised the objectives of systematic reviews of diagnostic test accuracy for a specific condition as follows.

Objectives

- Identification of number, quality and scope of primary studies.
- Overall summary of diagnostic accuracy.
- Comparison of different tests in terms of their accuracy.
- Determination whether (how) accuracy estimates depend on study quality.
- Determination whether accuracy differs in subgroups defined by patient and test characteristics (applicability or generalisablity).
- Raising further research issues and highlighting deficits.

The group also gives practical guidelines (checklists) for literature retrieval, quality appraisal, data presentation and analysis (the latter will be dealt with below).

Systematic reviews of test accuracy or effectiveness are included in the database of the NHS CRD, University of York, York Y01 5DD, UK.

As in the clinical trial setting, meta-analysis of diagnostic test accuracy involves three steps: find, appraise, and combine all studies relevant for a specific question. However, the various sources of bias in the assessment of diagnostic tests necessitate special efforts and methods to correct for them. Specifically, valid meta-analyses of test accuracy with the rationale 'to obtain valid summary estimates and provide information on factors affecting estimates to help readers decide how to generalise results to their settings' [Irwig *et al.* (9)] require:

1. No publication bias in the set of primary studies (see chapter 16).
2. No bias due to poor quality of the primary studies.
   - A good (acceptable) reference (gold) standard.
   - Test(s) and reference standard are to be read blind of each other.
   - Verification by reference standard for either all patients who underwent the index test or a stratified random sample of them (with adjustment for sampling fractions) (otherwise verification bias would result).
   - If two or more tests are compared they should each be performed on all patients or patients should be randomly allocated to them.
   - Assessment of the effect of reported design flaws on estimates of diagnostic accuracy (e.g. lack of blinding may cause overestimation of accuracy while a certain type of verification bias may render the contrary).
3. Variation in test threshold between the primary studies (if present) is accounted for.

[1] Visiting research fellow at the Department of Epidemiology and Public Health, University of Leicester, funded by the German Academic Exchange Service (DAAD).

4. The estimates from meta-analysis are generalisable (applicable) to the clinical problem at hand. This depends on:
   - the details of the test applied;
   - what other tests have been done before;
   - the patient spectrum (spectrum of abnormality and normality in the diseased and non-diseased groups, respectively); and
   - random effects models to account for between-study variability.
5. No bias caused by referral of false positives from primary to secondary or tertiary care [Sackett (9); Discussion].

The specific statistical methodology appropriate for meta-analysis of test accuracy will be outlined in the next section. This will be followed by a brief discussion and pointers for further research.

Statistical methods

Diagnostic tests may be differentiated by the type of their outcomes – binary, ordered categorical, or continuous (see chapter 14 for an explanation of these data types). For each outcome type various (specific) meta-analytic procedures have been proposed. The exposition presented in *Table 18* borrows from Irwig *et al.* (9).

TABLE 18 *2 x 2 table for a binary test*

		Reference test	
		+	–
Index test	+	***a***	***b***
	–	***c***	***d***
		n_1	n_2

Binary test results

Suppose, $n = n_1 + n_2$ subjects undergo both the index and the reference test. The binary test results may be condensed in a 2 × 2 table (see *Table 18*). The table corresponds to a certain threshold (cut-off, positivity criterion) c such as

$$\text{index test} = \begin{cases} + : \text{(latent) test variable} \geq c, \\ - : \text{(latent) test variable} < c, \end{cases} \qquad (21.1)$$

where the '(latent) test variable' is either observable or non-observable.

Widely used indices of test accuracy are

$$\begin{aligned} \text{TPR} &= a / n_1 \\ \text{FPR} &= b / n_2 \\ \text{FNR} &= c / n_1 \\ \text{TNR} &= d / n_2 \end{aligned} \qquad (21.2)$$

which are monotone in c. The OR

$$\text{OR} = \left(\frac{\text{TPR}}{1 - \text{TPR}} \right) \Bigg/ \left(\frac{\text{FPR}}{1 - \text{FPR}} \right) \qquad (21.3)$$

measures the discriminatory power of the index test and may also vary with the chosen threshold c.

Suppose a collection of k different 2 × 2 tables (studies) is to be summarised. Midgette *et al.* (13) suggested weighted averages of TPR_i and FPR_i $(i = 1,2,\ldots,k)$ provided these are not positively correlated (Spearman test) but homogeneous (χ^2 or extended exact Fisher test – see chapter 8). Heterogeneous data should not be combined (except maybe within subgroup analysis). If the TPR and FPR values are positively correlated (rendered by different thresholds) Midgette *et al.* (13) recommend estimation of a summary receiver operating characteristic (SROC) curve (see below).

The estimation of a SROC curve [various approaches by Kardaun and Kardaun (14), Moses *et al.* (15), Littenberg and Moses (16), Shapiro (11), Rutter and Gatsonis (17), Devries *et al.* (18)] using a linear model

$$D = \alpha + \beta S \qquad (21.4)$$

where

$$\begin{aligned} D &= \text{logit(TPR)} - \text{logit(FPR)} \\ &= \log(\text{odds(TPR)}/\text{odds(FPR)}) \\ &= \log(\text{OR}), \qquad (21.5) \\ S &= \text{logit(TPR)} + \text{logit(FPR)}, \\ \alpha &: \text{intercept}, \\ \beta &: \text{regression coefficient of } S. \end{aligned}$$

allows the combination of TPRs, FPRs and ORs with varying corresponding thresholds. The model (21.4) can be analysed in an unweighted, weighted or robust way. If the regression coefficientis near zero ($\beta \approx 0$) the accuracy for each primary study can be summarised by a common OR given by the intercept α. In this special case, other fixed or random effects approaches may be appropriate as well [Hasselblad and Hedges (10), Klassen and Rowe (19)]. Different diagnostic tests may be compared by examination of regression residuals (e.g. with a *t*-test) or introducing 'type of test' as covariate in (21.4). Inclusion of appropriate covariates in (21.4) also permits to determine whether study quality or patient characteristics affect test characteristics or to adjust for them as confounders in a comparison between tests. If data about two or more thresholds are available the use of GEE [Zeger and Liang (20)] may be useful for estimation of the (unweighted) SROC [Irwig *et al.* (9)].

Ordered categorical test results

Suppose, the test result Y can fall into one of J categories ('ratings'). Given k explanatory variables $x_1,\ldots,x_k$ the probability of Y falling in a given category j or below can be modelled as a non-linear function by means of the ordinal regression equation

$$g(\Pr(Y \leq j | x_1,\ldots,x_k)) = \frac{\theta_j - (\alpha_1 x_1 + \ldots + \alpha_k x_k)}{\exp(\beta_1 x_1 + \ldots + \beta_k x_k)} \qquad (21.6)$$

with cut-off values $\theta_1,\ldots,\theta_{J-1}$, location regression parameters , scale regression parameters , and a suitable link function g [Tosteson and Begg (21), Tosteson *et al.* (22), Peng and Hall (23)]. The (smooth) ROC curve of the test (using specific patient covariates) is obtained by plotting sensitivity versus 1–specificity for arbitrary cut-off values θ. The area under the ROC curve may easily be interpreted as the probability of correctly ranking a randomly chosen pair consisting of a diseased and a non-diseased subject. It is the most important summary index of the test's performance [Bamber (24), Hanley and McNeil (25)].

Meta-analysis can either make use of model (21.6) directly to combine data from appropriate studies [Mossman and Somoza (26), Tosteson and Begg (21), Tosteson *et al.* (22), Peng and Hall (23)] or pool the areas under the corresponding ROC curves using fixed or random effects models [McClish (27), Zhou(28)].

Dorfman *et al.* (29) approach the problem of modelling random sampling of readers and patients using the 'jack-knife' method.

Continuous test results

For continuously valued test outcomes – in particular, if they are normally distributed with equal variances – Hasselblad and Hedges (10) advocate the use of the standardised difference of empirical means

$$d = \frac{M_{\bar{D}} - M_D}{s_{\text{pooled}}} \qquad (21.7)$$

where

$M_{\bar{D}}$ sample mean of non-diseased,
M_D sample mean of diseased,
s_{pooled} pooled sample standard deviation

as a measure of discrimination or effectiveness. Alternatively, the area under the ROC curve can be estimated either parametrically or non-parametrically [Bamber (24), Hanley and McNeil (25)]. For either measure, fixed or random effects approaches can be used to combine information from a collection of studies.

Furthermore, the likelihood ratio, i.e. 'the ratio of the probability that a given level of a test result occurs in people with the disease to the probability of that test result in people without the disease' [Sackett *et al.* (30)] can be modelled by means of the linear model

$$\log(\text{LR}) = \log\left(\frac{N_{\bar{D}}}{N_D}\right) + \alpha + \beta x \qquad (21.8)$$

where

LR likelihood ratio,
$\log(N_{\bar{D}}/N_D)$ correction factor to convert log posterior odds to log(LR), i.e. ($\#_{\text{non-diseased in sample}}/\#_{\text{diseased in sample}}$), (where # represents the number of patients)
α intercept in logistic regression model with posterior odds of disease as dependent variable,
β regression coefficient for test measurement in logistic regression model with posterior odds of disease as dependent variable,
x test measurement

[see Albert (31), Irwig (32)]. Additional terms may be added to (21.8) in order to adjust for the effect of covariates. Equal calibration of the studies to be combined is an important assumption. If this is not met, the test data can be categorised and analysed using ordinal regression methods.

In practice, methods for ordered categorical data are often used for categorised continuous test results because costs-benefit arguments rule out collecting data for arbitrarily many threshold values (say more than 100).

Finally, three miscellaneous references regarding meta-analysis of diagnostic test accuracy: Sackett

[(9), Discussion] suggests the feedback of the number of patients one needs to treat in order to prevent one event (NNT) accompanied by summary estimates of diagnostic test accuracy from a high quality systematic review into the diagnostic process. Velanovich (33)[2] advocates the use of meta-analysis for estimating the true Bayesian posterior probability of a diagnostic test. Nierenberg and Feinstein (34) present results of a review concerning the dexamethasone suppression test to establish a five phase evaluation process for diagnostic tests.

Summary

The CMWG on Systematic Review of Screening and Diagnostic Tests [CMWG(1)] remarks that pooling of accuracy assessments within the Cochrane Collaboration will probably use dichotomised (binary) test data because, first, most primary studies present the data in this format and, second, further research on and developments of statistical methods for ordered categorical and continuous test outcomes is needed. Their method of choice is the analysis of the SROC curve in both the unweighted and weighted manner.

References

1. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended methods, updated 6 June. Available at http://www.som.flinders.edu.au/cochrane/, 1996.
2. Swets JA, Pickett RM. Evaluation of diagnostic systems. Methods from signal detection theory. New York: Academic Press, 1982.
3. Hanley JA. Receiver operating characteristic methodology: the state of the art. *CRC Crit Rev I Diagnostic Imaging* 1989;**29**:307–35.
4. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980s. *Stat Med* 1991;**10**:1887–95.
5. Abel U. Die Bewertung diagnostischer Tests. Stuttgart: Hippokrates-Verlag, 1993.
6. Campbell G. General methodology I. Advances in statistical methodology for evaluation of diagnostic and laboratory tests. *Stat Med* 1994;**13**:499–508.
7. Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, *et al.* Guidelines for meta-analyses evaluating diagnostic tests. *Ann Int Med* 1994;**120**:667–76.
8. Irwig L, Tosteson ANA, Gatsonis C. Meta-analyses evaluating diagnostic tests – response. *Ann Int Med* 1994;**121**:817–18.
9. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy (with discussion). *J Clin Epidemiol* 1995;**48**:119–30.
10. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;**117**:167–78.
11. Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol* 1995;**2**:37–47.
12. Ohlsson A. Systematic reviews – theory and practice. *Scand J Clin Lab Invest Suppl* 1994;**54**:25–32.
13. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. *Med Decis Making* 1993;**13**:253–7.
14. Kardaun JWPF, Kardaun OJWF. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Methods Inform Med* 1990;**29**:12–22.
15. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;**12**:1293–316.
16. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993;**13**:313–21.
17. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1996;**2**:48–56.
18. Devries SO, Hunink MGM, Polak JF. Summary receiver operating characteristic curves as a techniques for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Acad Radiol* 1996;**3**:361–9.
19. Klassen TP, Rowe PC. Selecting diagnostic tests to identify febrile infants less than 3 months of age as being at low risk for serious bacterial infection: a scientific overview (see comments). *J Pediatr* 1992;**121**:671–6.

[2] Velaanovich (33) describes a method for combining Bayesian posterior probabilities derived from diagnostic test studies. It should be stressed that the p-values been combined in this instance are essentially a measure of the effectiveness of the test. This method simply substitutes the probability estimate and its variance for treatment effect estimate and its variance in a random effects model, which is covered in chapter 11.

20. Zeger S, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986;**42**:121–30.

21. Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. *Med Decis Making* 1988;**8**:204–15.

22. Tosteson AN, Weinstein MC, Wittenberg J, *et al.* ROC curve regression analysis: the use of ordinal regression models for diagnostic test assessment. *Environ Health Perspect* 1994;**102**:73–8.

23. Peng F, Hall WJ. Bayesian analysis of ROC curves using Markov-chain Monte Carlo methods. *Med Decis Making* 1996;**16**:404–11.

24. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol* 1975;**12**:387–415.

25. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.

26. Mossman D, Somoza E. Maximizing diagnostic information from the dexamethasone suppression test: an approach to criterion selection using receiver operating characteristic analysis. *Arch Gen Psychiatry* 1989;**46**:653–60.

27. McClish DK. Combining and comparing area estimates across studies or strata. *Med Decis Making* 1992;**12**:274–9.

28. Zhou XH. Empirical Bayes combination of estimated areas under ROC curves using estimating equations. *Med Decis Making* 1996;**16**:24–8.

29. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalisation to the population of readers and patients with the jackknife method. *Invest Radiol* 1992;**27**:723–31.

30. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. 2nd edn. Boston: Little Brown, 1991.

31. Albert A. On the use and computation of likelihood ratios in clinical chemistry. *Clin Chem* 1982;**28**:1113–19.

32. Irwig L. Modelling result-specific likelihood ratios. *J Clin Epidemiol* 1992;**45**:1335–8.

33. Velanovich V. Meta-analysis for combining Bayesian probabilities. *Med Hypoth* 1991;**35**:192–5.

34. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test. Lessons from the rise and fall of dexamethasone suppression test. *JAMA* 1988;**259**:1699–702.

Chapter 22

Methods for combining other types of studies/data types

Combining a number of cross-over trials

Patient preference outcome

This is an outcome, special to crossover trials, and a simple methodology has been developed so patient preference outcomes can be combined across trials. In a simple two-period crossover design patients receive both treatments sequentially, and are asked which one they prefer.

P_A and P_B are calculated for each trial (the proportions preferring treatments A and B respectively – ignoring patients who did not express a preference). The difference $P_A - P_B$ is then calculated and is used as a (continuous) outcome measure. The variance of this outcome can be calculated using the below formula:

$$\mathrm{Var}(P_A - P_B) = (P_A + P_B - [P_A - P_B]^2)/n\,, \qquad (22.1)$$

where n = total number of patients in the study analysis. These estimates can then be combined using the standard inverse variance-weighted method (see pages 55–6), or by using any other weighting scheme in the usual manner (1).

Note: If a standard continuous outcome, rather than that of patient preference is used than standard methods for normal RCTs can be used to combine results.

Economic evaluation through meta-analysis

Introduction

Economic evaluation through meta-analysis are sometimes referred to as secondary economic evaluations, evaluations which use available data either alongside a review and meta-analysis of clinical trials as a summary of self standing evaluations (2).

Jefferson *et al.* (3) comment that despite the number of studies available, the process of reviewing and summing up economic evidence has been little developed. They go on to observe that:

> 'The distinction between primary and secondary research in economic evaluation is particularly difficult, given the range of steps and diversity of sources of data required for a typical economic evaluation.' (3)

There are many fewer meta-analytic economic evaluations than meta-analysis of RCTs. The latter appear to have greater scientific acceptance; it has been suggested that this is because RCTs use harder endpoints (3).

Jefferson *et al.* expand on the problems of carrying out a secondary economic evaluation:

> 'It is not clear whether 'systematic reviews' akin to those performed on epidemiological literature can be applied to economic evaluations. First, substantial methodological variations have been demonstrated between economic evaluations that are superficially comparable, raising questions about the reliability of cost-effectiveness 'league tables' and other comparative devices. Second, doubts have been expressed about the theoretical basis for transferring results from the setting in which an existing study was performed into a decision-making process in another setting. Third, it is not clear what technical options exist for summarising and transferring results from a number of economic evaluations.' (3)

Investigation into feasibility

As one may suspect from the introduction, methodology in this area is at an early stage of development.

Jefferson *et al.* (3) present work whose purpose is to explore the possibility of developing a methodology to sum up evidence about cost and cost-effectiveness from pre-existing work. They focus on two issues:

1. Whether, when comparing different studies performed at different times and places, the methods used for converting cost data into a standard currency make much difference to the conclusions of the analysis, in the sense of affecting the relative position of cost estimates from the different studies.
2. Exploring whether studies give enough detail about quantities of resource inputs associated with an intervention to allow these to be

compared directly, and the costs of care in different settings to be calculated from the resource data using a single set of unit prices.

They conclude there is scope for development:

> 'Progress may require a more coherent theoretical framework linked to cost and production function theory .... Inclusion criteria could be tied to increasingly explicit reporting guidelines now being urged for economic studies.' (3)

In addition, they comment:

> 'Systematic review in economic evaluation has not been widely used as a formal process, and it appears that funding agencies such as the UK Medical Research Council and NHS National Research and Development Programme are placing increasing emphasis on the prospective study design, in which economic evaluations are performed alongside and as part of randomised controlled trials. This may indeed be the best way forward ....
>
> ....(But,) such prospective studies may be costly, or be insufficiently powered to show significant cost differences. Questions will continue to be asked concerning the transferability of their results to different geographical settings.' (3)

Illustrative example of problems of combining cost-effectiveness information

Jefferson *et al.* (3) illustrate their work (summarised above) with an economic evaluation of vaccination against influenza A and B.

In 1990, the USA were considering an influenza immunisation programme, but the only available economic evaluation was done in Russia in 1980. A single vaccine cost 6 roubles in 1980 prices. Then, immunising 10,000 people resulted in 1500 fewer physician consultations at 20 roubles per consultation. It was concluded that one cannot directly translate this information, because items (including the vaccine, physician consultation and hospital admission) cost different amounts in the USA in 1990, and hence the ratio of benefits to costs would have altered in the translation.

Thus, two methods for standardising cost were compared:

1. Official exchange rates used to translate costs into US dollars, and the US Consumer Price index was used to place these costs on a 1987 basis.
2. Health specific price indices for each country of study were used to place costs on a 1987 basis, and health specific Purchasing Power Parities were used to translate costs into US dollars.[1]

The study concluded:

> 'All the economic evaluations found in the review reported favourable baseline benefit-cost ratios for vaccination against influenza, and hence at a first broad level of pooling it can be stated that there is general agreement in the available methodologically reliable literature that prevention of influenza A and B by vaccination is worthwhile, especially in high incidence scenarios' (3)

Importantly, however, Jefferson *et al.* went on to comment:

> 'Such a statement is however conditional on the level of clinical effectiveness of the vaccine. Vaccine effectiveness estimates used in the literature range from 30% to 80%, and the cost estimates rest upon these estimates. As no overview of effectiveness of influenza vaccination has been carried out, it must be concluded that the results of the economic literature, regardless of their quality, are only valid at the levels of clinical effectiveness assumed by the studies.' (3)

Further research (in economic evaluation through meta-analysis)

Leidl (4) considers the derivation of cost-effectiveness of a medical intervention at the population level rather than that reported at the patient level. He concludes that eventually the total health gains and the total budget impact expected in the reference population could be reported for a well-defined period. Yet, building on the information provided by single studies, meta-analyses and cross design synthesis can be used to achieve population level results. He goes on to comment:

> 'The methodological development of population based modelling that integrates epidemiology and economics is still in an early stage; a broader discussion of the usefulness of these methods for priority setting among medical interventions is still lacking.' (4)

References

1. Gotzsche PC. Patients' preference in indomethacin trials: an overview. *Lancet* 1989;**1**:88–91.
2. Jefferson T, DeMicheli V, Mugford M. Current issues. In: Elementary economic evaluation in health care. London: BMJ Publishing Group, 1996.

[1] Reasons for doing this are given in the original paper [(3) p. 158].

3. Jefferson T, Mugford M, Gray A, DeMicheli V. An exercise in the feasibility of carrying out secondary economic analysis. *Health Economics* 1996;**5**:155–65.

4. Leidl RM. Some factors to consider when using the results of economic evaluation studies at the population level. *Int J Technol Assess Health Care* 1994;**10**:467–78.

Chapter 23

Methods for correlated outcomes: combining multiple effect measures

Introduction

This section describes the methodology used to combine results, when several outcomes have been reported in each study. When this occurs, one can simply conduct separate analyses for each outcome, or discard all but one outcome (1), and combine the results using standard methods.[1] Raudenbush *et al.* explain why this latter approach may not be desirable:

> 'Because a single treatment may have different effects on different outcomes, it may be misleading to average such effects for each study or simply to choose a single outcome for analysis and ignore the others' (2)

Also, questions such as 'Does a treatment have larger effects on some outcomes than on others? Does the duration of treatment affect different outcomes differently?' are hard to answer (2) using this method. Further, these approaches are not optimally efficient as they do not use statistical information about the errors of estimation contained in the other estimated effect sizes (3).

An alternative approach would be to combine all outcome measures within each study and then combine these across studies. Raudenbush *et al.* points out a problem in doing this:

> 'Each study in a series of related studies typically uses a different set of outcome variables. Thus, a standard multivariate linear-model approach, requiring the same set of outcomes for each unit, cannot be used without either discarding data or imputing missing values.' (2) (Although pages 188–9 present a method that can be used when the same subscales are used in every study.)

Another problem when combining multiple outcomes is:

> 'It cannot, however, be assumed *a priori* that the relations between study features and effect sizes are the same for each outcome. Thus, a method of analysis should allow different predictors (covariates) of effect magnitude for different outcomes.' (2)

None of the above methods is capable of doing this.

Different effect-size estimates calculated for any one sample are typically correlated, so the statistical methods presented thus far, that assume independence between outcomes, may be inappropriate (2).[2]

In this report, two broadly defined situations are discussed where multiple outcomes may occur:

a) Where only two groups of subjects are being compared, but outcomes are reported on several variables.
b) Where several different treatments are being compared to one control (4).

An example of situation a) would be when one had subscales of one outcome and one may want to combine the variables into one outcome measure, e.g. different aspects of psychological well-being (3).

This chapter discusses the various approaches put forward to combine studies with multiple outcomes, retaining as much information as possible. It should be pointed out that many of these were developed by researchers in education, but similar situations do arise in health technology research, an example is given at the end of the chapter, see page 190, and for these instances recent modifications have been presented.

Methods for situation a) are covered on pages 188–91, and situation b) on page 191.

Gleser and Olkin (3) provide a discussion on when one can ignore correlations among estimated effect

[1] This approach is equivalent to standard multiple outcomes analyses applied to single studies.

[2] A related approach that has been used it to extract multiple effect sizes from a given study and then weight each one inversely proportional to the number of effects from that study. Although this allows for the differential effect of a treatment on different outcomes, it does not take into account the interdependence among the effects (2).

sizes, at the cost of being conservative, and use univariate approaches,[3] and when such univariate approaches are not advisable. They conclude that univariate approaches can be used for certain across-study inferences on individual effect sizes, but that multivariate methods are needed for most within-study inferences on effect sizes.

Approaches for dealing with multiple endpoint studies

Combining *p*-values

Strube (5) describes a method for combining significance levels (*p*-values) when the outcomes are non independent. An example from psychology is given, where two trials are considered in which both the patient and the therapist evaluated the treatment. If the four results were combined in the standard way (Stouffers's formula; see chapter 7), this would assume independence, which is clearly disputable with pairs of results coming from the same experiments and this will inflate the Type I error rate. To avoid this, terms for the covariance of the within study results are included in the denominator of Stouffers's formula [see (5) for details of a formula that generalises to *k* studies each with *k* findings].[4–6] From a practical point of view, the true correlation between the outcomes is unknown; however, the correlation between the two dependent variables provides an accurate and conservative estimate. However, this would require the correlation to be given in the report or the access to IPD. If neither of these are available, then it may be possible to estimate it from other results in the paper.[7] If this also is not possible, two other alternatives exist: 1) estimate the correlation using other studies where the value is known, or 2) carry out two analyses with upper and lower bounds for the correlation used.[8,9]

Method for reducing multiple outcomes to a single measure for each study

This method was developed for dealing with the situation where several subscales are to be combined into a single study estimate. Rosenthal and Rubin's (6) aim is to: (a) derive a single summary statistic incorporating the information from all the effect measures of a single study.[10] This statistic could then be combined with, and compared to, the results from other studies using standard meta-analytic procedures. In addition they wish (b), to test specific hypotheses about the relative magnitudes of effects on different covariates, and estimate the magnitude of these contrasts. This methodology is developed for combining either significance levels or measures of effect magnitude.[11] The method combines the df-1 contrasts from a single study and then uses this value in a standard meta-analysis. To use this method, the correlations of the within study results have to be known (or estimated) along with the degrees of freedom for that study.[12]

This method is slightly more complex than that given previously; the underlying formula is given below:

[3] Such procedures are described in the text.

[4] An alternative procedure is also suggested – averaging the *Z* scores within studies before combining. This, however, tends to produce a conservative solution compared to the above.

[5] A modification to the contrasting significance methodology of Rosenthal and Rubin (11), to take into account the correlations is also given (see chapter 7).

[6] An adjustment to Rosenthal's file drawer calculation for significance levels (see pages 127–32), due to correlated outcomes is also given in (5). This calculation assumes similar number of outcomes per study in the unpublished studies to those published.

[7] Examples are given in Strube (5).

[8] Methods for establishing these are discussed in the original paper.

[9] Rosenthal and Rubin (6) comment on the above method. Using estimates for the correlations among the variables to estimate correlations among the *Z* values applies only in situations in which the degrees of freedom are large and the ts are fairly small [more details in (11)]. Below they present an alternative, more accurate, procedure (6) that can be used when these conditions are not met, however, this requires the additional knowledge of the degrees of freedom used.

[10] It should be stressed that this may not be an appropriate procedure in some instances. It makes most sense when outcomes are parallel measures of a single construct (2)

[11] Paper only describes methodology for continuous outcomes and *p*-values.

[12] Hedges and Olkin [(1), p. 210] present an alternative rendering of the above. This includes homogeneity tests to consider whether it is appropriate to combine estimates within a study, and combine estimates across studies taking the correlation into account.

Composite effect size e_c,

$$e_c = \frac{\Sigma\lambda_i t_i / I}{\left[\rho(\Sigma\lambda_i)^2 + (1-\rho)\Sigma\lambda_i^2\right]^{1/2}}, \qquad (23.1)$$

where I is the index of the size of the study, t_i the test of significance of the effect of the independent variable on the ith dependent variable, λ_i the weight we wish to assign to the importance of the ith dependent variable and ρ the typical inter-correlation among the dependent variables.[13]

Many technical details on how to calculate this are given in the original paper (6).

This method requires knowledge or guesses of the correlations between effect measures, and for those correlations to be 'fairly' homogeneous (see page 188).

A more generalised form of the above is presented by Gleser and Olkin (3), using a GLS regression approach (see original for details). The sections below present some more recent developments in the subject of multiple outcome meta-analysis.

Combining vectors of effect size estimates

Hedges and Olkin (1) comment that the above method (pages 188–9) may not be appropriate when there is little reason to believe that the effect sizes on the different constructs are identical. In these situations a method is presented to combine a vector of effect sizes from each study. A limitation is that the method assumes the same outcomes are measured on all studies to be combined. A homogeneity test for vectors of outcomes is also given. This method is not presented in full as it was superseded by the model of Raudenbush *et al.* (2) (see below), who acknowledge Hedges and Olkin's work, and that of Rosenthal and Rubin (6), as forming a foundation for their more sophisticated analysis.

Development of a multivariate model

The model presented by Raudenbush *et al.* (2) is more general than either of the two methods outlined above and it has formed the base on which further work has developed (these extensions are presented in the following sections). Their model uses a GLS regression approach which allows different outcomes (and different numbers of outcomes) to be measured across studies, and also different covariates to be used in regression models to explain the variation in effect sizes for each outcome. Essentially, therefore, this reports an extension of the meta-regression methods of chapter 11.

The model is quite complicated, though a clear and thorough explanation is given in (2), so due to space constraints is not reproduced here.

It should be noted, as was the case for the methods presented earlier, the correlations between outcomes need estimating. Methods previously discussed for doing this apply to this model also (see page 188).[14] Several limitations were noted in the original paper including the danger of model mis-specification, and that random effects and thus a mixed modelling approach (see chapter 12) had not been developed for dealing with multiple outcomes. Several hypotheses can be tested using the model, broadly speaking tests of the model fit, tests of significance of the entire set of covariates, and tests about the individual effects in the model can all be carried out.

An alternative formulation of essentially the same type of GLS regression model is given by Gleser and Olkin (3). In this presentation the authors point out a mistake in the approach of Hedges and Olkin (1) that is carried over into the generalised regression model of Raudenbush *et al.* (2) presented above. They say the large-sample correlation between two estimated effect sizes in a given multiple-endpoint study is equal to the observed correlation of the outcomes. It is now known that this is not the case and for this reason the latter formulation must be recommended.

Extension of the GLS approach

Berkey *et al.* (7) consider multiple-outcome meta-analysis in a clinical trials setting. In the method of Raudenbush *et al.* (2) presented above the outcome scale considered was the standardised treatment difference between the two groups in the study. This method handles data from studies that report

[13] Hence, the test statistic, t_c is given by:

$$t_c = \frac{\Sigma\lambda_i t_i}{\left[\rho(\Sigma\lambda_i)^2 + (1-\rho)\Sigma\lambda_i^2 + (1-\rho^2)\Sigma\lambda_i^2 t_i^2/2df\right]^{1/2}}$$

[14] The paper calls for correlations to routinely be reported in research reports.

different subsets of the outcomes. If a further complication is added, namely that we wish 'to include more than two treatment types in a single meta-analysis, and no single treatment or control group appears in every study to serve as the common group for the computation of effect sizes' (7). In this situation a GLS analysis in terms of effect sizes becomes substantially more complicated. Essentially, the problem being addressed is one where multiple outcomes and multiple treatments are considered simultaneously in a single meta-analysis. It is a model for this problem that Berkey *et al.* consider. In their presentation they acknowledged the model by Dear (8) (see chapter 20) for meta-analysis of survival data, in which the multiple outcomes are survival proportions reported at multiple time points, and use the model of Raudenbush *et al.* (2) as a starting point.

Using this methodology, the authors point out that the model can include more than two treatment groups from multi-arm trials, and can also include a single arm from randomised trials that include only one of the treatments. In doing so, it allows the meta-analysis to use more of the available data. If a common placebo, or before and after treatment data is available for all studies, then within trial comparisons can be directly analysed [using the original units rather than standardised effect size used by Raudenbush *et al.* (2)]. If this is not the case, an analysis can still proceed using simply the outcomes of each arm. Adjustment by study level and treatment-group level covariates when evaluating treatment effectiveness are both possible.

As is the case for all the methods in this section, the correlation between outcomes is required. As previously discussed these can be gained from trial reports, externals sources, IPD from some of the trials in the meta-analysis. Also, assumed values can be used and their impact assessed by a sensitivity analysis.

For illustration purposes the authors present an example from studies of rheumatoid arthritis, where tender joint count, erythrocyte sedimentation rate and grip strength, all reflect important aspects of the disease state. The model is applied to a dataset with a clear explanation of every stage of the analyses.[15,16]

It should be noted that this model would be inappropriate for binary multiple outcomes (e.g. in this example carious versus sound surfaces). It would appear that no methodology has been developed to incorporate binary variables in multiple outcomes meta-analysis. As with the Raudenbush *et al.* model (2), no random effects option is available. The following section describes a situation where this model has been implemented.

Application of extension of GLS approach

Berkey *et al.* (9) apply the model of Berkey *et al.* (7) discussed above to a relatively simple periodontal dataset. Here surgical and non-surgical periodontal treatments were compared, using outcomes of probing depth and attachment level. The primary goal is the estimation of treatment effectiveness for the two outcome measures. 'Analysing both outcome measures simultaneously will improve the accuracy and efficiency of each estimate, so that the standard error of each estimate will be smaller than (or equal to) the standard error from the separate analyses approach.' Since the correlations between outcomes were not reported in the papers, estimates were derived from the individual patient data of one of the studies being combined. All five studies used a split-mouth design, and each quadrant of teeth was randomised to a different treatment. Both outcomes were measured on a continuous scale, and since all studies used the same two outcomes there was no need to standardise these so they were combined on the original metric (see pages 63–4), simplifying the interpretation of the results. A thorough sensitivity analysis of the correlation values was also carried out.

This paper, which includes a clear and detailed appendix for the technical aspects of the analysis, gives a very good insight into the use of these kinds of models. The model fitted is relatively simple, as all five studies measured the same two outcomes and no covariates are included in the model.

Methodology for multiple outcome studies using correlation coefficients as outcome variables

Roth and Sackett (10) discuss the situation where study outcome measures are correlation coefficients (see pages 112–13) and these are

[15] An appendix clearly explains how to set up data for the GLS regression model.

[16] The GLS model of Berkey *et al.* (7) cannot be fit using a standard regression package. However, SAS/IML code is available from the first author.

correlated. This idea can get confusing! Put another way, the context considered is that of studies reporting multiple outcomes, each of which is a correlation coefficient (as opposed to some other continuous measure of effect magnitude). A fictitious example of where this situation may occur is: five student doctors each take a reading on the same group of patients (e.g. blood pressure) and these are compared to a gold standard measure (say the reading of an experienced doctor). Then, if this study were replicated at different hospitals, and one wanted to combine the agreement of each student with the experienced doctor, the correlations between students examining the same patients would have to be accounted for. This paper develops methodology for doing this, which decomposes the correlations to look at differences of a) students within each hospital and b) differences between hospitals (i.e. mean correlations). Extensions to this setting are also discussed (see original paper of mathematical derivation and formula). Monte Carlo simulations were done to test the methods presented.[17]

Approaches for dealing with multiple treatment studies

Gleser and Olkin (3) present a framework for combining studies where patients are assigned to more than two different treatments (situation b of the introduction). The model given can deal with varying numbers of treatments to be combined from each of the studies. As with the approach of Raudenbush *et al.* (2), a GLS model is used (see page 189). A test of the goodness of fit of the model is given which is equivalent to a test of homogeneity (chapter 8).[18] This methodology only deals with outcomes measured on a continuous scale. See original paper for computational details.

Further research

- The need for random effects and mixed modelling approaches to meta-analysis of multiple outcome studies.
- Methodology for the incorporation of binary outcomes into multiple outcome meta-analysis.

Summary

In order to combine multiple effect measures, the correlations/covariances between outcomes are needed for most methods. If these are not available then one must make a guess at them, and assess the impact of their choice using a sensitivity analysis, or alternatively estimate them from external sources, or IPD from some of the trials.

References

1. Hedges LV, Olkin I. Statistical methods for meta-analysis. London: Academic Press, 1985.
2. Raudenbush SW, Becker BJ, Kalaian H. Modeling multivariate effect sizes. *Psychol Bull* 1988;**103**:111–20.
3. Gleser LJ, Olkin I. Cooper H, Hedges LV, editors. Stochastically dependent effect sizes. In: The handbook of research synthesis. New York: Russell Sage Foundation, p. 339–56, 1994.
4. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.
5. Strube MJ. Combining and comparing significance levels from nonindependent hypothesis tests. *Psychol Bull* 1985;**97**:334–41.
6. Rosenthal R, Rubin DB. Meta-analytic procedures for combining studies with multiple effect sizes. *Psychol Bull* 1986;**99**:400–6.
7. Berkey CS, Anderson JJ, Hoaglin DC. Multiple-outcome meta-analysis of clinical trials. *Stat Med* 1996;**15**:537–57.
8. Dear KBG. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* 1994;**50**:989–1002.
9. Berkey CS, Antczak-Bouckoms A, Hoaglin DC, Mosteller F, Pihlstrom BL. Multiple-outcomes meta-analysis of treatments for periodontal disease. *J Dent Res* 1995;**74**:1030–9.
10. Roth L, Sackett PR. Development and Monte-Carlo evaluation of meta-analytic estimators for correlated data. *Psychol Bull* 1991;**110**:318–27.
11. Rosenthal R, Rubin DB. Comparing significance levels of independent studies. *Psychol Bull* 1979;**86**:1165–8.

[17] The methodology of chapter 22 was presented in the social science literature and, to the best of the authors of this reports knowledge, has not been used in a health setting.

[18] A modification presented shows that if the assumption of homogeneity of control and treatment standard deviations within each study holds, greater accuracy can be obtained in the estimation of the effect sizes by replacing the control sample standard deviation with the pooled estimate of the common standard deviation. Another formula is given, for use, when this assumption fails to hold.

Chapter 24

Meta-analysis of individual patient data

Introduction

With one or two notable exceptions, until recently, it was customary for all meta-analyses to be carried out using the aggregated summary results of studies, these were obtained from journal articles and where these were not sufficient, or the study was not published, the necessary summary data was requested directly from the research group who originally carried out the work. With continually improving technology and communication, formal registration of RCTs, and increasing awareness of the benefits, it is becoming more feasible to obtain the whole study datasets from the original researchers, making a synthesised overview using information at the patient level possible. This has become known in the literature as MAP, although other terms such as 'mega-analysis' (1) have also been used.

There are several motivating reasons for carrying out an ambitious analysis of this type; these are discussed below. Firstly, several comparative studies have been carried out to compare the results of a standard meta-analysis using aggregated data to ones which uses IPD, several of these (2–4) have found discrepant results between the methods.[1] Stewart and Parmar (2) compared summary data with IPD for cisplatin-based therapy for ovarian cancer. The two methods did give different results; however, different studies were used in the analysis so the two analyses were not directly comparable. Jeng *et al.* (3) compared literature (MAL) to IPD (MAP) meta-analysis for paternal cell immunisation for recurrent miscarriage, and found that meta-analysis using literature based data over-estimated the treatment effect. They conclude:

> 'Results using the MAL approach can differ from those using the MAP approach because of publication bias, short follow-up, or lack of adjustment for significant confounders' (3)

Other benefits include (5):

1. The ability to carry out detailed data checking and ensure the quality of randomisation and follow-up;
2. Ensure the appropriateness of the analyses;
3. Update follow-up information;
4. Undertake subgroup analyses for important hypotheses about differences in effect. This is possible with aggregate data, but is generally easier with IPD;
5. Survival and other time-to-event analyses can be carried out in a more satisfactory manner. Specific techniques used to analyse data of this type are given in the survival analysis chapter (chapter 20). It has been suggested that IPD analysis is the only satisfactory way to carry out meta-analysis with survival endpoints. It is also worth noting that longer follow-up times than those published in reports may be available if IPD is collected. Stewart and Parmar observed (2) that pressure to publish quickly often results in short follow-up, so meta-analyses of published data tend to focus on early time-points, which may be inappropriate in a chronic disease. For example, in the treatment of breast cancer by chemotherapy, benefit was greatest 5–10 years after treatment.

To obtain data at the patient level it will usually be necessary to contact all groups of investigators who carried out the original trials to be combined. These necessary collaborations may have several 'knock-on' benefits (5):

1. More complete identification of relevant trials;
2. Better compliance with providing missing data;
3. More balanced interpretation of the results;
4. Wider endorsement and dissemination of the results;
5. Better clarification of further research;
6. Collaboration on further research.

The statistical methods for synthesis at the patient level are similar to those used to analyse multi-centre clinical trials. This literature is beyond the scope of this report and the interested reader is referred to Pocock (6). The main feature that differentiates this kind of analysis from that of a single trial is that a covariate can be included to indicate which study a patient came from. This

[1] Little work has been done to explain why this happens or under which circumstances it is likely to occur (10).

covariate could take the form of either a fixed or random effect. The rest of this section outlines the approaches used and discusses the benefits and disadvantages of carrying out MAP.

Procedural methodology

The methods described in this section are a summary of those derived by the Cochrane Working Group (CWG) on meta-analysis using IPD. A fuller account can be found in (5).

Much of the new methodology can be viewed as procedural, such as ways of nurturing collaboration and collecting/checking the (usually large quantities) of data. Many of the procedural methods for synthesis of aggregated data, outlined in chapter 3 are also relevant in this situation. Refer to (5) for details of extra planning/organising that needs doing when planning a synthesis on IPD and note the increased time/costs implied here over use of aggregated results are usually very substantial.

Data collection

The following list has been identified as the minimum data that can be collected to carry out an IPD meta-analysis:

Patient identifier, treatment allocated and outcome(s), together with the date of randomisation and date of outcome if time to event is to be calculated. Also, it is often important to collect additional baseline variables, even when subgroup analysis are not planned, because these data are extremely useful in checking the integrity of the randomisation process.

It is appreciated that collecting old datasets can be a difficult and slow process. The working group concluded:

> 'When a large proportion of the total randomised evidence (perhaps 90–95%) has been collected, the missing data may be considered unlikely to alter importantly the meta-analysis results.' (5)

This can be checked by using a sensitivity analysis, such as examining the effect of including extreme results for the missing data.

Checking data

It is very important that the analysis should be based on the 'intention-to-treat' principle (as is the case for RCTs) and therefore that data should be collected, and analysis based on all randomised patients.[2]

The group suggest simple procedures for checking correct randomisation and follow-up (see the original paper for details).

Discussion of issues involved in carrying out a MAP

Pros

Many of the advantages of a MAP over one using aggregated data were outlined in the introduction to this section. Indeed, Chalmers *et al.* commented:

> 'they are yardsticks against which the quality of other systematic reviews of randomised controlled trials should be measured.' (7)

Some additional advantages have been noted by Gueyffier *et al.* (8): 1) the ability to check whether the treatment effect is constant over time, an assumption which may not be true but which is necessary to calculate an overall estimate when the treatment effect is reported at different time points from one study to another; and 2) the ability to identify interactions between the treatment effect and patient profiles. Three methods for doing this are presented in the appendix of the paper (8).

Cons

They are however, costly, and time consuming to carry out; the CWG noted:

> 'It is perhaps not generally appreciated just how much time and effort is involved in performing an IPD meta-analysis. It is not something to be undertaken lightly, and since a variety of clinical, scientific, statistical, computing and data management skills are required, it is generally not something to be undertaken by an individual.' (5)

A MAP by Pignon *et al.* (9) took approximately 3 years to complete and the CWG concluded that any IPD meta-analysis is unlikely to reach its first publication in much less time (5). As for costs, the CWG suggested around £1000 per trial or £5–10 per patient (in 1994), whichever was less. Although, it has been argued that while the cost of reviews such as these are substantial, they are small relative to the total amount invested in health care and are clearly a smart investment (10).

[2] Indeed, if the original study did not do this and excluded patients from the results, they can now be reintroduced into the analysis (11).

This method relies heavily on the international co-operation between the individuals and groups who have conducted relevant trials (5). To run and report a trial takes much hard work and thus objections to data sharing from trialists are understandable. However, as Oxman *et al.* point out:

> 'patients do not consent to participate in trials for the benefit of researchers or corporate profit. .... it is unethical for trialists, pharmaceutical companies, or others to withhold data for private interests.' (10)

So on ethical grounds, researchers should supply their data when requested for the purposes of a well planned synthesis. Efforts should be taken to make the collecting of data as easy as possible, by measures such as assuring confidentiality, clearly explaining aims of the meta-analysis, making the report manuscript available to all contributors and being flexible to the format of the data requested (11).

There is a problem of how to proceed if all persuasion fails, or the data for the study is simply lost or destroyed. As mentioned in the methodology section this may not be too serious if say 90–95% of the data has been obtained. If on the other hand more than this is missing, aggregate data provided by trialists or data extracted from publications could be used. However, the CWG comment that it is not clear whether it is desirable to do so (5). A suggestion is to investigate the effects of including the aggregated data in a sensitivity analysis (5).[3] See (12) for very recent advances on this subject.

To conclude, many potential advantages of making the extra effort to obtain IPD have been discussed; however, little is known about the actual magnitude of gains that can be achieved.

For this reason, the CWG call for additional empirical evidence of the relative values of the different techniques involved in such reviews should be sought and published (5). A list of explicit suggestions is given below.

Further research on when the extra efforts of doing a meta-analysis using IPD compared with aggregate data are worthwhile.

The CWG comment:

> 'Given that the central collection, checking and analysis of individual patient data from all relevant trials can require a considerable amount of time, personnel and financial resources, further research is needed to determine when it is most appropriate to adopt this approach and what the most appropriate alternatives are if sufficient resources are not available.' (5)

CWG Research agenda (1994):

1. Comparison of individual patient data with summary data supplied by trialists.
2. Comparison of individual patient data with published data.
3. Comparison of individual patient data after extensive data-checking with individual patient data supplied initially.
4. Comparison of trial quality as assessed using the individual patient data with quality as assessed from the published report.

More details of these are given in appendix 6 of (5).

Further research

As well as the issues raised by the CWG research agenda above, the following are relevant:

Methodology to combine individual patient data with aggregate data (i.e. when a proportion of the IPD is unobtainable) is needed. This clearly could be possible using multi-level modelling. See also Collette *et al.* (12) for recent methodology.

Very recently, Higgins and Whitehead (13) proposed a method of including both patient level and study-level covariates in a MAP which uses a Bayesian approach (see chapter 13).

This chapter has dealt with meta-analysis methodology for epidemiological studies. Analysis of IPD may have rewards for observational studies as well as RCTs. For instance, confounding on a patient level could be adjusted for using IPD covariates. For example, consider two observational studies, one which originally adjusted for subject age, and one that did not. If age is available in both studies, using IPD – a meta-analysis could be carried out either adjusting (data from both studies) by age or combining both studies unadjusted.

[3] However no details in the paper were given on how to do this.

Summary

There are several advantages of carrying out a MAP, over a standard meta-analysis using aggregated data. These include the ability to: 1) carry out detailed data checking, 2) ensure the appropriateness of the analyses and 3) update follow-up information. This has led to the comment that MAP data are the yardsticks against which the quality of other systematic reviews of RCTs should be measured (7).

These benefits do not come without a cost however, as MAPs are very time consuming and costly. Currently there is little empirical evidence regarding the actual magnitude of the gains, and it is yet to be established whether the extra effort is worthwhile, in given situations.

References

1. Fortin PR, Lew RA, Liang MH, Wright EA, Beckett LA, Chalmers TC, *et al.* Validation of a meta-analysis: the effects of fish oil in rheumatoid arthritis. *J Clin Epidemiol* 1995;**48**:1379–90.
2. Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet* 1993;**341**:418–22.
3. Jeng GT, Scott JR, Burmeister LF. A comparison of metaanalytic results using literature vs individual patient data – paternal cell immunization for recurrent miscarriage. *JAMA* 1995;**274**:830–6.
4. Pignon JP, Arriagada R. Meta-analysis. *Lancet* 1993;**341**:418–22.
5. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Stat Med* 1995;**14**:2057–79.
6. Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.
7. Chalmers I, Sandercock P, Wennberg J. The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann N Y Acad Sci* 1993;**703**:156–65.
8. Gueyffier F, Boutitie F, Boissel JP, Coope J, Cutler J, Ekbom T, *et al.* INDANA: a meta-analysis on individual patient data in hypertension. Protocol and preliminary results. *Therapie* 1995;**50**:353–62.
9. Pignon JP, Arriagada R, Ihde DC, Johnson DH, Perry MC, Souhami RL, *et al.* A meta-analysis of thoracic radiotherapy for small-cell lung cancer. *N Engl J Med* 1992;**327**:1618–24.
10. Oxman AD, Clarke MJ, Stewart LA. From science to practice. Meta-analyses using individual patient data are needed (editorial 1995; comment). *JAMA* year?;**274**:845–6.
11. Clarke MJ, Stewart LA. Systematic reviews – obtaining data from randomized controlled trials – how much do we need for reliable and informative meta-analyses. *BMJ* 1994;**309**:1007–10.
12. Collette L, Suciu S, Bijnens L, Sylvester R. Including literature data in individual patient data meta-analyses for time-to-event endpoints. *Controlled Clin Trials* 1997;**18**:188S.
13. Higgins JPT, Whitehead A. Inclusion of both patient level and study-level covariates in a meta-analysis. *Controlled Clin Trials* 1997;**18**:84S.

Part G:

Results VI – extensions of meta-analytic methods

Chapter 25

Cumulative meta-analysis

Introduction

Cumulative meta-analysis has been defined as the process of performing a new meta-analysis every time a new trial is published (1).

Lau *et al.* give a broader and more detailed definition:

> '... performing a new meta-analysis every time a new trial is added to a series of trials. The contribution of individual studies to the cumulatively pooled results can be determined. The accumulation may proceed according to the year of completion or publication of each study,[1] the event rate in the control group, the size of each study, the size of the difference between the treatment and the control groups in each study, some quality score that has been assigned to each study, or other covariates such as drug dosage or time-to-treatment. With each criterion, the studies may be sequentially pooled in ascending or descending order. When studies are accumulated by their publication year, the earliest year at which the treatment effect becomes statistically significant can be established.' (2)

It worth noting the distinction between updating a meta-analysis and cumulative meta-analysis: in the latter the results are presented, as each study is added, with the plot produced being an integral part of the analysis, rather than simply re-analysing an updated set of trials. Lau *et al.* comment:

> 'The accumulating results are looked at as a whole for the picture that the trends present and the impact of a published or planned study on the overall result can be assessed.' (2)

However, it would appear that this distinction has become somewhat blurred in the literature.

One of the clearest advantages of this method is expressed by Antman *et al.:*

> 'Cumulative meta-analysis offer the caregiver and the health-care consumer with answers regarding the effectiveness of a certain intervention at the earliest possible date in time' (3)

The updating nature of cumulative meta-analysis makes it naturally Bayesian (2) (see page 198 and chapter 13 for a description of Bayesian methods in meta-analysis).

The earliest example of such an analysis is by Baum *et al.* (4) on use of antibiotic prophylaxis in colon cancer surgery.

A related topic, prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials, is discussed on pages 213–14.

Methodology

Cumulative meta-analysis requires no new statistical techniques to combine study estimates. The trials are ordered by a given criteria (e.g. publication date), then the studies are combined, starting with the first two, and systematically repeating the analysis including the next study in sequence each time, until all the studies available are combined. If the sequential results of this procedure are plotted on the same graph a plot such as (5), *Figure 1,* can be obtained. Here RCTs investigating the effect of intravenous streptokinase (IVSK) for acute myocardial infarction (AMI) have been combined in chronological order. On the left, a conventional meta-analysis is plotted with the point estimate and 95% CI for each trial, and the combined estimate and 95% CI for the combined effect given at the bottom. On the right hand side is the corresponding cumulative meta-analysis, showing the combined result as each trial is included.[2] Producing a diagram like this enables the researcher to assess the impact of each new study on the pooled estimate of the treatment effect. One can see the general trend of a reduction in the width of the CI as the number of studies increases, which is to be expected. The cumulative plot shows that as early as 1971, when only four trials had been completed, the benefit of the treatment reached nominal statistical significance. However, as the next three trial results were included the treatment effect became non-significant again before permanently regaining its significance. As Lau *et al.* point out, this plot is particularly valuable because:

[1] This is the most common usage, by far.

[2] Note the different scales on the horizontal axes of the two plots.

> 'The cumulative method indicates that intravenous streptokinase could have been shown to be lifesaving almost 20 years ago (circa 1972) long before its submission to and approval by the Food and Drug Administration and its general adoption in practice.' (5)

The Bayesian approach to cumulative meta-analysis

The introduction mentioned that the idea of cumulative meta-analysis was amenable to a Bayesian approach. This idea is explored further here. Bayesian analysis quantifies the use of the past history in a prior distribution expressing our belief in the value of the parameter being measured before the information from the current data is incorporated (2). So as each trial is added, the last posterior distribution could become (or provide the basis for) the next prior distribution. The interpretation of results is also different from that in the frequentist approach, as explained by Lau *et al.:*

> 'In the Bayesian paradigm (and therefore in cumulative meta-analysis) the 95% confidence interval describes an interval of highest posterior probability in which we can be 95% certain that the true effect lies. This is also the shortest such interval. This definition contrasts with the common frequentist one which states that in an infinite number of repeated trials, 95% of the 95% confidence intervals will include the true effect. Note that this frequentist definition says nothing about the current confidence interval we are concerned with, but rather makes a general statement referring to intervals that could have been, but where not, observed.' (2)

However, despite this, Lau *et al.* (6) used standard classical statistical methods to pool studies; therefore, although the above philosophy is correct, in practice, their claim that their CIs (which should be credible intervals anyway) have a Bayesian interpretation is not correct.

Cumulative meta-analysis: ordering on variables other than trial publication date

Lau *et al.* (2) discuss the potential benefits of carrying out a cumulative meta-analysis using variables other than trial publication date to order the trials. Each of these, as discussed by Lau *et al.* (2) is discussed below.[3]

Ordering by control group event rate

It can be difficult to establish treatment efficacy in studies or conditions with low control group event rates. Under these conditions random variation are likely to make the treatment worse on occasion. Using the streptokinase studies discussed above, Lau *et al.* (2) show that studies with high control rates are more likely to demonstrate a higher estimate of treatment efficiency (by combining the highest control rates first the treatment effect systematically reduces as more trials are combined). 'Provided that the randomisation process is balanced, a cumulative meta-analysis ordered by ascending or descending control rates provides another approach for exploring the heterogeneity among the studies. Study of control rates may also provide insights into the severity of illness in the study population as well as disease and treatment trends over time.' (2)

Ordering by study size

Smaller studies are generally subject to greater variability, and withholding publication is also more likely to occur. In the streptokinase example, cumulative meta-analysis by study size was compared to analysis by study year. This showed that the result became statistically significant with fewer patients when ordered by study size compared with publication year. This implies that the bigger the studies are, fewer patients need to be randomised.

Ordering by the size of the difference between treatment and control

Lau *et al.:*

> 'Cumulative meta-analysis based on the effect size can be used to highlight the heterogeneity of treatment effect and help to identify studies that may be very different either by protocol design or patient characteristics. The impact of current or potential negative or inconclusive studies on the overall pooled result can be estimated.' (2)

Ordering by the quality of the individual trial
Lau *et al.:*

> 'The overall usefulness of the quantitative estimate of the quality of individual studies has yet to be determined. There have not been enough examples to demonstrate a consistent correlation of quality scores and estimates of treatment effects.' (2)

However, this does not mean that a measure of study quality cannot be used in a sensitivity analysis (see pages 209–10).

Ordering by study covariates

Although subject to ecological fallacy in principle, this may lead to useful clinical insights. Lau *et al.*

[3] For example plots of these methods see paper [Lau *et al.* (2)].

(2) discovered in trials of thrombolytic treatment a cumulative meta-analysis clearly displays a trend toward a reduced estimate of overall efficacy as studies with longer mean time-to-treatment are added.

Discussion

Advantages of cumulative meta-analysis

Lau *et al.* boldly state, and back up by an example, that:

> 'Cumulative meta-analysis can show under what conditions a treatment has been proven effective, ineffective or harmful.', and vaguely postulate, 'Obviously its ultimate usefulness will be the application to interpretation of future randomised control trials of the same or changing treatments.' (2)

They conclude: 'Performing a new meta-analysis every time a new trial becomes available and rearranging the order of the trials in various other ways is in our opinion the best way to utilize the information that can be obtained from clinical trials to build evidence for exemplary medical care.' (2)

Discussion of problems regarding cumulative meta-analysis

Lau *et al.* comment (2) that it has been argued that displaying 'proof', by cumulative meta-analysis, that a treatment was effective before a number of large or 'mega' trial were carried out is a classic example of hindsight reasoning. Lau *et al.* defend this (2), reporting that they alone have carried out over 100 cumulative meta-analyses and found that the large studies mostly echoed the results of meta-analyses of small studies. They add:

> 'We should begin to examine prospectively the role of additional studies when other studies are available.' (2) (see pages 209–10)

Another problem related to cumulative meta-analysis which has been raised is whether a correction factor for multiple testing needs to be taken into account as with stopping rules for clinical trials. Lau *et al.* (2) state that one is required under classical (frequentist) analyses, but argue it is irrelevant under the Bayesian interpretation.

Another aspect of the multiplicity problem that has been highlighted is that, even if there is no treatment difference, a cumulative meta-analysis will eventually lead to statistical significance. Berkey *et al.* (7) suggest the need of a stopping rule for cumulative meta-analyses. At the moment alpha (the probability of a Type I error) approaches 1 (as opposed to 0.05 used in clinical trials). They conclude that no good general method for doing this is possible as each trial has its own unique sample size.

Lau *et al.* discuss the influence of a cumulative meta-analysis on future action:

> 'When there is a clear-cut trend it makes no sense to act as if there was no prior information and to base the sizing of a study on the assumption that each treatment is equally effective. Realistically, does an investigator need as much data to be convinced when previous studies have demonstrated (or indicated) efficacy, as when no studies have been undertaken? The traditional approach taken in textbooks, may have resulted in thousands of patients being relegated to ineffective treatments when far fewer randomised patients would have confirmed that the conclusions to be drawn from the past trials are probably correct.' (2)

However, the authors feel that information from past trials should not be considered when devising stopping rules for current clinical trials because of the danger of compounding of overestimation. They are also cautious of the use of cumulative meta-analysis for future trials sample size calculation for a similar reason.

In a similar vein, Henderson *et al.* (8) discuss how 12 small trials came to completion during the course of one long one. It considers if a cumulative meta-analysis of the smaller ones had been carried out, whether the conduct of the larger one could have been influenced for the better. They conclude:

> 'Our thesis is that if related published trials are available, a meta-analysis should be started in the planning stages of a clinical trial, continued through the ongoing conduct of the trial, and performed as one analysis among many in the final analysis of the trial.' (8)

Finally, Borzak *et al.* (9), commenting on the results of the treatments for myocardial infarction, cumulative meta-analyses (3), state that cumulative meta-analysis is not the 'be all and end all' as it does not assess the combinations of drugs[4] that may be used (implying that factorial trials are needed). Antman *et al.* (10) in their reply acknowledge the point put suggest that the use of a factorial design would be unethical because one would have to have placebo arm when drugs being investigated are known to have a beneficial effect.

[4] But in principle it could.

Further research

Flather *et al.* (11) suggest the need to validate cumulative application of meta-analysis as a research methodology.

The stopping rule problem [despite Berkey *et al.* (7) concluding that there is no good general method for doing this, as each trial has its own unique sample size; some group sequential methods are robust to this].

Summary

Cumulative meta-analysis is valuable as an exploratory/sensitivity analysis tool. There are questionable gains if it is done in real time, and a correction for multiplicity is needed for the frequentist approach.

References

1. Whiting GW, Lau J, Kupelnick B, Chalmers TC. Trends in inflammatory bowel disease therapy: a meta-analytic approach. *Can J Gastroenterol* 1995;**9**:405–11.
2. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials: builds evidence for exemplary medical care. *J Clin Epidemiol* 1995;**48**:45–57.
3. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA* 1992;**268**:240–8.
4. Baum ML, Anish DS, Chalmers TC, Sacks HS, Smith H, Fagerstrom RM. A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls. *N Engl J Med* 1981;**305**:795MCMC9.
5. Lau J, Antman EM, Jimenez-Silva J, Kupelink B, Mosteller SF, Chalmers TC, *et al.* Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* 1992;**327**:248MCMC54.
6. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials. *Stat Med* 1987;**6**:315MCMC25.
7. Berkey CS, Mosteller F, Lau J, Antman EM. Uncertainty of the time of first significance in random effects cumulative metaanalysis. *Controlled Clin Trials* 1996;**17**:357MCMC71.
8. Henderson WG, Moritz T, Goldman S, Copeland J, Sethi G. Use of cumulative meta-analysis in the design, monitoring, and final analysis of a clinical trial: a case study. *Controlled Clin Trials* 1995;**16**:331MCMC41.
9. Borzak S, Rosman H, Antman EM, Lau J, Kupelnick B, Mosteller F, *et al.* Cumulative meta-analyses and the problem of multiple drug effects. *JAMA* 1993;**269**:214.
10. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative metaanalyses and the problem of multiple-drug effects MCMC reply. *JAMA* 1993;**269**:214.
11. Flather MD, Farkouh ME, Yusuf S, Julian D, Braunwald E, editors. Meta-analysis in the evaluation of therapies. In: Management of acute myocardial infarction. London: WB Saunders, 1994, p. 393MCMC406.

Chapter 26

Generalised synthesis of evidence

Introduction – combining different types of study design

So far in this report, the methodology has been concerned with combining a single type of study. Much of the emphasis has been on RCTs; however, chapter 19 dealt with issues exclusive to observational studies. This chapter reviews methodology for combining results from different types of study designs. Methodology for combining matched and unmatched data has already been covered in chapter 19. The combination of cross-over clinical trials (where patients receive more than one treatment regimen) with (one-period) RCTs is discussed first. Next, single arm studies, i.e. studies which use historical controls are examined, and methodology is presented for combining them into a meta-analysis with other RCTs. This is followed by two general methods of combining evidence from different sources (or different groups of studies), namely the confidence-profile method and cross-design synthesis. Sources may include randomised trials, observational studies, animal experiments or database analyses. Hlatky (1) observes that using databases to compare alternative treatments is controversial; additionally, many biostatisticians contend that observational studies are so inherently biased that only data from randomised controlled clinical trials can be used to compare therapies (2–4). However, others have argued that while randomised trials are clearly a superior methodology, it is simply too difficult, costly and time consuming to perform a randomised study to answer every question of clinical interest (5–7). Further, discussion of the pros and cons of combining evidence from sources other than RCTs is provided on pages 203–5. A related topic to this chapter is methods for combining information from disparate toxicological studies; this is covered on page 220.

Combining cross-over trials with other studies

No new methodology is needed to do this, provided the cross-over trials and the other designs use the same outcome measure. In fact, the results of the trials can be combined ignoring the difference in study design. If there is a treatment carry-over effect, then the second period of treatment should be excluded from the analysis. Fortin *et al.* provides a good example of this, they combined 10 trials, eight parallel and two cross-over, and report:

> 'One of the cross-over trials was analyzed as a parallel study, only including data from the first arm of the trial when it was found that the carryover effect exceeded the washout period.' (8)[1]

Single-arm studies (model for incorporating historical controls)

Begg and Pilote (9) present a model to estimate an overall treatment effect when some comparative studies are to be combined with non-comparative, historical control studies. The model differs from the standard random effects model presented thus far (chapter 10) for two reasons: 1) an estimate for each treatment is found (rather than a difference or a ratio); and 2) in this model the treatment effects are fixed but a random effect baseline term is included. The implications of this are that uncontrolled (non-comparative) studies can now be included in the analysis, thus creating the potential to combine extra information over the standard model. It is the authors intention that this model should be used primarily when a dominant proportion of information on a treatment exists from the uncontrolled studies (which may be less reliable than that of the controlled).

Two extensions to this model are proposed: firstly, a preliminary test for systematic bias is given; and secondly, methodology is given to include a random effects term for the treatment effect as well as the baseline effect. In the discussions, the authors stress caution at the interpretation stage when using this method.

Li and Begg extend this (10) by presenting a more general theory removing the need for distributional assumptions and they use EB

[1] Cross-over designs are notoriously designed inappropriately (30) and analysed wrongly. If a trial has been carried out inappropriately then its results may be biased.

estimators (see chapter 13) for the variance terms. The authors state:

> 'The relative contribution of the uncontrolled studies is directly related to the degree of homogeneity in the studies, as evidenced by the closeness of the estimated baseline effects.' (10)

For formulae and computational details see original papers.

The confidence profile method (see also page 103)

Introduction

The confidence profile method was first presented by Eddy (11) in 1989. It has been put forward as a Bayesian method for assessing health technologies (11), and described as:

> 'a set of quantitative techniques for interpreting and displaying the results of individual experiments; exploring the effects of biases that affect the internal validity of experiments; adjusting experiments for factors that affect their comparability or applicability to specific questions (external validity); and combining evidence from multiple sources' (12)

An application of this model has been cited as one of the first examples of bringing together evidence from randomised and non-randomised studies. Eddy *et al.* (13) combined information from RCTs, case–control studies, and simple observational studies to assess the value of mammography screening in women under the age of 50 years. The method was cited as influencing the cross-design synthesis approach (14) (see pages 203–5).

This method can accommodate all the problems, listed below, that may occur when trying to synthesise evidence: multiple pieces of evidence, different experimental designs, different types of outcomes, different measures of effect, biases to internal validity, biases to comparability and external validity, indirect evidence, mixed comparisons, gaps in experimental evidence. (12)

Despite often being described as a Bayesian method, it can be formulated under classical conditions where MLEs and covariances for the parameters in a problem can be derived (12).

Methodology

A key feature of this method is that it models biases explicitly. Hence, the result of an analysis by this method, a posterior distribution for the parameter of interest,[2] incorporates all the uncertainty the assessor chooses to describe about any of the parameters used in the analysis.

Hasselblad notes:

> 'Misclassification rates, measurement error probabilities, and contamination rates are easily included in the model. These biases may not be known precisely, but some evidence usually exists, either in the study itself or in studies of similar populations or outcomes' (15)

As with more standard Bayesian analyses (see chapter 13), it is possible to incorporate subjective judgements into the model in a structured way, though this is not a requirement of the method (16).

There are four fundamental methodological steps for carrying out the confidence profile method, namely:

- define problem precisely
- obtain all available evidence
- define the model parameters and their relationships
- solve model (using one of a selection of given methods).

Later steps can give information requiring the modification of earlier steps, so this process could be considered as being iterative (15).

Hasselblad also claims that the method eliminates the need for sensitivity analysis, saying:

> 'Every parameter, if properly modelled, contains all the information about biases and uncertainty. Thus the final answer includes uncertainty about all of the parameters in the model.' (15)

A piece of software, which is available commercially, called FAST*PRO (17) has been developed to carry out the confidence profile analysis.

This is the briefest of overviews of this method, which is complex. It is well documented, with a whole book describing the methodology (12), and various papers describing its use also (11,16,18).

[2] For more information on posterior distributions see chapter 13.

Cross-design synthesis (see also page 103)

Introduction

The initial work on cross-design synthesis was carried out by the Program Evaluation and Methodology Division of the US General Accounting Office (GAO) (19). Their idea was to create methodology for a new form of meta-analysis that aims at capturing the strengths of multiple-study designs while minimising their weaknesses. Its purpose is to provide better answers to difficult research questions (20), and, 'relates to improving cost-effectiveness of health services, against a backdrop of concern about perceived spiralling health care expenditure.' (21)

Technical issue arises from the difficulties of applying results of RCTs to clinical practice, e.g. patients seldom conform to the characteristics of participants in RCTs. Meta-analysis can make it harder to move from judging whether a treatment is, in principle, efficacious, to deciding how to manage a particular patient, and subgroup analyses of (individual) RCTs lead to *post-hoc* verdicts of questionable reliability (21). However, observational data cannot provide definite answers to questions about therapeutic effectiveness (22).

Droitcour *et al.* (20) state that definitive answers about the effects of various treatments in medical practice can be provided only by a body of research that meets two key criteria: (a) scientific rigor in comparing treatment outcomes and (b) generalisability to the conditions of medical practice. Randomised controlled trials are designed to provide unbiased comparisons of outcomes following treatment, but often fall short of meeting the generalisability criterion. Conversely, statistical analyses of databases are uniquely suited to covering outcomes across the full range of patients, but they rarely provide convincing evidence of unbiased comparison (20). Thus, most RCTs and most database analyses probably fail to meet at least one of the two criteria for providing valid answers to questions about a treatment's effect in medical practice. However, if the strengths of complementary study designs can be combined both criteria can be met; this was the initial motivation for cross-design synthesis. It is interesting to note that Droitcour *et al.* (19) comment that although their work reviews methods for assessing, adjusting, and combining study results, its greatest emphasis is placed on methods for assessing study weaknesses.

Methodology[3]

Droitcour *et al.* (20) acknowledge the previous work of Rubin (23) (see pages 214–15) and Eddy (11,12,16,18) (see page 202), who separately explored ways of synthesising results from studies with a diversity of designs. They say cross-design synthesis builds on these directions but with two key differences: 1) cross-design synthesis focuses on combining results from studies with complementary designs; 2) cross-design synthesis uses a two-pronged approach to study assessment. The first prong consists of an overall quality assessment of each study (see chapter 6). The second prong is a focused assessment of the potential biases that derive from the primary weakness(es) inherent in a study's design. This second prong is the heart of the strategy of cross-design synthesis; its findings are used only to a) adjust the results of an individual study, and b) identify each study's most appropriate contribution to a synthesis model.

As mentioned in the introduction, RCTs and database analyses have complimentary strengths, but one cannot assume that in combining their study results, their strengths will be preserved while their weaknesses counteract each other. For this reason Droitcour *et al.* (20) devised the following three stage strategy for minimising weaknesses of study designs:

- **Focused assessment of the study biases that may derive from characteristic design weaknesses.**

  > 'In-depth and relatively narrow assessments of key biases are conducted in addition to general assessments of quality that are more common in meta-analysis .... The purpose of focused assessments is not to eliminate studies of overall low quality from the synthesis (or otherwise to disregard their results) but rather to provide the information needed to compensate for specific weaknesses.' (20)[4]

- **Individual adjustment of each study's results to 'correct for' identified biases**

  > 'Cross-design synthesis advocates standardising each RCT's results to relevant patient population distributions. For example, patient age-sex distributions from a medical-practice database could be used to correct RCT results for over- or underrepresentation of certain age-sex groups. Similarly, our strategy calls

[3] The methodology for cross-design synthesis is presented here in a conceptual way. For explicit details the reader is recommended the original report (19).

[4] See (19) for explicit details of how to go about these assessments.

> for each treatment effect that is estimated by a database analysis to be adjusted upward or downward, as appropriate' (20)[5]

- **Development of a synthesis framework and an appropriate model for combining results (within and across designs) in light of all assessment information**

Droitcour *et al.* (20) comment that despite secondary adjustments, there is a possibility that the weaknesses of each design may continue to bias study results. This may be because some patient groups may have been totally excluded from randomised studies, which is a problem that cannot be fixed by standardising individual studies' results to correct for over- or under-representation. Similarly, focused assessment of a database analysis may not detect every imbalance in the comparison groups.

The solution put forward to this problem is to devise a framework for organising, analysing, and combining results from different categories of study designs.[6] This framework: a) stratifies the observed effects of treatment on both study design and population coverage, and b) fine-tunes the population coverage strata (so that the results for randomised studies and database analyses can be compared for matched patient groups).[7]

Once this has been done, the investigator must decide whether to: a) present results from each stratum separately; b) present only estimates from certain strata (e.g. strata that contain only those studies deemed to be of high quality); or c) combine estimates across strata using adaptations of the various methods of meta-analysis.

Discussion

Droitcour *et al.* (20) point out three major strengths of cross-design synthesis: 1) it can draw upon different kinds of studies that, in combination, can tell more about how medical treatments work than any single type of study can; 2) it can be applied to existing results in several areas because diverse study designs are increasingly being used to evaluate treatment effectiveness; 3) it has the ability to produce the generalisable information needed to support credible medical practice guidelines.

A limitation, the authors point out, is the necessity of relying on investigator judgement for many decisions. Until refinements of this strategy are developed, GAO believes it is best applied by those knowledgeable about both a specific medical treatment and evaluation methods in general (24).

An anonymous editorial in the *Lancet* was cautious about this new methodology, arguing:

> 'The risk with cross design synthesis is that the more expensive, time-consuming, and reliable component – RCTs – will increasingly be replaced by database analyses.' (21)

Chelimsky *et al.* (24) disagreed with this, commenting that RCTs were a necessary part of cross-design synthesis.

Further developments in cross-design synthesis

The original methodology proposed for cross-design synthesis, by Droitcour *et al.* (19,20) dealt with combining RCTs and database analyses. The spirit of this new methodology was not to limit the inclusion to just these two specific types of study, but to incorporate a broader range of designs. Indeed, Droitcour *et al.* (20) conclude their paper by commenting that studies such as those using a case–control design could provide relevant information.

Abrams and Jones (25) comment that although RCTs may be the 'gold standard' against which the value of evidence from other types of study is judged, in some clinical areas, such as surgery and reproductive medicine, it can be difficult to perform RCTs, for ethical and other reasons. In other areas, such as cancer, the participation rate of patients in RCTs is low and hence the generalisability of the results obtained in them is somewhat limited. Furthermore, changes of resourcing in

[5] This adjustment may be performed using either secondary analysis (and primary adjustment procedures) or the sort of adjustment procedure described by Eddy and colleagues (12). This latter approach suggests that the investigator specify 'a ratio for the outcome parameter that applies to individuals in the treated group compared with individuals in the control group, in the absence of intervention'.

[6] Although this general approach has its roots in meta-analysis, the framework derives directly from the work of Hlatky (1).

[7] See GAO (19) for specific details of methodology, including ways of projecting results to patients not covered in RCTs. Projecting results to patients not covered by RCTs is consistent with the approach to meta-analysis advocated by Rubin (31) (see chapter 28).

the NHS are being cited as making it increasingly difficult to organise and perform RCTs. If the ideal RCT could be conducted, there would be no need for subjective interpretations or for systemic methods of supplementing RCT results (24); however, in these, and other, situations evidence from studies other than RCTs is valuable. The factors motivated the methodology below.

Smith *et al.* (26) have developed a model to include studies with disparate designs into a single synthesis. The authors point out that whilst it may be appropriate to consider randomised studies alone when assessing the efficacy of an intervention, when considering the effectiveness of such an intervention within a more general population evidence from non-randomised studies should be considered as well. They also point out that in certain situations the randomised evidence may be less than adequate due to economic, organisational or ethical considerations. Although this work follows in the spirit of Droitcour *et al.* (19), the methodology used is somewhat different and more specific/operational than that of Droitcour *et al.* A Bayesian hierarchical model approach is taken (see chapter 13). The hierarchical nature of the model specifically allows for the quantitative within and between sources heterogeneity, whilst the Bayesian approach can accommodate *a priori* beliefs regarding qualitative differences between the various sources of evidence.[8] This model can be viewed as an extension of the standard random effects model of chapter 10, but with an extra level of variation to allow for variability in effect sizes between different sources. The method is illustrated in the context of screening for breast cancer, where evidence is available from both RCTs and non-randomised studies. *Figure 6* [reproduced from (25)] outlines the three parameter levels: i) the overall population effect of screening μ, ii) type-of-study parameters θ_i *(i* = 1, 2, 3, where *i* = 1 denotes the effect associated with randomised studies, ... and so on), and iii) φ_{ji} *(i* = 1, 2, 3, *j* = 1, ..., *n*$_i$*)* study-specific parameters, there being n_1 RCTs.[9,10]

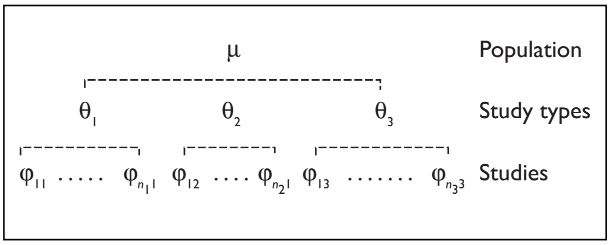


FIGURE 6 *Hierarchical model for breast-cancer-screening synthesis*

The authors point out an issue that remains unresolved, namely whether evidence from randomised and non-randomised studies is to be treated in an equal manner; they conclude this will often depend on the situation under consideration. However, using the Bayesian approach, beliefs about the relative merits of individual studies or types of study can be incorporated in the model. For example, beliefs about the relative value of RCTs, cohort-study and case–control study results may be modelled explicitly and the dependence of the conclusions of the review on these beliefs investigated (25).

Glasziou and Irwig model

In a similar vein to this work, Glasziou and Irwig (27) consider generalising randomised trial results using additional trial information.

They consider the following equation:

Net benefit = (risk level × risk reduction) – harm

This model suggests potential benefit increases with risk, but that harm will remain relatively fixed. Thus at low levels of risk, the benefits will not outweigh the harm and we should refrain from intervening, but at higher levels the benefit will outweigh the harm. Completing the above equation for population subgroups generally requires several sources of data. The authors suggest that the estimate of RRR should come from (a meta-analysis of) randomised trials, the adverse event rates may come from both randomised trials and other epidemiological studies; risk level will usually come from multivariate risk equations derived from large cohort studies.[11]

[8] Prior distributions may represent subjective beliefs elicited from experts, or they might represent other data-based evidence, which though pertinent to the issue in question is not of a form that can be directly incorporated, e.g. data from animal experiments (32).

[9] The paper gives code for the Bayesian analysis package BUGS (33) to carry out this type of analysis, extensions including incorporating prior constraints, prior beliefs, and covariates is also included.

[10] Abrams and Jones (34) point out that parameter estimate for a model of this type can be obtained from several other methods including, GLS, Bayesian methods, classical multilevel models (35), or confidence profiling methods (12). However, GLS and EB methods have been shown to fail to take into account fully all the uncertainty in the model (36,37).

[11] This work, to the authors' best knowledge is currently only available in abstract form, hence the short account.

Application of generalised synthesis of evidence

Tweedie and Mengersen (28) investigate the relationship between lung cancer and passive smoking. Previously, two approaches had been taken for investigating this: 1) the biochemical approach, using cotinine in the main as a marker; and 2) the epidemiological approach. The paper uses both sorts of studies in one meta-analysis. The authors comment on using the now-standard 'Wald adjustment' (29) for differential misclassification; this estimates the effect of differential bias introduced by the misclassification of smokers and non-smokers. The motivation for this is adjustment is 'because smokers tend to marry smokers, if a study contains subjects who are assessed as non-smokers when they are not, they are more likely to be assessed as exposed to ETS: and thus the estimate of relative risk of exposure to ETS will be exaggerated, due to the association of lung cancer with active smoking for this group of 'deceivers'' (28).

Further research

Many of the issues in the further work section of chapter 13 will be relevant here also. In addition:

- Guidelines on when it is appropriate to include studies with designs other than RCTs into a meta-analysis.
- When using multi-level/hierarchical models, establishing when variance estimates of parameters (at the different levels) are (reasonably) reliable. This is particularly relevant when there are a small number of sources of evidence.
- As with Bayesian methods generally, the inclusion of subjective beliefs regarding the credibility of the different sources of evidence require careful elicitation, and further work is required in this area.
- Development of systematic approaches to quality assessments of non-RCTs, and their use in meta-analysis and cross-design synthesis.
- Further experience is required as regards the practical implementation and use of these methods. Sharing of experience in the use of cross-design synthesis approaches for the combination of data from studies with differing designs through a workshop of researchers active in the field, would be desirable.

Summary

Methodology is becoming available for combining studies of different designs. Many of the techniques utilise modern statistical models, including Bayesian methods, and hence many of the methods are extensions of those of chapter 13. Their implementation is, however, facilitated by recent advances in computer software. When such analyses are appropriate, is still open to debate, as there is concern that including studies with poorer designs will weaken the analysis, though this issue is partially addressed by conducting sensitivity analyses under various credibility assumptions.

References

1. Hlatky MA. Using databases to evaluate therapy. *Stat Med* 1991;**10**:647–52.
2. Byar DP. Why data bases should not replace randomized clinical trials. *Biometrics* 1980;**36**:337–42.
3. Mantel N. Cautions on the use of medical databases. *Stat Med* 1983;**2**:355–62.
4. Green SB. Patient heterogeneity and the need for randomized clinical trials. *Controlled Clin Trials* 1982;**3**:189–98.
5. Hlatky MA, Lee KL, Harrel FEJ, Califf RM, Pryor DB, Mark DB, Rosati RA. Tying clinical research to patient care by use of an observational database. *Stat Med* 1984;**3**:375–84.
6. Hlatky MA, Califf RM, Harrel FEJ, Lee KL, Mark DB, Muhlbaier LM, *et al.* Clinical judgment and therapeutic decision making. *J Am Coll Cardiol* 1990;**15**:1–14.
7. Feinstein AR. Current problems and future challanges in randomized clinical trials. *Circulation* 1984;**70**:767–74.
8. Fortin PR, Lew RA, Liang MH, Wright EA, Beckett LA, Chalmers TC, Sperling RI. Validation of a meta-analysis: the effects of fish oil in rheumatoid arthritis. *J Clin Epidemiol* 1995;**48**:1379–90.
9. Begg CB, Pilote L. A model for incorporating historical controls into a meta-analysis. *Biometrics* 1991;**47**:899–906.
10. Li ZH, Begg CB. Random effects models for combining results from controlled and uncontrolled studies in a metaanalysis. *J Am Statist Assoc* 1994;**89**:1523–7.
11. Eddy DM. The confidence profile method: a Bayesian method for assessing health technologies. *Operations Research* 1989;**37**:210–28.
12. Eddy DM, Hasselblad V, Shachter R. Meta-analysis by the confidence profile method. San Diego: Academic Press, 1992.

13. Eddy DM, Hasselblad V, McGivney W, Hendee W. The value of mammography screening in women under the age of 50 years. *JAMA* 1988;**259**:1512–9.

14. John DN, Wright T, Berti C. Is there an economic advantage in the use of SSRIs over TCAs in the treatment of depression? *J Serotonin Res* 1996;**2**:225–35.

15. Hasselblad VIC, Mccrory DC. Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making* 1995;**15**:81–96.

16. Eddy DM, Hasselblad V, Shachter R. A Bayesian method for synthesizing evidence: the confidence profile method. *Int J Technol Assess Health Care* 1990;**6**:31–55.

17. Eddy DM, Hasselblad V. FastPro: Software for MetaAnalysis by the Confidence Profile Method (computer program). San Diego, California: Academic Press. 3.5-inch disk, IBM-PC, 1992.

18. Eddy DM, Hasselblad V, Shachter R. An introduction to a Bayesian method for meta-analysis: the confidence profile method. *Med Decis Making* 1990;**10**:15–23.

19. General Accounting Office. Cross design synthesis: a new strategy for medical effectiveness research. Washington, DC. GAO, 1992.

20. Droitcour J, Silberman G, Chelimsky E. Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *Int J Technol Assess Health Care* 1993;**9**:440–9.

21. Editorial. Cross design synthesis: a new strategy for studying medical outcomes? *Lancet* 1992;**340**:944–6.

22. Kassirer JP. Clinical trials and meta-analysis. What do they do for us? *N Engl J Med* 1992;**327**:273–4.

23. Rubin D, Wachter KW, Straf ML, editors. A new perspective. In: The future of meta-analysis. New York: Russell Sage Foundation, 1992, p. 155–65.

24. Chelimsky E, Silberman G, Droitcour J. Cross design synthesis (letter 1993; comment). *Lancet* 1993;**341**:498.

25. Abrams KR, Jones DR. Meta-analysis and the synthesis of evidence. *IMA J Math Appl Med Biol* 1995;**12**:297–313.

26. Smith TC, Abrams KR, Jones DR. Using hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. Department of Epidemiology and Public Health Technical Report. University of Leicester, 1995.

27. Glasziou P, Irwig L. Generalizing randomized trial results using additional epidemiologic information. *Am J Epidemiol* 1995;**141**:S47.

28. Tweedie RL, Mengersen KL. Lung cancer and passive smoking: reconciling the biochemical and epidemiological approaches. *Br J Cancer* 1992;**66**:700–5.

29. Wald NJ, Nanchahal K, Thompson SG, Cuckle HS. Does breathing other people's tobacco smoke cause lung cancer? *BMJ* 1986;**293**:1217–22.

30. Khan KS, Daya S, Collins JA, Walter SD. Empirical-evidence of bias in infertility research – overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertil Steril* 1996;**65**:939–45.

31. Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography: a meta-analysis. *JAMA* 1995;**273**:149–54.

32. Abrams KR, Hellmich M, Jones DR. Bayesian approach to health care evidence. Department of Epidemiology and Public Health Technical Report 97-01, University of Leicester, 1997.

33. Thomas A, Spiegelhalter DJ, Gilks WR, Bernardo J, Berger J, Dawid A, Smith A, editors. BUGS: A program to perform Bayesian inference using Gibbs sampling. In: Bayesian statistics 4. Oxford: Oxford University Press, 1992.

34. Schatzkin A, Longnecker MP. Alcohol and breast cancer. where are we now and where do we go from here? (review). *Cancer* 1994;**74**:1101–10.

35. Goldstein H. Multilevel models in educational and social research. London: Griffin, 1987.

36. Lambert PC, Abrams KR. Meta-analysis using multilevel models. *Multilevel Modelling Newsletter* 1996;**7**:17–19.

37. Carlin JB. Meta-analysis for 2 x 2 tables: a Bayesian approach. *Stat Med* 1992;**11**:141–58.

Chapter 27

Special issues and problems in meta-analysis

Subgroup analysis

Two different types of subgroup analysis are possible in a meta-analysis. One can investigate subsets of studies. Studies being pooled may differ with respect to treatments applied, control groups, patient eligibility, quality control, study conduct, and follow-up maturity (1). See pages 45–60 for further discussion of this procedure.

Alternatively, one can consider subsets of patients within the studies being pooled. Put formally, this type of subgroup analysis can be defined as 'the investigation of the influence of factors other than treatment factors on the response variables or treatment effects in clinical trials' (2).

This procedure is commonly carried out on single RCTs, but is also sometimes possible in a meta-analysis.

Yusuf *et al.* comment:

> 'When reasonably uniform data are pooled from many studies, the statistical power to detect a subgroup effect may be high enough to establish the likely existence of differential subgroup effects when the individual trials did not .... Conversely, the pooled data from many trials may refute the claim of a subgroup effect from a single trial.' (3)

When conducting subgroup analyses within a meta-analysis, Yusuf *et al.* (3) comment that clear definitions of the subgroups are essential. For example, subgroups defined by a specific ejection fraction for all trials, say 35%, are preferable to vague categorisations of patients, say 'high risk,' the definition of which may vary greatly between trials.

Gelber and Goldhirsch (1) discuss subset analysis and observe that meta-analysis leads to larger subsets than analysis of individual trials. They do comment however that one does need to be wary of misclassification, dilution and bias.

Counsell *et al.* (4) carried out an experiment to determine whether inappropriate subgroup analysis together with chance could change the conclusion of a systematic review of several randomised trials of an ineffective treatment. Trials were simulated by throwing fair dice and recording outcomes which were either death or survival. Publication bias was also simulated. The results showed that analysis of subsets to be misleading, with chance influencing the outcomes of clinical trials and systematic reviews of trials much more than many investigators realise.

Oxman *et al.* (5) state that the extent to which a clinician should believe and act on the results of subgroup analyses of data from randomised trials or meta-analyses is controversial. Their paper provides guidelines for making these decisions – these are reproduced below:

- Is the magnitude of the difference clinically important?
- Was the difference statistically significant?
- Did the hypothesis precede rather than follow the analysis?
- Was the subgroup analysis one of a small number of hypotheses tested?
- Was the difference suggested by comparisons within rather than between studies?
- Was the difference consistent across studies?
- Is there indirect evidence that supports the hypothesised difference?

Sensitivity analysis

As the Cochrane Handbook states, sensitivity analysis:

> 'provides reviewers with an approach to testing how robust the results of the review are, relative to key decisions and assumptions that were made in the process of conducting a review. Each reviewer must identify the key decisions and assumptions that are open to question, and might conceivably have affected the results, for a particular review.' [(6), p. 83]

It has been argued that sensitivity analysis should in fact permeate all stages of a meta-analysis (7). Indeed, many sections of this report have commented on the use of methods as a form of sensitivity analysis.

The Handbook also provides a list of factors one may want to investigate. These are:

- changing the inclusion criteria (including the types of participants, interventions and outcome measures, and methodological cut-points)
- including or excluding studies where there is some ambiguity as to whether they meet the inclusion criteria
- excluding unpublished studies
- excluding studies of lower methodological quality. Additionally, Blair *et al.* (7) suggest using quality scores (see chapter 6) to determine the effect of confounders on outcomes (in epidemiological studies)
- reanalysing the data using a reasonable range results (say the upper and lower limits) for studies where there may be some uncertainty about the results (e.g. because of inconsistencies in how the results are reported that cannot be resolved by contacting the investigators or because of differences in how outcomes are defined or measured)
- reanalysing the data imputing a reasonable range of values for missing data
- reanalysing the data using different statistical approaches (e.g. using both fixed and random effects models).

Additionally, simulations of extra trials can be carried out to assess the robustness of the results. These may be particularly useful if one knows of trials currently underway, and one could assess how a range of likely outcomes of these trials would effect the conclusions of a meta-analysis.

Graphical displays used in sensitivity analysis

Thompson (8) considers the impact of the choice of statistical methods in more detail. He explains that a random effects analysis can be viewed as simply changing the percentage of weight allocated to each trial, as compared to the fixed effect analysis. He takes the sensitivity analysis further than just doing fixed and random effects analyses to determining the pooled OR as a function of the between study variance. Hence, a value of zero for the between study variance corresponds to a fixed effects analysis; in a particular example 0.023 (see original paper), a random effects analysis; and infinity would give the studies equal weighting. Thompson constructs a graph of the pooled OR over this range of values for the between study variance, to see how sensitive it is (see original paper for graph).

The reliability of meta-analysis

Cappelleri *et al.* (9) compare the results of meta-analyses of small trials with those of large trials.[1] The motivation for doing this came after the disagreement between existing trial results and the findings of a mega-trial for the treatment of AMI with intravenous magnesium. The authors pose the questions: How well do large and smaller studies agree in their results? How frequent are the significant disagreements? Why do these disagreements occur? Are the disagreements clinically important?

Chalmers *et al.* (10) and Villar *et al.* (11) also investigate how well the results of a meta-analysis of smaller randomised trials predicted the results of a large 'gold standard' trial. Villar *et al.* (11) looked at 30 meta-analyses in perinatal medicine, covering 185 RCTs and compared the results of the meta-analyses, with largest study removed, to the results of the largest study alone. They found 24 correctly predicted the direction of the treatment effect, but only 18 of the 30 both showed an effect in the same in direction of treatment effect as the largest trial and were statistically significant.

Cappelleri *et al.* conclude that the results of meta-analyses of smaller trials are usually compatible with the results of larger trials:

> 'clear-cut discrepancies do occur and their frequency is more substantial when the results are analysed without considering the variability of treatment effect among different smaller trials (i.e. with a fixed effects model[2]) .... Potential explanations for most of the genuine disagreements may be identified in control rate differences, specific protocol or study differences, and publication bias, as well as methodological factors such as the quality of primary studies. Clinically important disagreements without identifiable explanations are uncommon.' (9)

They do warn, however, that their investigation is retrospective and not designed to decide whether a meta-analysis of smaller trials is sufficient or a mega-trial is warranted in general.

Effect of early stopping rules in clinical trials on meta-analysis

Hughes *et al.* (12) investigate the effect early stopping rules for clinical trials have on the

[1] A large trial was defined (and analysed) in two ways by its size and by its power.

[2] When a random effects model was used the disagreement between large and smaller studies was halved (i.e. by taking the heterogeneity into consideration).

heterogeneity and estimation in a meta-analysis. They conclude that for any overview of reasonable size, the estimate of effect is approximately unbiased. This is because it is the trials with extreme results which stop early, and hence have less precise estimates of effect. Therefore, they carry much less weight in the analysis than those with more representative results because these continue longer and achieve greater precision.

They do, however, conclude:

> 'When undertaking or interpreting overviews, one should ascertain whether stopping rules have been used (either formally or informally) and should consider whether their use might account for any heterogeneity found.' (12)

Green *et al.* (13) also investigate the effect of early stopping rules on overviews of clinical trials. As well as investing three different stopping rule methods, they also considered the effect of the inappropriate use of early stopping rules, specifically when no significance level adjustments have been made, to take into account multiple testing. The authors report that the bias induced by this latter mechanism would inflate the effect of the new treatment in overview results, in a similar way to that induced by publication bias (see chapter 16). They conclude that combining results greatly diminishes the effect of inappropriate multiple testing on level; additional follow-up dilutes this effect even more, so that this error should be of limited concern.

It is interesting to note that in the discussion of this paper (13), Peto comments that when a therapy is new and the early results of trials are combined, the effect will be greater. He suggests that perhaps, overviews very early on of enthusiastic trials, are not appropriate.

Using multiple comparisons to detect non-exchangeability across studies

The National Research Council report [(14), p. 149] discusses the use of multiple comparisons in meta-analysis. The authors comment that the *Q* test (pages 39–40) is an omnibus test and hence does not point out which studies/outcomes contribute to the heterogeneity. Multiple comparisons can be useful; explicitly they 'can be exploited as a screening device to find studies that cluster together, thereby helping the analyst to decide which studies, if any, may profitably be pooled'. (14)

The report outlines classical, empirical, and full Bayes approaches to multiple comparisons. To the authors' knowledge, these procedures have never been used in health services research.

Further research

The National Research Council report (14) highlights several research opportunities on the topic of multiple comparisons. Concerning Bayesian methods, they state further work on the elicitation of prior distributions in the context of multiple means is needed. (For further information on prior elicitation see chapter 13.) For the classical approach, three sets of research opportunities have been identified: 1) graphical representation of all-pairwise comparisons, 2) multiple comparisons in the general linear model, 3) multiple comparisons when the variances are unequal. [Further details of all these are given in (14), p. 156.]

The reason one carries out a meta-analysis is to inform current clinical practice. How to use the results of a meta-analysis to treat individual patients is an important topic (15). Increased dissemination of how this is done through workshops and other means would be beneficial.

References

1. Gelber RD, Goldhirsch A. Interpretation of results from subset analyses within overviews of randomized clinical trials. *Stat Med* 1987;**6**:371–8.
2. Schneider B. Analysis of clinical trial outcomes: alternative approaches to subgroup analysis. *Controlled Clin Trials* 1989;**10**:176S–86S.
3. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;**266**:93–8.
4. Counsell CE, Clarke MJ, Slattery J, Sandercock PAG. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *BMJ* 1994;**309**:1677–81.
5. Oxman AD, Guyatt GH. A consumers guide to subgroup analyses. *Ann Int Med* 1992;**116**:78–84.
6. Oxman AD, editor. The Cochrane Collaboration handbook: preparing and maintaining systematic reviews. 2nd edn. Oxford: Cochrane Collaboration, 1996.
7. Olkin I. Re: 'A critical look at some popular meta-analytic methods' (comment). *Am J Epidemiol* 1994;**140**:297–9.

8. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction (review). *Stat Methods Med Res* 1993;**2**:173–92.

9. Cappelleri JC, Ioannidis JPA, Schmid CH, Deferranti SD, Aubert M, Chalmers TC, *et al.* Large trials vs metaanalysis of smaller trials – how do their results compare. *JAMA* 1996;**276**:1332–8.

10. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I: control of bias and comparison with large co-operative trials. *Stat Med* 1987;**6**:315–25.

11. Villar J, Carroli G, Belizan JM. Predictive ability of meta-analyses of randomised controlled trials. *Lancet* 1995;**345**:772–6.

12. Hughes MD, Freedman LS, Pocock SJ. The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* 1992;**48**:41–53.

13. Green SJ, Fleming TR, Emerson S. Effects on overviews of early stopping rules for clinical-trials. *Stat Med* 1987;**6**:361.

14. National Research Council. Combining information: statistical issues and opportunities for research. Washington DC: National Academy Press, 1992.

15. Mant J, Hicks N, Rosenberg W, Sackett D. How to use overviews of prevention trials to treat individual patients. *Cerebrovasc Dis* 1996;**6**:34–9.

Chapter 28

New developments and extensions in meta-analysis

Prospective meta-analysis

Margitic *et al.* (1) discuss a new approach to meta-analysis which they call prospective meta-analysis. They comment on potential disadvantages with standard (retrospective) meta-analysis, namely: possibility of biases (selection, trial-related, publication), trial heterogeneity (non-concurrent, different protocols and data, varying outcomes, and data quality), and incomplete access to the databases of the individual trials comprising a particular meta-analysis. Margitic *et al.* state:

> 'By combining elements of a multicenter clinical trial with specific features of a retrospective meta-analysis, the prospective meta-analysis reduces a number of the disadvantages inherent in a retrospective study. It is a compilation of common data collected prospectively from trials testing related interventions.' (1)

Prospective meta-analysis gives the following benefits:

- Selection and publication biases are minimised.
- A complete, pooled database is obtained with more consistent quality data across sites.
- Allows pre-trial statement of objectives, definition of the population of interest, and *a priori* hypothesis testing.

The paper suggests procedural methodology that may be useful when planning and conducting a prospective meta-analysis. Some of the key points were summarised by Probstfield and Applegate (2); these are set out below. Much fuller details are given in the original paper (1).

- Establish collaboration while the trials are being designed.
- Develop a detailed protocol and standardises staff training.
- Standardise collection of some key common data.
- Establish external study oversight by an unbiased group.
- Establish clear publication procedures and priorities.

Friedenreich (3) discusses prospective meta-analysis designs for epidemiological studies, and states that by planning data pooling during the design phase, combined analyses are easier to conduct, since the studies being combined will have similar designs and standardised methods. This approach has already been used by the International Agency for Research on Cancer for a number of cohort and case–control studies.

A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials

Whitehead (4) combines the methodology of sequential designs (specifically the triangular test), and that for combining studies. This methodology was developed in the context of a series of studies, following broadly similar protocols, each comparing the same form of new treatment with a control treatment.

> 'For example, in the evaluation of a drug for the prevention of a serious side-effect resulting from chemotherapy given to cancer patients, different studies may deal with cancers at different sites. However, the primary efficacy variable, which is the occurrence or not of the specific side-effect, is the same in all studies[1] .... In such cases, individual fixed sample size studies may be designed for the primary efficacy variable, but the safety variable would be analysed according to a sequential design with stopping boundaries. Significant evidence demonstrating that the new treatment was harmful could then lead to the stopping of all the studies.' (4)

The author goes on to comment that one may wish to treat the design as a multicentre trial, or stratified study, rather than planning each trial with enough power to detect, say, side-effects individually. One would want to know as quickly as possible if there was a difference (i.e. there was a side-effect), so a sequential design would be suitable. If a fixed effects model is used in this situation, no new methodology is necessary, however, new methodology is needed if random effects are included in the analysis.

[1] Two other practical situations where this methodology could be used are given in the paper (4).

The paper focuses on the sequential design and analysis used for the cumulative meta-analysis. See original paper (4) for the methodology details.

Meta-analysis using surrogate markers

Very little literature exists on this. However, the little that does exist contributes substantially to the meta-analysis methodology applicable to HTA, and thus is outlined here.

Definition of a surrogate endpoint

A surrogate marker can be used in the place of the real outcome, when this is difficult to measure (5). An example would be use of CD4 count as a marker for time until the onset of AIDS in a human immunodeficiency virus (HIV) trial (6), allowing the evaluation of an intervention to be done in a much shorter space of time.

A'Hern *et al.* (7) investigated the relationship between response rate and median survival in advanced breast cancer. The authors used the results of RCTs and investigated the association between the OR that summarised the difference in response rates in pairs of arms within the same study and the corresponding ratio of median survivals.

Several limitations of the method they used for doing this have been highlighted. Daniels and Hughes (6) comment that although the precision in estimating the treatment difference on the clinical outcome is allowed for, the method did not take into account the level of precision in estimating the treatment difference on the response variable. Torri *et al.* (8) also made the same criticism of the method.

This problem is addressed independently in the two approaches below, by Torri *et al.* (8) and Daniels and Hughes (6).

Model of Torri *et al.*

Torri *et al.* (8) present a method to determine the relationship of anti-tumour response and survival in patients with advanced ovarian cancer treated with chemotherapy. Firstly, a correlation coefficient (Kendall's tau) was calculated for this relationship for each study using values for median response. These were then combined using standard methods (see pages 112–13 combining correlation coefficients; this is not Person's, however). A secondary analysis was also performed to quantify the relationship between the magnitude of a treatment effect on median survival and its effect on response rate. The model incorporates between and within study sources of variability in the estimates of response and survival. See paper for details of the new model used.

Evaluating potential surrogate markers using meta-analysis

The aim of Daniels and Hughes work (6) is slightly different. They present a method to model the association between the treatment difference on a potential surrogate marker and the treatment difference on the clinical outcome of interest. Then, this model is used to determine the surrogate's reliability for predicting the treatment difference on clinical outcome given an observed difference on the surrogate marker. The approach taken to modelling is a Bayesian one, and this model is a development of that of DuMouchel (9), used for more traditional meta-analysis applications (see chapter 13).[2]

This model is applied to the example given in the definition section for HIV, using CD4 count as a surrogate for the time to AIDS onset. It is very interesting to note that the 15 trials included used different interventions. This is permissible, since treatment effect is not of interest here, it is the relationship between CD4 and AIDS onset which is under investigation. Several of the trials had three and four arms which could all be incorporated in the model.

Estimating and extrapolating a response surface – a new approach to meta-analysis

In 1990 Rubin presented a new perspective for meta-analysis that was deliberately provocative (10). The crux of this method is that it aims to estimate the effect of an ideal study, rather than to calculate average effects over the studies being combined. He questioned the aims of a meta-analysis, stating there are two kinds: those for literature synthesis and those for understanding the underlying science. He believed that at the time the current view was to carry out literature synthesis, with the aim being of summarising all existing studies by their average population effect.

[2] For the mathematical description of the model see (6).

Presented in the chapter is a contrasting view which is called 'building and extrapolating a response surface'. Details of the model are not given, just a conceptual framework which is summarised below:

Let *Y* denote the outcome variable(s). Let there be two treatments: experimental (E) versus control (C). The variable to be estimated is the effect of E versus C on *Y*. Let (*X*, *Z*) define factors describing studies that might be used to estimate the effect of E versus C on *Y*.

Two different factors are defined: *X*s consists of scientific factors (e.g. the gender of subjects, the age of subjects etc.); *Z*s are scientifically uninteresting design variables (e.g. the sample sizes of the studies and the type of controls being used).[3]

Thus, the response surface of interest is the effect of treatment E versus C on *Y* as a function of the two kinds of factors, *X* and *Z*. This is a whole surface that expresses the typical treatment effect as a function of scientific factors and design factors such as we conceptualise them. From this model, the effect of E versus C on *Y* needs estimating when *Z* is Z_0, where Z_0 indicates values of the design variables for a perfect study. Hence, the ideal answer is a whole function of the scientific factors, with design factors fixed at perfect studies.

Rubin emphasises the distinction between this approach and that is normally taken (10), namely that in his new perspective answers are conditional on those factors that describe the science, *X*, and an ideal study $Z = Z_0$:

> 'In contrast, the current view .... pursues averages of this response surface, average answers over the values of the scientific factors the designs that current investigators have, for some reason, chose to use.' (10)[4]

Recently, Vanhonacker (11) presented an implementation of Rubin's conceptual idea and derived a least squares approach to meta-analysis response surface extrapolation. The paper, in fact, has three goals: 1) to bring conceptual clarity to what is being estimated in meta-analysis models; 2) to refocus attention on what is of scientific interest and how meta-analysis can help in our understanding of a modelled phenomenon; 3) to provide an unbiased response-surface-extrapolation estimator of the effects of intrinsic scientific value.

The model derived (see paper for details) is applied to an empirical illustration investigating the question how advertising affects product sales using 128 primary studies.

Combining meta-analysis and decision analysis

The vast majority of the time a meta-analysis is performed and the results reported. Others are left to assess its implications for effecting changes in medical practice. However, the study below incorporates the treatment effect estimates, provided by a meta-analysis, into a decision model assessing when the treatment should be used.

Midgette *et al.* (12) present a combined meta-analysis and decision analysis of the effects of infarct location, and of likelihood of infarction. They assess the effectiveness of IVSK on short-term survival after suspected AMI. Since the diagnosis of AMI is sometimes uncertain, this needs taking into account in the model. Using a meta-analysis of the effects of IVSK on short-term mortality in patients with different locations of infarction a simple decision tree was developed to compare IVSK with conservative treatment for AMI. Short-term mortality, costs, and marginal cost-effectiveness ratios as a ratio of additional dollars per additional live saved were all predicted.[5]

Meta-analysis of single case research

A substantial number of papers have been written on the subject of meta-analysis of studies with single case designs (13–18) (and this list, we suspect, is not exhaustive). Single case designs are often used in psychotherapy, but their use in assessing health technologies is somewhat rarer. There is considerable debate in the literature over appropriate measures of effect size. As well as effect differences, percentage non-overlapping data and other more complex measures have been used as outcome

[3] Rubin notes that the dichotomy between *X* and *Z* is vague. There are certain factors that sometimes would be design factors and other times would be more usefully thought of as factors of scientific interest.

[4] The chapter gives detailed notes on the advantages of this method over the normal approach.

[5] See paper for details of methodology, results and discussion.

variables. For a good review and introduction to the area see Allison and Gorman (13).

Best evidence synthesis: an alternative to meta-analysis

Slavin has proposed an alternative method to meta-analysis for synthesising results in a particular field (19,20). He calls this method best evidence synthesis, and claims it combines the strengths of meta-analytic and traditional reviews.

Put simply, this method does not combine all the studies carried out, but considers only the best of them and only combines those. Criteria for 'best' would be defined *a priori*. This is linked with a more formal discussion of the studies/results. Slavin states:

> 'In a meta-analysis, the presentation of the 'results' is essentially the end point of the review. In a best-evidence synthesis, the table of study characteristics and effects sizes and the results of any pooling are simply a point of departure for an intelligent, critical examination of the literature.'
>
> 'Best-evidence synthesis incorporates the quantification and systematic literature search methods of meta-analysis with the detailed analysis of critical issues and study characteristics of the best traditional reviews in an attempt to provide a thorough and unbiased means of synthesising research and providing clear and useful conclusions.' (20)

He criticises meta-analysis for the potential for serious errors, and states that in a meta-analysis the reader has no way of forming his or her own opinion as it is rare that they describe even one study in any detail. He also comments that biases in the primary research are too often reflected in the meta-analysis. In addition, Slavin states:

> 'Meta-analysis was developed to replace the artistic narrative review with a scientific and systematic method. Yet in fear of allowing bias to creep in, meta-analysis is typically mechanistic, driven more by concerns about reliability and replicability than about adding understanding of phenomena of interest.' (20)

As a result of these misgivings, he proposed 'best-evidence synthesis[6]' as an alternative to meta-analysis (19), to incorporate many of the important contributions of meta-analysis, but also to retain many of the features of intelligent and insightful narrative reviews.

Letzel (21) discusses in detail the uses and differences between best-evidence synthesis and meta-analysis.

Further research

Meinert (22) calls for the need for methodology to control the timing of when a meta-analysis, in a given field, is performed.

References

1. Margitic SE, Morgan TM, Sager MA, Furberg CD. Lessons learned from a prospective meta-analysis (see comments). *J Am Geriatr Soc* 1995;**43**:435–9.
2. Probstfield J, Applegate WB. Prospective meta-analysis: ahoy: a clinical trial? (editorial 1995; comment). *J Am Geriatr Soc* year?;**43**:452–3.
3. Friedenreich CM. Methods for pooled analyses of epidemiologic studies (review). *Epidemiology* 1993;**4**:295–302.
4. Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Stat Med* 1997;**16**:2901–13.
5. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med* 1989;**8**:415–25.
6. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med* 1997;**16**:1965–1982.
7. A'Hern RP, Ebbs SR, Baum MB. Does chemotherapy improve survival in advanced breast cancer? A statistical overview. *Br J Cancer* 1988;**57**:615–18.
8. Tori V, Simon R, Russek-Cohen E, Midthune D, Friedman M. Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. *J Natl Cancer Inst* 1992;**84**:407–14.
9. DuMouchel W. Hierarchical Bayes linear models for meta-analysis. Research Triangle Park, NC: National Institute of Statistical Sciences, 27, 1994.
10. Rubin D, Wachter KW, Straf ML, editors. A new perspective. In: The future of meta-analysis. New York: Russell Sage Foundation, 1992, p. 155–65.
11. Vanhonacker WR. Meta-analysis and response surface extrapolation: a least squares approach. *Am Statist* 1996;**50**:294–9.

[6] Author comments that it could just as well have been seen as a set of standards for meta-analysis designed to describe the full richness and unique contributions made by the most important studies.

12. Midgette AS, Wong JB, Beshansky JR, Porath A, Fleming C, Pauker SG. Cost-effectiveness of streptokinase for acute myocardial-infarction – a combined metaanalysis and decision-analysis of the effects of infarct location and of likelihood of infarction. *Med Decis Making* 1994;**14**:108–17.

13. Allison DB, Gorman BS. Calculating effect sizes for meta-analysis: the case of the single case (review). *Behav Res Ther* 1993;**31**:621–31.

14. Scruggs TE, Mastropieri MA, Casto G. The quantitative synthesis of single subject research: methodology and validation. *Remedial and Special Education* 1987;**8**:24–33.

15. White OR. Some comments concerning 'The quantitative synthesis of single-subject research'. *Remedial and Special Education* 1987;**8**:34–9.

16. Corcoran KJ. Aggregating the idiographic data of single subject research. *Social Work Research and Abstracts* 1985;**21**:9–12.

17. White DM, Rusch FR, Kazdin AE, Hartmann DP. Applications of meta analysis inindividual subject research. *Behavioral Assessment* 1989;**11**:281–96.

18. Jayaratne S, Tripodi T, Talsma E. The comparative analysis and aggregation of single case data. *J Appl Behav Sci* 1988;**24**:119–28.

19. Slavin RE. Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educ Res* 1986;**15**:5–11.

20. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis (review). *J Clin Epidemiol* 1995;**48**:9–18.

21. Letzel H. 'Best-evidence synthesis: an intelligent alternative to meta-analysis': discussion. A case of 'either-or' or 'as well' (comment). *J Clin Epidemiol* 1995;**48**:19–21.

22. Meinert CL. Meta-analysis: science or religion? *Controlled Clin Trials* 1989;**10**:257S–63S.

Chapter 29

Unusual/developing areas of application in meta-analysis

Meta-analysis for estimation of admixture component in genetic epidemiology

Li (1) presents a multiplicative[1] random effects model for meta-analysis. This is in contrast with the additive models used in this report to combine treatment differences (see chapter 10). A requirement for a multiplicative model stems from estimation of the admixture component in human genetics, used in genetic epidemiology and DNA fingerprinting. This is the motivating example for this model. A random effects multiplicative model is proposed for estimation of the admixture component by combining unique allele frequencies from several loci.

Paper (1) gives a description of the model which uses the EM algorithm to find MLEs under an EB framework.

Meta-analysis of animal experiments

Freedman (2) describes the meta-analysis of animal experiments designed to investigate the effects of dietary fat intake upon mammary tumour development. Logistic regression models (fixed effect) are used to relate the mammary tumour incidence in different groups of rodents to their dietary intake. An experiment 'effect' is included to ensure that estimated nutrient effects are based only upon within-experiment comparisons. Freedman notes that the nutrient effects may not be estimable from each individual experiment, but only from combinations of experiments. (see also chapter 26 for a model that incorporates single arm studies into a meta-analysis).

The author flags up that no ideal method for testing for heterogeneity of effects is available for this situation, and that if heterogeneity were found, a random-effects version of their models would need to be developed.

Meta-analysis in pharmacokinetics

Keller *et al.* (3) present a comparison of the statistical methods for meta-analysis to standardise the different results of studies in pharmacokinetics using an example based on renal insufficiency data. Four methods are compared:

- **Method 1:** A number of linear regression equations are incorporated into a single regression function by a *Z*-transformation [see also Hedges and Olkin (4) for details];
- **Method 2:** The standardised correlation coefficient *(R)* from *k* different linear regression equations and correlation coefficients *(r)* can be calculated with an ML method [see also Hedges and Olkin (4) for details];
- **Method 3:** The weighted arithmetic mean value can be determined from the extreme values [see (3) for details];
- **Method 4:** Using a non-parametric method, regression data can be assessed together with published extreme values [see (3) for details].

In this particular example, method 4 was the most reliable. The authors go on to state:

> 'A new method is needed with which the pharmacokinetic changes in renal insufficiency reported in published studies can be submitted to meta-analysis and subsequently summarised in a standardised database.' (3)

Environmental risk studies

In recent decades, there has been mounting interest in the threat of human disease resulting from the exposure to man-made environmental agents, and also from diet and personal habits such as cigarette smoking (5). Several environmental risk models are available, usually to estimate a definition of safe dose. These include: standard survival models, biologically motivated stochastic models, and the two-event clonal expansion model.

[1] For more on multiplicative models see Hedges and Olkin [(4), p. 315].

These are reviewed in (5). Below is a very brief discussion of the methodology used to combine such information, followed by an example.

Combining environmental information

Cox *et al.* (6)[2] report from a workshop investigating the following data problems:

1. Combining environmental data from multiple and diverse sources: statistical reporting on environmental conditions and trends in aquatic, terrestrial, and atmospheric settings, and combining design-based ecological data and observed data for environmental assessment purposes.
2. Combining environmental epidemiological studies for hazard identification and risk assessment. Problems such as assessing risk of exposures to nitrogen dioxide using Bayesian methods to model uncertainty in effect estimation.
3. Forming environmental indicators and indexes, including issues of aggregation, combined mapping procedures, and multiple data source conformance.

Methodology for combining probability based and non-probability based monitoring data is presented. Also, issues in combining spatially referenced monitoring and assessment data are discussed, as well as developing and combining ecological indicators and indexes.

More immediately relevant is a section on combining information in environmental epidemiology. A discussion of combining *p*-values is presented (see chapter 7), a method for combining *p*-values where the samples are of material having multiple (and correlated) toxic features is discussed [see Mathew *et al.* (7) also for more information].

An example investigating the effect of efforts to reduce childhood exposure to lead is discussed. Data were available from three cities; the authors report:

> 'To estimate changes in blood levels during the abatement period, and to incorporate site differences encountered as each city addressed and implemented its abatement strategy, city-specific structural equations were modelled to account for different lead pathways into the bloodstream.' (6)

There were many problems with this analysis:

> 'the complexity of the different models overwhelmed any gains in sensitivity that data combination was able to provide.' (6)

Application of dose–response in non-cancer toxicity is also discussed. This combined cross-sectional epidemiological cohort study data with sub-chronic laboratory rodent toxicity data, and was implemented using the confidence profile method (see page 202).

Stratified ordinal regression: a tool for combining information from disparate toxicological studies

Cox and Piegorsch discuss the development of methodology for combining studies on acute inhalation assessment:

> 'This project involves combination of data from studies of inhalation damage from various airborne toxins in order to estimate human health risk. The studies vary greatly in their endpoints: short- and long-term exposures in laboratory animals, acute exposures to humans in chemical and/or community accidents, chronic exposure studies in urban areas, etc. .... The research goal is to develop methodology for data combination that incorporates the range of endpoint severity, exposure concentrations, and exposure durations. Particular emphasis is directed at acute exposures, since these are thought to be more common than chronic, long-term exposures in many human situations. The paradigm is based on severity modelling, wherein concentration, duration, and response are integrated to determine potential risks to humans after acute inhalation exposure to some environmental toxin. The method groups the response data into ranked severity categories, and assumes that duration and concentration are independent explanatory variables for predicting response. This is essentially an ordinal regression, using a logistic or another discrete-data regression model for the concentration–duration response. From the regression, one wishes to estimate the level at which an exposed subject will respond with a small probability, say 10%. This is the 10% effective dose, or ED10, for which a lower bound is calculated, at say, 95% confidence. For risk assessment this lower bound is the divided by an arbitrary "safety factor".' (6)

Details are given in the report (8) of the stratified, random effects, ordinal logistic regression model proposed, along with an illustrative example.

Hierarchical model applications in benchmark dose analysis

Piegorsch and Cox (9) discuss methods for establishing a benchmark dose through combining

[2] This work has now been written up as a paper by Cox and Piegorsch (12); however, the content in some sections is different.

information from different studies. They suggest a Bayesian hierarchical modelling approach may be suitable in some situations; however, classical hierarchical models could also be used. Several different applications are discussed (see paper for further details).

Combining parametised models using forecasting examples from AIDS research

The National Research Council report on combining information presents an example of synthesising results in a forecasting setting [(5), p. 165]. Their example is the problem of forecasting the future size of the AIDS epidemic in the United States (or world-wide). One approach to the problem put forward by Taylor (10) involves formulating stochastic models that include two components: the growth of the HIV infection epidemic over time, and the distribution of lag-time from HIV infection to AIDS. Taylor identified 21 plausible lag-time distributions and five possible parametric models for the growth of the HIV epidemic. This gives rise to $(21 \times 5) = 105$ different possible forecasts. The problem thus posed is:

> 'It is natural to suppose that one can do better in prediction and uncertainty assessment by combining the information in these forecasts in some way, but how is this to be done sensibly?' (5)

Three possible solutions are put forward: sensitivity analysis, weighted average of individual forecasts and model mixing. Each of these is discussed in the report (5).

More recently, in a similar vein, Cooley *et al.* (11) conducted a meta-analysis of estimates of the AIDS incubation distribution. Information was combined from 12 studies to estimate the distribution with greater precision than is possible from a single study. The modelling approach to the incubation distribution was through a hazard function of which the form was unknown for times in the future.

References

1. Li Z. A multiplicative random effects model for meta-analysis with application to estimation of admixture component. *Biometrics* 1995;**51**:864–73.
2. Freedman LS. Meta-analysis of animal experiments on dietary fat intake and mammary tumours. *Stat Med* 1994;**13**:709–18.
3. Keller F, Erdmann K, Giehl M, Buettner P. Nonparametric meta-analysis of published data on kidney-function dependence of pharmacokinetic parameters for the aminoglycoside netilmicin (review). *Clin Pharmacokinet* 1993;**25**:71–9.
4. Hedges LV; Olkin I. Statistical methods for meta-analysis. London: Academic Press, 1985.
5. National Research Council. Combining information: statistical issues and opportunities for research. Washington DC: National Academy Press, 1992.
6. Cox LH, Piegorsch WW. Combining environmental information: environmetric research in ecological monitoring, epidemiology, toxicology, and environmental data reporting. Research Triangle Park, NC: National Institute of Statistical Sciences, #12, 1994.
7. Mathew T, Sinha BK, Zhou L. Some statistical procedures for combining independent tests. *J Am Statist Assoc* 1993;**88**:912–19.
8. Carroll RJ, Simpson DG, Zhou H, *et al.* Stratified ordinal regression: a tool for combining information from disparate toxicological studies. Research Triangle Park, NC. National Institute of Statistical Sciences, #26, 1994.
9. Piegorsch WW, Cox LH. Combining environmental information. 2. Environmental epidemiology and toxicology. *Environmetrics* 1996;**7**:309–24.
10. Taylor JM. Models for the HIV infection and AIDS epidemic in the United States. *Stat Med* 1989;**8**:450–8.
11. Cooley PC, Myers LE, Hamill DN. A meta-analysis of estimates of the AIDS incubation distribution. *Eur J Epidemiol* 1996;**12**:229–35.
12. Cox LH, Piegorsch WW. Combining environmental information. 1. Environmental monitoring, measurement and assessment. *Environmetrics* 1996;**7**:299–308.

Part H:

Results VII – summary, recommendations and further research

Chapter 30

Summary, recommendations, and further research

Recommendations for systematic review practice by health service researchers

For the most part, recommendations to health services researchers or health technology assessors undertaking systematic reviews and meta-analyses follow standard and widely agreed approaches to these methods in other contexts. Greater latitude in the nature of studies potentially eligible for review – including non-randomised studies and the results of audit exercises, for example – may, however, be appropriate. The key stages are (with extensions and/or less widely agreed aspects in parentheses):

1. Specification in a protocol of the objectives, hypotheses (in both biological and healthcare terms), scope, and methods of the systematic review, before the study is undertaken.
2. Compilation of as comprehensive a set of reports as possible of relevant primary studies, having searched for all potentially relevant data, clearly documenting all search methods and sources.
3. Assessment of the methodological quality of the set of studies (the method being based on the extent to which susceptibility to bias is minimised, and the specific system used reported). Any selection of studies on quality or other criteria should be based on clearly stated *a priori* specifications. The reproducibility of the procedures in 2) and 3) should also be assessed.
4. Identification of a common set of definitions of outcome, explanatory and confounding variables, which are, as far as possible, compatible with those in each of the primary studies.
5. Extraction of estimates of outcome measures and of study and subject characteristics in a standardised way from primary study documentation, with due checks on extractor bias. Procedures should be explicit, unbiased and reproducible.
6. Perform, where warranted by the scope and characteristics of the data compiled, quantitative synthesis of primary study results (meta-analysis) using appropriate methods and models (clearly stated), in order to explore and allow for all important sources of variation (e.g. differences in study quality, participants, in the dose, duration, or nature of the intervention, or in the definitions and measurement of outcomes). This will often involve the use of mixed/hierarchical models, including fixed covariates to explain some elements of between-study variation, in combination with random effects terms.
7. Performance of a narrative or qualitative summary, where data are too sparse, or of too low quality, or too heterogeneous to proceed with a statistical aggregation (meta-analysis). In such cases, the process of conduct and reporting should still be rigorous and explicit.
8. Exploration of the robustness of the results of the systematic review to the choices and assumptions made in all of the above stages. In particular, the following should be explained or explored:
   - the impact of study quality/inclusion criteria;
   - the likelihood and possible impact of publication bias;
   - the implications of the effect of different model selection strategies, and exploration of a reasonable range of values for missing data from studies with uncertain results.
9. Clear presentation of key aspects of all of the above stages in the study report, in order to enable critical appraisal and replication of the systematic review. These should include a table of key elements of each primary study. Graphical displays can also assist interpretation and should be included where appropriate. CIs around pooled point estimates should be reported.
10. Appraisal of methodological limitations of both the primary studies and the systematic review. Any clinical or policy recommendations should be practical and explicit, and make clear the research evidence on which they are based. Proposal of a future research agenda should include clinical and methodological requirements as appropriate.

Summary of review findings

At the end of many of the chapters in this report the main findings are summarised. For convenience, these are compiled below.

Chapter 3 Procedural methodology

This section is not intended to be anything more than a brief overview of the issues that are important when one is considering carrying out research synthesis. It may help the researcher who is new to the subject to get a feel for the discipline, and serve as a springboard into later sections of this report as many of the issues touched on here are expanded in later sections.

Chapter 4 Searching the literature and identifying primary studies

This section has concentrated on searching the literature and identifying primary studies that might potentially be included in a systematic review or meta-analysis. The main point identified is that there is no one single search strategy that would provide adequate results, and that in performing reviews researchers should maintain a healthy degree of scepticism about any or all their searches. However, a second key point is that all searches/ methods that are used should be sufficiently well documented so that they may be replicated by other researchers. This latter point is equally important as regards study inclusion/exclusion.

Finally, changes are happening rapidly in terms of electronic publishing and databases. Such changes will undoubtedly have profound implications for conducting systematic reviews in the future.

Chapter 6 Study quality

This chapter has considered both the assessment and use of quality of scores in meta-analysis. Whilst a number of methods have been proposed for assessing study quality (of primary studies) in a meta-analysis, no consensus appears to have developed as to which method is most appropriate, or indeed whether such an exercise is appropriate at all. As far as the use to which such quality scores can be put, a number of possibilities exist, but in specific situations the meta-analyst should not be totally reliant on any one method, in addition that is to an unadjusted analysis.

Chapter 7 Simple methods for combining studies

This chapter has considered principally two basic methods for synthesising evidence; vote counting and the combination of *p*-values. Whilst vote counting is one of the simplest methods available, it should only be used if absolutely necessary. By contrast, although the combination of *p*-values does convey some aspect of effect size, there are a number of disadvantages to the use of such a method. As a result, it should only be used with caution, since it may mask some fundamental differences in the studies.

Chapter 8 Heterogeneity

In conclusion, we are some way off agreeing upon the best strategy for dealing with heterogeneity. It seems essential to look for it and test for it and sensible to explore possible reasons for its presence. When a sizeable amount of unexplained heterogeneity is still present after this, a judgement has to be made on whether it is appropriate to combine the results; if so with what model; and what conclusions can be drawn from it. Presently, these decisions require a large degree of subjectivity on the part of the reviewer. Whatever approach is used, ‘it is invalid to delete from the set of studies to be meta-analysed those whose results are in the ‘wrong direction,’ for the opportunity for bias in identifying the ‘deviant’ studies is too great’ Fleiss (1).

Chapter 9 Fixed effects

This chapter has considered the so-called fixed effect approach to meta-analysis. This assumes that all the studies in a meta-analysis are estimating the same underlying unknown true intervention effect. A variety of estimation methods have been proposed for such models, whilst in many situations they give qualitatively similar results, in some circumstances differences can be serious. In terms of binary data, problems with a number of methods occur if there are zero events in any treatment arms in any study. In such circumstances there has been some empirical work reported on the various methods advocated for overcoming this problem. Meta-analysts should report precisely what methods have been used in such circumstances.

Chapter 10 Random effects

At this present time, it would seem neither fixed nor random effect models could be considered the ideal analysis, beyond any dispute, for a given situation. Indeed, it has been illustrated that both methods have their shortcomings. As the point estimates of effect size given by both methods are usually very similar, the only time the choice of model will be critical is if its significance is marginal using a fixed effect model. Here there is a chance that the more conservative CI given by the random effects approach would consider the effect to be non-significant. It is interesting to note that Peto, one of the strongest opponents of random effects models takes 3 standard deviations rather than 2 (1% not 5%) as his critical value when considering the significance of a (fixed) effect in an overview, considering 2 standard deviations to be not stringent enough for the magnitude of the implications of an overview. ('.........we are messing around if we take two standard deviations, two-and-a-half standard deviations, as serious evidence. We get so much nonsense mixed up in with the sense that it is just irresponsible. I think we've got to get better standards of evidence than we normally have, and this means in the individual trials and in overviews. I think you need to go to at least three standard deviations.' (2) The point in mentioning this is that one of the world leaders in the field, although conceptually at poles with the advocators for random effects, through this more stringent cut-off point is actually making an adjustment with practical implications very similar to those inherent by the use of a random effects model. While it would appear that the conceptual debate over the correct model is some way off a conclusion, a practical line to take may be to say: use whichever strategy (single analysis or several) you yourself feel is most appropriate for the situation, but if there is evidence of heterogeneity (significant or not) and a fixed effect analysis is the sole analysis carried out and the result is only marginally significant (5% level) then extreme caution is needed when reporting and interpreting the results. Another key point to consider here relates to the clinical significance rather than the statistical significance of the pooled estimate obtained. One should be concerned about estimates and their SEs, rather than *p*-values. It should be pointed out that other models do exist for meta-analysis, chapter 12 covers mixed models, and chapter 13 Bayesian models. It is interesting that the National Research Council (3) take the approach of calling random (and fixed) effects models a special case within a hierarchical model framework, of which other models (such as mixed, cross-design synthesis (chapter 26) are simply extensions. Another point worthy of note is that when using a Bayesian approach, one does not necessarily have to choose between the two models (fixed and random), but rather we can average across models using BFs (see chapter 13).

Chapter 11 Meta-regression

This chapter has extended the methods of chapter 9 (fixed effects) to take account of the fact that there are often covariates at either the study-level or patient level available, and that these can be important in helping to explain any heterogeneity present. Such an analysis should be seen as a fundamental component of any meta-analysis, but as with any modelling exercise due care and attention should be paid to the verification of any assumption the models make. One of the potential advantages of this approach is that estimates of the relative benefits of treatments for patients with different combinations of covariates can be derived, or more information on the relative effect of different forms of delivering the intervention. This is the sort of data that is very relevant to clinical practice, where overall average effects may be too general to be useful for particular situations.

Chapter 12 Mixed models

This chapter has extended the methods of meta-regression in chapter 11, to allow for the existence of between study heterogeneity that cannot be adequately modelled by fixed covariates in a meta-regression model. The simplest models simply allow for a single random effect term, whilst more complicated models can allow for different levels of between-study heterogeneity associated with differing levels of a factor using a hierarchical modelling framework.

Chapter 13 Bayesian methods in meta-analysis

This chapter has summarised the general use of empirical and fully Bayesian methods with respect to meta-analysis, and in particular a number of specific areas in which there has been considerable research over the last few years, and in which Bayesian methods have a potential role to play. Although currently much research is been put into these methods, so far their use in practice is far from routine. Distinct advantages of the Bayesian approach include the ability to incorporate *a priori* information which would otherwise be excluded in a classical analysis. However, when such *a priori* evidence is based on subjective beliefs, the issue of whose prior beliefs to use is raised. Though many of the computational difficulties that have plagued the application of Bayesian methods in practice have been partially solved by recent development in MCMC methods, these should not be seen as 'black box' methods since they raise issues concerning convergence.

Chapter 14 Combining other measures

This chapter presents other scales commonly used when assessing outcomes in medical research. One needs to be aware that scales other than ORs and standardised mean differences exist and can be used to combine studies. Additionally, it is important to note that since different studies may report outcomes on different scales then it may be necessary to transform a proportion of them before synthesis can proceed. Methods for doing this are presented in the next chapter.

Chapter 15 Issues concerning scales of measurement when combining data

This chapter has considered some of the issues that must be considered when deciding which scales of measurement are to be used when combining data. Though there are specific statistical methods that can be employed when the studies in a meta-analysis use a variety of measurement scales, so as to produce a single unified scale of measurement, a number of issues should be considered. Firstly, that different scales may lead to different results, both quantitatively and qualitatively. Secondly, the most convenient common scale, statistically, may not be the most appropriately clinically. Finally, where possible sensitivity analyses should be performed to check the inter-dependence between the quantitative result obtained and the measurement scale used.

Chapter 16 Publication bias

In conducting a meta-analysis, researchers should always be aware of the potential for publication bias, and make efforts to assess to what extent publication bias may affect their meta-analysis. In terms of the inclusion of unpublished studies, a sensitivity analysis should be performed to assess the likely impact of including unpublished data.

The intention of above sections was to give the reader a brief but relatively complete overview of the methods proposed to deal with publication bias. It has already been noted that many of the methods are new and exploratory.

Chapter 17 Missing data

Not a lot has been written on the problem of missing data in meta-analysis. Most of the methods discussed here have been adapted from other situations. Many of the advanced methods have not been used extensively in a meta-analysis setting (4). Pigott suggests that the current development of computer programs that implement the procedures described by Little and Rubin (5) should advance the development of sensible methods for handling missing data in research synthesis (4).

Cooper and Hedges state (6) that missing data is ‘perhaps the most pervasive practical problem in research synthesis’. The also observe that ‘the prevalence of missing data on moderator and mediating variables influences the degree to which the problems investigated by a synthesis can be formulated’, and predict that new methods will evolve, and that:

> ‘Much of this work will likely be in the form of adapting methods developed in other areas of statistics to the special requirements of research synthesis. These methods will produce more accurate analyses when data are not missing completely at random but are well enough related to observed study characteristics that they can be predicted reasonably well with a model based on data that are observed.’ (6)

When covariate information is missing this can be a problem when analysing heterogeneity using meta-regression (see chapter 11) as Pigott explains:

> ‘A synthesist may try several different analyses with the data to determine if any of a study’s characteristics relate to the effect magnitude of the study. In each of these analyses, only studies with complete information on relevant variables may be included. Each of these analyses may utilize a different set of studies that may not be representative of the sample originally chosen for the synthesis and may not correspond with each other. The results of each analysis may not generalise to the population of studies on a topic nor to any of the other samples of studies used in the analyses.’ (4)

It should be stressed that whatever method is used to deal with missing data, a careful sensitivity analysis of the modelling assumptions on the conclusions should be performed as a final step.

Chapter 18 Reporting the results of a meta-analysis

This chapter has given a brief overview of methods used to report a systematic review. It is recommended for researchers to include tables of all studies considered in a review, so possible to see which were excluded. The bottom line on reporting a review is that enough information should be provided so people can replicate, or carry out changes/updates to it.

Chapter 19 Meta-analysis of observational studies

Most of the considerations for combining observational studies are the same as those outlined in the rest of the report for RCTs. One new question that needs addressing is ‘Has proper control or adjustment been made for the biases that frequently occur in epidemiological studies, such as sociodemographic or clinical differences between study populations, misclassification of subjects with regard to case–control status and to levels of exposure, factors other than the level of exposure that may affect whether a subject is a case or a control (i.e. confounding variables).’ (1) Key references on this subject are the seminal paper by Greenland (7) and the set of guidelines reported by Blair *et al.* (8). The use of sensitivity analysis to deal with the above problems is emphasised.

Chapter 20 Meta-analysis of survival data

Survival analysis data requires specialist meta-analysis techniques (as well as specialist statistical methods in general) because of data censoring. If this censoring is ignored, this may bias the overall estimates. Other than this problem, the various standard approaches for meta analysis are possible. In such instances methods such as finding summary measures for survival data (such as the hazard ratio), and then combining those is possible.

Chapter 21 Meta-analysis of diagnostic test data

The CMWG on Systematic Review of Screening and Diagnostic Tests (9) remarks that pooling of accuracy assessments within the Cochrane Collaboration will probably use dichotomised (binary) test data because, first, most primary studies present the data in this format and, second, further research on and developments of statistical methods for ordered categorical and continuous test outcomes is needed. Their method of choice is the analysis of the SROC curve in both the unweighted and weighted manner.

Chapter 23 Methods for correlated outcomes: combining multiple effect measures

In order to combine multiple effect measures the correlations/covariances between outcomes are needed for most methods. If these are not available, then one must make a guess at them, and assess the impact of their choice using a sensitivity analysis, or alternatively estimate them from external sources, or IPD from some of the trials.

Chapter 24 Meta-analysis of individual patient data

There are several advantages of carrying out a meta-analysis of IPD, over a standard meta-analysis using aggregated data. These include the ability to: 1) carry out detailed data checking, 2) ensure the appropriateness of the analyses, and 3) update follow-up information. This has led to the comment that MAP data are the yardsticks against which the quality of other systematic reviews of randomised controlled trials should be measured (10).

These benefits do not come without a cost, however, as IPD meta-analyses are very time consuming and costly. Currently, there is little empirical evidence regarding the actual magnitude of the gains, and it is yet to be established whether the extra effort is worthwhile, in given situations.

Chapter 25 Cumulative meta-analysis

Cumulative meta-analysis is valuable as an exploratory/sensitivity analysis tool. There are questionable gains if it is done in real time, and a correction for multiplicity is needed for the frequentist approach.

Chapter 26 Generalised synthesis of evidence

Methodology is becoming available for combining studies of different designs. Many of the techniques utilise modern statistical models, including Bayesian methods, and hence many of the methods are extensions of those of chapter 13. Their implementation is however facilitated by recent advances in computer software. When such analyses are appropriate, is still open to debate, as there is concern that including studies with poorer designs will weaken the analysis, though this issue is partially addressed by conducting sensitivity analyses under various credibility assumptions.

Agenda for further research

This section summarises areas and issues requiring further (methodological) research effort. One or two priority areas are explicitly indicated for further HTA funding. The detailed background to specific recommendations in the list below can be found in the relevant chapters from which they are drawn. For convenience, the source chapter is included after each recommendation.

Priority areas for further HTA funding

1. Sensitivity analysis of the impact of many aspects of the design and analysis of the systematic review, and in particular of the meta-analysis, has been advocated. The result is a complex set of inter-related sensitivity analyses. Research into optimum, or at least efficient, strategies of multi-dimensional sensitivity analysis in these contexts would thus be useful.
2. Evaluation of the role in HTA of meta-analysis of observational studies, and cross-design synthesis (which often features the inclusion of non randomised evidence), possibly through systematic research and workshops of researchers active in the field.

Heterogeneity, publication[1] and related biases, study quality

1. Investigation into the relevant dimensions of methodological quality and empirical research which establishes the relative importance of these dimensions in different contexts (diseases, endpoints, study designs etc.). This should then lead validated instruments to assess these dimensions individually, which should finally lead to the development of rigorous, validated, and parsimonious scales for the (repeatable) assessment of study quality.
2. Exploration of the role of study quality as an explanation of heterogeneity
3. Empirical investigation into the basis for choice of cut-off values for exclusion of studies on grounds of quality, and of robustness of this value.
4. Development of systematic approaches to quality assessments of non-RCTs, and their use in meta-analysis and cross-design synthesis.
5. Further investigation of the relationships between heterogeneity and publication bias, including the development of methods for assessing heterogeneity taking into account selection bias and vice-versa.
6. Development of guidelines/recommendations for the identification and exploring heterogeneity.
7. Investigation of degree of heterogeneity (both quantitative and qualitative) beyond which combining of all the studies should not be considered.
8. Investigation into the effects of choice of measurement scale (e.g. choice between OR and RR measures, or use of a mixture of the two), from both: a) a statistical perspective, and b) a clinical perspective.
9. Further investigation of aspects of publication bias, including:
   – the relative merits of methods based on the funnel graph versus methods based on weighted distribution theory;
   – assessing the impact of the pipeline problem;
   – development of a test that is sensitive to either the magnitude of the estimate of effect size in a primary study or the significance level of the test for treatment or other effects in primary studies, since either may influence publication bias;
   – development of methods for estimation of chances of failing to detect a bias that would have a profound effect on the results of a meta-analysis;
   – investigation of power to detect publication bias, and in particular of the influence of the number of primary studies;
   – empirical study of degree and mechanisms of publication bias in meta-analysis of epidemiological and other non-randomised studies;
   – investigation into the extent the use of a prospective register for trials minimises publication bias;
10. In general, identification of the most sensitive approaches for detecting publication bias.
11. General investigation of the impact of missing values, and extension of currently available methods to a wider range of circumstances with missing data, including the use of Bayesian methods.
12. Development of the use of simulation of results of new studies before they are published or of hypothetical studies to allow their impact on meta-analysis to be assessed.
13. Further development of detailed publication guidelines to encourage uniform reporting of the results of studies, particularly of types other than RCTs.
14. Investigation of which methods are superior for investigating the baseline risk of patients in studies.

Approaches to modelling and analysis

1. Investigation of the relative merits of the different approaches to combining studies in which some arms report no events (zeros in 2 × 2 tables).
2. Comparison of new methods for random effects modelling incorporating all modelling parameter uncertainty.
3. Investigation of robustness of random effects models to departures from normality, and further consideration of use of likelihood methods or of other distributional assumptions.
4. Empirical investigation of model attributable weights with particular reference to over-weighting of large samples, in some models.
5. Investigation of the impact of missing data at both the study level and patient-level.
6. Development of experience with practical applications of mixed models, including criteria for the identification of covariates as fixed or random, and specification of any hierarchical structure.

[1] Note: HTA programme has commissioned a separate review in this area.

7. Development of methodology for combining IPD with study level data.
8. Investigation of the role of cumulative/sequential application of meta-analysis as a research methodology, including:
   - development of criteria for stopping rules for sequential meta-analysis, perhaps based on group sequential methods;
   - formalisation of Bayesian approaches to sequential/cumulative meta-analysis and investigations of their properties.
9. Investigation of the incorporation of data of different types.
10. Further development of methods for integration of qualitative assessments of studies with quantitative estimates of the results.
11. Development of random/mixed effects models for meta-analysis of survival data.
12. Use and implications of the exact methods (fixed effects models); should they be used? If so, when?
13. More extensive but critical use of Bayesian methods, including:
    - encouragement of **expository papers** in the applied literature on the application of Bayesian methods;
    - more research on the use of **elicited prior beliefs**.
14. More research into the use of meta-analytic techniques in conjunction with decision analysis methods, that take into account the uncertainty associated with any meta-analysis findings.
15. More research into extrapolation and use of results of a meta-analysis to clinical practice and healthcare policy making.

References

1. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127–39.

2. Peto R. Why do we need systematic overviews of randomised trials? *Stat Med* 1987;**6**:233–40.

3. National Research Council. Combining information: statistical issues and opportunities for research. Washington DC: National Academy Press, 1992.

4. Pigott TD, Cooper H, Hedges LV, editors. Methods for handling missing data in research synthesis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 163–76.

5. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley, 1987.

6. Cooper H, Hedges LV, Cooper H, Hedges LV, editors. Potentials and limitations of research synthesis. In: The handbook of research synthesis. New York: Russell Sage Foundation, 1994, p. 521–30.

7. Greenland S. Quantitative methods in the review of epidemiological literature. *Epidemiol Rev* 1987;**9**:1–30.

8. Blair A, Burg J, Foran J, Gibb H, Greenland S, Morris R, *et al.* Guidelines for application of meta-analysis in environmental epidemiology. ISLI Risk Science Institute. *Regul Toxicol Pharmacol* 1995;**22**:189–97.

9. Irwig L, Glasziou P. The Cochrane methods working group on systematic review of screening and diagnostic tests: recommended methods. Online: http://www.som flinders edu au/cochrane/, 1996.

10. Chalmers I, Sandercock P, Wennberg J. The Cochrane collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann N Y Acad Sci* 1993;**703**:156–65.

Appendix 1

Strategies used to search the electronic databases to identify relevant publications for this report

General approach to searching electronic databases

It was perceived that there are three types of publications relevant to the subject of review synthesis, namely:

1. References concerned with the methodology of research synthesis/meta-analysis.
2. References reporting the results of a particular synthesis of evidence/meta-analysis, these contain no new methodology content.
3. References presenting the results of a synthesis/meta-analysis but, in the process developed new methodology which is presented also.

A Venn diagram *(Figure 7)* illustrates this situation.

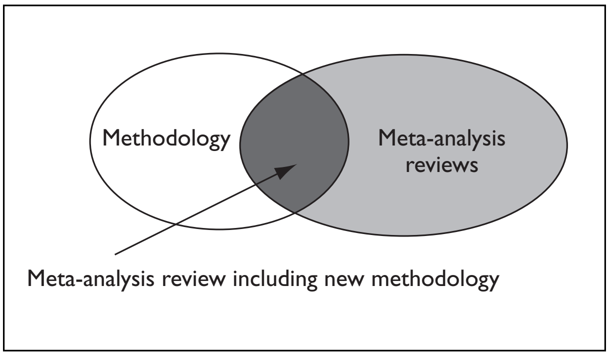


FIGURE 7 *A Venn diagram*

For the purposes of this review, publications of type 1 and 3 above are of interest, while those in group 2 are of much lesser importance, and are generally ignorable, except in so far as they provide illustrative examples.

A preconception that was confirmed during the searching was that to obtain a large proportion of 1, 2 and 3 would not be too difficult. The difficult task would be to separate group 2 from group 1. Even more difficult would be identification and separation of group 3 (as distinct from group 2), even if an abstract was available.

Hence, the search strategy process was conceptualised in two stages:

- **Stage 1**: a strategy to retrieve all papers concerned with systematic reviews and/or meta-analysis, and then:
- **Stage 2**: a strategy to split the methodology (groups 1 and 3) from the non-methodology (group 2).

The following strategies were all implemented in July/August 1996.

Stage 1: strategy

Two approaches were considered for stage 1. The first was based on a union of all the words we could think of to describe systematic reviews/ meta-analysis (our strategy). The second made use of the search strategies published by the CRD[1] (1); two search strategies, broad (CRD broad)) and less broad (CRD short), for identifying systematic reviews in MEDLINE. These are designed to 'maximise the recall of potentially systematic reviews and seek to minimise the recall of reviews which appear to be non-systematic or narrative' (1). The three strategies for use with MEDLINE are reproduced below.

CRD broad

This is the first of the CRD searches. Key points from the CRD notes which accompany it are as follows:

> 'employs a very broad search strategy for identifying reviews in Medline. It is designed to maximise the recall of potentially systematic reviews and seeks to minimise the recall of reviews which appear to be non-systematic or narrative....Systematic reviews are not consistently indexed by Medline so the search strategy includes textwords in addition to MeSH headings.'

[*] The versions used were from the website: http://www.york.ac.uk/inst/crd/search.htm.
These are slightly different from the published versions.

> 'Case reports and historical reviews are explicitly excluded. Further limits in Medline could be applied to restrict searches depending on the specific search requirements. The search strategy above will usually return a large volume of references which will then need to be sifted.' (1)

Table 19 shows the system used.

Search locations:	ti = words in title
	sh = MeSH subject headings
	tw = textwords
	pt = publication type
Key:	$ = the truncation symbol
	adj4 = adjacent (within four words)
	ab = words in abstract
	exp = explode term (include all sub terms also)
Applied to:	Database: MEDLINE (CD-ROM)
	Dates: 1992–1996 August Disc

Comments

The term in bold is an alteration necessary to get it to run on the system we used (Ovid CD-ROM).

This search results in a massive final count of 142,711 references.

CRD short

This is the second of the CRD searches. Key points of notes which accompany it are as follows:

> 'A more precise version of CRD broad is shown below. This excludes the majority of non-MESH search terms and is more reliant on the consistency of MEDLINE indexing.' (1)

Table 20 shows the system used.

Search locations:	ti = words in title
	sh = MeSH subject headings
	tw = textwords
	pt = publication type
Applied to:	Database: MEDLINE (CD-ROM)
	Dates: 1992–1996 (August disc)
Key:	$ = the truncation symbol
	adj4 = adjacent (within four words)
	ab = words in abstract

Comments

There are many fewer references, 23,538 (only 16% of those found by CRD broad).

Our strategy

Where possible, MeSH terms as well as text words were included. The facility adj4 (within the adjacent four words) is also taken advantage of where possible. *Table 21* shows the system used.

TABLE 19

Line number	Search term	Number found
1	meta-analysis.sh.	927
2	meta-analy$.tw.	1628
3	metaanaly$.tw.	78
4	(systematic$ adj4 (review$ or overview$)).tw.	304
5	meta-analysis.pt. meta analysis.pt.	1508
6	exp review literature.sh.	283
7	review.pt.	171,724
8	review.ti.	13,394
9	review literature.pt.	6943
10	(overview adj4 trial$).tw.	100
11	consensus development conference.pt.	944
12	case report.sh.	154,470
13	historical article.pt.	15,590
14	review of reported cases.pt.	12,834
15	review, multicase.pt.	2333
16	or/1-11	177,729
17	or/12-15	173,639
18	16 not 17	156,929
19	animal.sh.	456,180
20	human.sh.	1,066,147
21	19 not (19 and 20)	341,003
22	18 not 21	142,711
23	Your subject specific search terms	*
24	22 and 23	**

*Key: *Entering terms on a specific subject here and combining in ** will find meta-analyses on the specific subject*

Search locations:	ti = words in title
	sh = MeSH subject headings
	tw = textwords
	pt = publication type
Key:	$ = the truncation symbol
	adj4 = adjacent (within four words)
	ab = words in abstract

TABLE 20 *System used for CRD sheet*

Line number	Search term	Number found
1	(meta-analysis or review literature).sh.	1013
2	(meta-analy$ or (meta adj anal$)).tw.	1628
3	metaanal$.tw.	78
4	meta-analysis.pt.	1508
5	review, academic.pt.	17,402
6	review literature.pt.	6943
7	case report.sh.	154,470
8	letter.pt.	98,341
9	historical article.pt.	15,590
10	review of rep[orted cases.pt.	12,834
11	review, multicase.pt.	2333
12	or/1-6	26,952
13	or/7-11	251,858
14	12 not 13	25,956
15	animal.sh.	456180
16	human.sh.	1,066,147
17	15 not (15 and 16)	341,003
18	14 not 17	**23,538**
19	Your subject specific search terms	*
20	18 and 19	

*Key: *Entering terms on a specific subject here and combining in ** will find meta-analyses on the specific subject*

Applied to: Database: MEDLINE (CD-ROM)
Dates: 1992–1996 (August disc)

Comments

The terms used in this search strategy were derived from consulting known key references in the field to see which terms were used for keywording and generally commonly used in the text.

Note the inclusion of **meta-analysis** as a publication type. As this was only introduced in 1993, however, it cannot be relied upon exclusively.

To illustrate the point that publication types and subject headings are not enough to rely on; when **meta-analysis.sh.** was combined (using Boolean argument AND) with **meta analysis.pt.** only 256 papers were indexed using both terms!

TABLE 21 *System used in this search strategy*

Line number	Search term	Number found
1	meta-analysis.sh.	927
2	meta analysis.pt.	1508
3	meta-analy$.tw.	1628
4	met-analy$.tw.	6
5	metanaly$.tw.	10
6	metaanaly$.tw.	78
7	met analy$.tw.	6
8	meta analy$.tw.	1628
9	or/1–8 (group1)	2595
10	overview$.tw.	4809
11	(synthesis adj4 evidence).tw.)	389
12	quantitative review$.tw.	17
13	quantitative synthes$.tw.	7
14	review synthes$.tw.	27
15	quantitative overview$.tw.	11
16	(systematic$ adj4 (review$ or overview$)).tw.	304
17	research synthes*.tw.	14
18	quantitative pooling.tw.	1
19	(combin* adj4 estimates).tw.	9
20	(pool$ adj4 results$).tw.	526
21	(pool$ adj4 estimate).tw.	162
22	or/10–18 (group 2)	757
23	or/19–21 (group 3)	766
24	9 or 22 (groups 1, 2)	3248
25	9 or 22 or 23 (groups 1, 2, 3)	3914

The proportions in each group are similar to those for the BIDS search (see later in appendix 1) with groups 2 and 3 providing relatively few references.

An investigation to see how many references these three strategies had in common was carried out.

Comparing strategies

This tabulation was designed to determine how many references each of the three searches had in common *(Table 22)*.

***TABLE 22** Comparing references retrieved from the various strategies*

Term	Number of references returned
CRD broad	23,538
CRD short	142,711
Our strategy	3914
CRD broad AND CRD short	23,538
CRD broad AND Our strategy	2862
CRD broad OR Our strategy	143,763
CRD short AND Our strategy	2368
CRD short OR Our strategy	25,084

Comments
The CRD short search is a perfect subset of the CRD long search.

Results

The three strategies, 'CRD long', 'CRD short', and, 'Our strategy', brought up 142,711, 23,583 and 3914 hits respectively varying by an **order of magnitude**.

Our strategy returns 1546 references that the CRD short search does not find and 1052 that CRD broad does not find either. It was quite alarming to find that a search, with the same aim, which retrieves 142,711 references, but does not find 27% of the references found in a much more specific search.

The safest way to proceed would be to use both the CRD broad and the extra references found by 'Our strategy'. However, this returns a group of 143,763 references for the last 3.5 years alone. We believed that earlier than 1991 the synthesis/meta-analysis literature was much smaller. However, we believe this would not mean the CRD broad returns would be proportionately decreased as it clearly has a very low specificity.

The next stage was to develop strategies to retrieve the methodology papers from the results of this first stage.

Stage 2: strategy

From considering the results of stage 1, it was clearly necessary to reduce the number of papers to allow manual scanning of their abstracts. The second stage was to develop a strategy to retrieve the methodology papers. Approximately 20 areas of research had been highlighted in the grant application for the project. These titles plus other associated terms were used as the basis of this second stage. This had clear limitations; firstly the strategies would need many terms and be very long and complex. As a result, many of the on-line engines had trouble coping with them. Secondly, and more fundamentally, this search is only for topics known to exist. This implies areas or research unknown to us could be missed. A pilot using a small proportion of these terms was carried out to assess which stage 1 strategies would be feasible. This is reported below.

Search: stage 2 – preliminary *(Table 23)*

Applied to: Database: MEDLINE (CD-ROM)
Dates: 1992–1996 (August disc)

***TABLE 23** Preliminary stage of search*

Term	Number of references found overall	∩ Our strategy	∩ CRD short	∩ CRD broad
statistics.sh.	1936	34	37	149
publication bias.sh.	60	18	16	25
bayes$.tw.	554	21	27	65
sensitivity analysis.tw.	297	17	17	36
predict$.tw.	56,978	259	855	4382
homogen$.tw.	15,911	89	139	508

Comments
It seems that the number of papers returned for some terms, namely *predict$.tw.* would be impractical to search through for the CRD broad strategy. For many of the terms with fewer references namely *publication bias.sh., bayes$.tw.* the broad search did not bring up many more returns.

It would seem sensible to suggest making a list of terms to search on both CRD broad and Our strategies and a list to search only on Our (and possibly the CRD short) strategy/ies.

MEDLINE stage 2 strategy *(Table 24)*

This strategy uses all the subject specific terms intersected with 'Our strategy'.

Every term suspected of having a MeSH heading was looked up in the index, and if a suitable one was found, this was used as well as the text word.

Search locations: ti = words in title
sh = MeSH subject headings
tw = textwords
pt = publication type

TABLE 24 *MEDLINE stage 2 strategy*

Term	**Search term**	**Total**	**Total ∩ 'Our strategy'**
Procedural methodology			
∇procedural methodology	*procedural methodology.tw.*	2	0
procedure/s	*procedur$.tw.*	62,836	202
guidelines	*guide$.tw.*	17,642	120
guidance	*guideline.pt.*	2640	17
framework/s	*guidelines.sh.*	1478	11
data extraction	*health planning guidelines.sh*	137	0
individual patient data	"	5250	21
study level data	*framework$.tw.*	629	141
level of aggregate/ion	*data extraction.tw.*	37	25
	individual patient data.tw.	0	–
	study level data.tw.	11	2
	level of aggregat$.tw.		
Publication bias			
publication bias	*publication bias.tw.*	67	35
literature bias	*publication bias.sh.*	60	18
reporting bias	*literature bias.tw.*	1	1
	reporting bias.tw.	32	1
Statistical issues			
∇statistical issues	*statistical issues.tw.*	41	2
∇statistical methodology	*statistical methodology.tw.*	0	–
statistics	*statistic$.tw.*	41,876	502
calculation/s	*statistics.sh.*	1936	34
	exp statistics.sh.	132,649	1152
	data interpretation, statistical.sh.	2839	87
	calculation$.tw.	7107	51
Fixed effect(s) approaches			
fixed effect/s	*fixed effect$.tw.*	118	24
homogeneity	*homogen$.tw.*	15,911	89
homogeneous	"	1072	2
∇classical method	*(classical and (approach or method).tw.*	8542	20
classical	*classical.tw.*		
Random effect(s) approaches			
random effect/s	*random effect$.tw.*	194	48
heterogeneity	*heterogene$.tw.*	14,262	142
heterogeneous	"		
Mixed effect approaches			
mixed effects	*mixed effect$.tw.*	111	7
	mixed-effect$.tw.	111	7
Study quality			
study quality	*study quality*	143	28
quality of studies	*(quality of studies ∪ quality of study)*	120	38
Influence of specific studies			
influence of specific studies	*influence of specific studies.tw.*	0	–
influence of a specific study	*influence of a specific study.tw.*	0	–
study influence	*study influence.tw.*	680	1
deletion method/s	*deletion method$.tw.*	5	0
sensitivity analysis	*sensitivity analys$.tw.*	418	26
	'sensitivity and specificity'.sh.	19,934	121

Key: ∇ *term is a subset of another term and not necessary for the search but included for interest*

continued

TABLE 24 contd *MEDLINE stage 2 strategy*

Term	**Search term**	**Total**	**Total ∩ 'Our strategy'**
Meta-regression			
meta-regression	*meta-regression.tw.*	5	5
meta-regression	*metaregression.tw.*	4	4
∇meta regression	*meta regression.tw.*	5	5
regression	*regression.tw.*	20,471	199
explaining variation	*regression analysis.sh.*	12,250	107
covariates	*exp regression analysis.sh.*	21,161	199
	explaining variation.tw.	17	0
	covariates.tw.	825	17
Cross design synthesis			
∇cross design synthesis	*cross design$ synthesis.tw.*	3	2
cross design	*cross design$.tw.*	404	4
Confidence profiling approach			
confidence profiling	*confidence profil$.tw.*	2	2
confidence profile	"		
Missing data			
missing data	*missing data.tw.*	157	3
missing value/s	*missing value$.tw.*	35	0
incomplete data	*incomplete data.tw.*	84	2
Scales of measurement			
scales of measurement	*scale$ of measurement$.tw.*	68	0
measurement scale/s	*measurement scale$.tw.*	106	0
effect/s measure/s	*effect measure$.tw.*	384	6
survival data	*survival data.tw.*	492	14
survival analysis	*survival analysis.tw.*	798	10
survival measure/s	*survival analysis.sh.*	7987	102
ordinal data	*survival measure$.tw.*	109	2
ordinal outcome/s	*ordinal data.tw.*	23	0
ordinal measure/s	*ordinal outcome$.tw.*	2	0
continuous outcome	*ordinal measure$.tw.*	4	0
number needed to treat	*continuous outcome.tw.*	9	0
economic	*number needed to treat.tw.*	14	4
∇cost effectiveness	*economic.tw.*	5413	43
cost	*cost effective$.tw.*	4385	71
	cost$.tw.	20,552	162
Reporting of the results			
reporting of the result/s	*report$ of the result$.tw.*	2666	40
report/ing the result/s	*report$ the result$.tw.*	2666	40
result/s reporting	*result$ report$.tw.*	1548	16
Prediction			
predict/ion	*predict$.tw.*	56,978	259
∇effect predict/ion	*effect predict$.tw.*	51	1
Cumulative meta-analysis			
cumulative	*cumulative.tw.*	3	0
sequential	*sequential.tw.*	8373	24
chronological	*chronological.tw.*	748	6
Multi-level modelling			
multi-level model/s	*'multi-level' model$.tw.*	3	0
multi level model/s	*multi level model$.tw.*	3	0
multilevel model/s	*multilevel model$.tw.*	15	0
hierarchical model/s	*hierarchical model$.tw.*	49	0
Key: ∇ term is a subset of another term and not necessary for the search but included for interest			

continued

TABLE 24 contd *MEDLINE stage 2 strategy*

Term	Search term	Total	Total ∩ 'Our strategy'
Empirical Bayes			
empirical bayes	*empirical bayes$.tw.*	42	3
Full Bayes			
bayes$	*bayes theorem.sh.*	536	16
	(bayes$) NOT *(empirical)* .tw.	647	21
MCMC			
markov chain	*markov chain.tw.*	85	1
monte carlo	*monte carlo.tw.*	949	3
monte-carlo	*'monte-carlo'.tw.*	949	3
Gibbs sampling			
gibbs sampling	*gibbs sampl$.tw.*	34	0
bugs	*bugs.tw.*	72	2
simulation methods	*simulation methods.tw.*	32	0

Key: ∇ term is a subset of another term and not necessary for the search but included for interest

Key: $ = the truncation symbol
adj4 = adjacent
(within four words)
ab = words in abstract
exp = explode MeSH term
Applied to: Database: MEDLINE (Ovid)
Dates: 1992–1996 (August disc)

It was decided to cross the more specific subject terms with CRD broad in the hope of finding most papers missed by crossing with 'Our strategy' *(Table 25)*.

Search locations: ti = words in title
sh = MeSH subject headings
tw = textwords
pt = publication type
Key: $ = the truncation symbol
Applied to: Database: MEDLINE (Ovid)
Dates: 1992–1996 (August disc)

This search strategy was employed for the last 5 years. Each papers database entry retrieved by stage 1 and stage 2 was scanned manually to decide its relevance and hence whether it should be included. Where there was uncertainty to its relevance the whole paper was obtained for inspection before a decision was made.

Searches of other databases

In addition to MEDLINE, similar search strategies were used for ISI Science and Social Science and EMBASE accessed through the BIDS system. Due to a less sophisticated search engine, the strategies are

TABLE 25

Search term	Total	Total ∩ CRD broad
publication bias.sh.	60	25
statistic$ method$.tw.	846	118
random effect$.tw.	194	50
mixed effect$.tw.	111	10
study quality.tw.	143	41
sensitivity analys$.tw.	418	52
meta-regression.tw.	5	5
metaregression.tw.	4	4
meta regression.tw.	5	5
cross-design synthesis.tw.	3	3
number needed to treat.tw.	14	4
multi-level model$.tw.	3	1
multi level model$.tw.	3	1
multilevel model$.tw.	15	4
empirical bayes$.tw.	42	6
baye$ NOT empirical.tw.	(lost)	72
bayes theorem.sh.	(inc above)	2
markov chain.tw.	85	32
monte carlo.tw.	949	2
monte-carlo.tw.	(inc above)	5
gibbs sampl$.tw.	34	
bugs.tw.	72	

simplified versions of those used for MEDLINE as given above. These are reproduced below.

ISI/EMBASE Stage 1 search strategy

Due to lack of MeSH terms or publication type field (or equivalents) direct adoption of the CRD searches was not possible. A simplified version of 'Our strategy' was used exclusively in stage 1. Searches were done for the previous 5 years, 1991–1996.

The list of terms below were used to search in the title, keywords and abstracts of the databases, unless otherwise specified. Each term was combined with the OR command to retrieve all references at this stage.

*meta analy**
'meta-analy'* (quotes needed, otherwise '–' acts as a Boolean NOT operand)
*metaanaly**
*quantitative review**
*quantitative overview**

*quantitative synthes**
*review synthes**
*research synthes**
*systematic overview**
*systematic review**
quantitative pooling

synthes of evidence*
*evidence synthes**
combin + estimate* (title only)*
pooling (title only)
combin estimate**
combin of estimate**

ISI/EMBASE Stage 2 search strategy *(Table 26)*

All searches were done on the title, keyword and abstract fields unless otherwise indicated.

Other searches

At this point it was decided to stop database searching. Clearly other databases, such as those for specialist education and psychology literatures could have been searched but it was felt diminishing returns for the time and effort invested would have been obtained.

A much simplified search of EMBASE and ISI Science was done a number of times from winter 1996 through summer 1997 with the aim to retrieve references recently published.

Results

It was very difficult to locate group 3 references (applications with new methodology). Several of these were found, mostly through exploding reference lists and personal communication with other researchers in the field. It would seem that locating these through electronic databases alone would be almost impossible.

Reference

1. Deeks J, Glanville J, Sheldon T. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. Centre for Reviews and Dissemination, York. York Publishing Services Ltd, Report #4, 1996.

TABLE 26 *ISI/EMBASE stage 2 search strategy*

Term	Search term	Total ∩ Stage 1
Procedural methodology		
∇procedural methodology	*procedural methodology*	0
procedure/s	*procedur**	500
guidelines	*guide**	188
guidance	"	24
framework/s	*framework**	141
data extraction	*data extraction*	29
individual patient data	*individual patient data*	0
study level data	*study level data*	1
level of aggregate/ion	*level of aggregat**	
Publication bias		
publication bias	*publication bias*	32
literature bias	*literature bias*	0
reporting bias	*reporting bias*	0
Statistical issues		
∇statistical issues	*statistical issues*	2
∇statistical methodology	*statistic* method**	44
statistics	*statistic**	813
calculation/s	*calculation**	82
Fixed effect(s) approaches		
fixed effect/s	*fixed effect**	16
homogeneity	*homogen**	90
homogeneous	"	34
classical	*classical*	
Random effect(s) approaches		
random effect/s	*random effect**	29
heterogeneity	*heterogene**	172
heterogeneous	"	
Mixed effect approaches		
mixed effects	*mixed effect**	2
	'mixed-effect'*	2
Study quality		
study quality	*study quality*	28
quality of studies	*(quality of studies ∪ quality of study)*	10
Influence of specific studies		
influence of specific studies	*influence of specific studies*	0
influence of a specific study	*influence of a specific study*	0
study influence	*study influence*	0
deletion method/s	*deletion method**	0
sensitivity analysis	*sensitivity analys**	29
Meta-regression		
meta-regression	*'meta-regression'*	5
metaregression	*metaregression*	2
(meta regression	*meta regression*	2
regression	*regression*	234
explaining variation	*explaining variation*	0
covariates	*covariates*	16
Cross design synthesis		
cross design	*cross design**	4
	'cross-design'*	1
Confidence profiling approach		
confidence profiling	*confidence profil**	2
confidence profile	"	
Missing data		
missing data	*missing data*	5
missing value/s	*missing value**	0
incomplete data	*incomplete data*	1
Scales of measurement		
scales of measurement	*scale* of measurement**	0
measurement scale/s	*measurement scale**	2
effect/s measure/s	*effect measure**	8
survival data	*survival data*	7
survival analysis	*survival analysis*	14
survival measure/s	*survival measure**	1
ordinal data	*ordinal data*	0
ordinal outcome/s	*ordinal outcome**	0
ordinal measure/s	*ordinal measure**	0
continuous outcome	*continuous outcome*	0
number needed to treat	*number needed to treat*	17
economic	*economic*	69
∇cost effectiveness	*cost effective**	145
cost	*cost**	398
Reporting of the results		
reporting of the result/s	*report* of the result**	10
report/ing the result/s	*report* the result**	24
result/s reporting	*result* report**	9
Prediction		
predict/ion	*predict**	288
∇effect predict/ion	*effect predict**	0
Cumulative meta-analysis		
cumulative	*cumulative*	61
sequential	*sequential*	23
chronological	*chronological*	7
Multi-level modelling		
multi-level model/s	*'multi-level' model**	0
multi level model/s	*multi level model**	0
multilevel model/s	*multilevel model**	0
hierarchical model/s	*hierarchical model**	1
Empirical Bayes		
empirical bayes	*empirical bayes**	5
Full Bayes		
bayes*	*(bayes*)* NOT *(empirical)*	23
MCMC		
markov chain	*markov chain*	4
monte carlo	*monte carlo*	4
monte-carlo	*'monte-carlo'*	0
Gibbs sampling		
gibbs sampling	*gibbs sampl**	0
bugs	*bugs*	1
simulation methods	*simulation methods*	0

Key: ∇ *term is a subset of another term and not necessary for the search but included for interest*

Appendix 2

Relevant material located too late to include in the review

The following is a list of references that were located too late in the project to review in the main text. They have been separated from the main Bibliography to make it easier for the reader to identify papers of further interest not covered by the report.

Anonymous. Meta-analysis of drug abuse prevention programs. Proceedings of a meeting. July 26–27. *NIDA Res Monogr* 1993;**170**:1–252.

Bailar JC, 3rd. The promise and problems of meta-analysis (editorial; comment). *N Engl J Med* 1997;**337**:559–61.

Bangert-Drowns RL. Some limiting factors in meta-analysis. *NIDA Res Monogr* 1997;**170**:234–52.

Bayarri MJ, DeGroot M. A Bayesian view of weighted distributions and selection models. Technical Report #375; Department of Statistics, Carnagie Mellon University, 1986.

Bayarri MJ, DeGroot M. The analysis of published significant results. Technical Report #91-21; Department of Statistics, Carnagie Mellon University, 1991.

Beck CT. Use of meta-analysis as a teaching strategy in nursing research courses. *J Nurs Educ* 1997;**36**:87–90.

Berlin JA. Commentary: summary statistics of poor quality studies must be treated cautiously (comment). *BMJ* 1997;**314**:337.

Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trials data. *Online J Curr Clin Trials* 1994; Doc No 13.

Chow SC, Liu J. Meta-analysis for bioequivalence review. *J Biopharm Stat* 1997;**7**:97–111.

Cleary RJ. An application of Gibbs sampling to estimation in meta-analysis: accounting for publication bias. *Journal of Education and Behavioral Statistics* 1997;**22**:141–54.

Conn HO. Interpretation of data from multiple trials: a critical review (review). *J Int Med* 1997;**241**:177–83.

Cook RJ, Walter SD. A logistic model for trend in 2 x 2 x kappa tables with applications to meta-analyses. *Biometrics* 1997;**53**:352–7.

Coste J, Bouyer J, Job-Spira N. Meta-analysis – 'does one bad apple spoil the barrel?' (letter). *Fertil Steril* 1997;**67**:791–2.

Counsell C. Formulating the questions and locating the studies for inclusion in systematic reviews. *Ann Int Med* 1996;**127**:380–7.

Devine EC. Issues and challenges in coding interventions for meta-analysis of prevention research. *NIDA Res Monogr* 1997;**170**:130–46.

Egger E, ZellwegerZahner T, Schneider M, Junker C, Lengeler C. Language bias in randomised controlled trials published in English and German. *Lancet* 1997;**350**:326–9.

Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**:629–34.

Eysenck HJ. Meta-analysis of best-evidence synthesis? (review). *J Eval Clin Pract* 1995;**1**:29–36.

Frongillo E. Combining information using hierarchical models. PhD Dissertation. Biometrics Unit, Cornell University, Ithaca, NY, 1995.

Goodman C. Step 1: Specify the assessment problem. In: Literature searching and evidence interpretation for assessing health care practices. Stockholm, SBU: The Swedish Council on Technology Assessment in Health Care, 1993.

Hansen WB, Rose LA. Issues in classification in meta-analysis in substance abuse prevention research. *NIDA Res Monogr* 1997;**170**:183–201.

Heaney RP. Some questions about 'Epidemiologic association between dietary calcium intake and blood pressure: a meta-analysis of published data' (letter). *Am J Epidemiol* 1997;**145**:858–9.

Hedges LV. Improving meta-analysis for policy purposes. *NIDA Res Monogr* 1997;**170**:202–15.

Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. *Can Med Assoc J* 1997;**156**:1411–16.

Johnson BT, Carey MP, Muellerleile PA. Large trials vs meta-analysis of smaller trials (letter; comment). *JAMA* 1997;**277**:377–8.

Klebanoff MA, Levine RJ, Dersimonian R. Large trials vs meta-analysis of smaller trials [letter; comment]. *JAMA* 1997;**277**:376–8.

Lancaster T. Systematic reviews (letter). *Fam Pract* 1997;**14**:90.

Law S. Diary of a novice Cochranite. Random thoughts on the third annual Cochrane Collaboration Colloquium (editorial). *Can Fam Physician* 1997;**43**:401–2, 410–12.

Lelorier J, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med* 1997;**337**:536–42.

Lewis G, Churchill R, Hotopp M. Systematic reviews and meta-analysis (editorial). *Psychol Med* 1997;**27**:3–7.

Lipsey MW. Using linked meta-analysis to build policy models. *NIDA Res Monogr* 1997;**170**:216–33.

Mathew T, Sinha BK, Zhou L. Some statistical procedures for combining independent tests. *J Am Statist Assoc* 1993;**88**:912–19.

Matt GE. Drawing generalized causal inferences based on meta-analysis. *NIDA Res Monogr* 1997;**170**:165–82.

Moher D. Assessing the quality of randomized controlled trials: implications for the conduct of meta-analyses. NHS HTA 93/52/04, 1997.

Perry PD. Realities of the effect size calculation process: considerations for beginning meta-analysts. *NIDA Res Monogr* 1997;**170**:120–9.

Pigott TD. The application of maximum likelihood methods to missing data in meta-analysis. Dissertation, University of Chicago, 1992.

Rice JP. The role of meta-analysis in linkage studies of complex traits. *Am J Med Genet* 1997;**74**:112–14.

Silagy C. Systematic reviews (letter). *Fam Pract* 1997;**14**:90–1.

Smith BJ, Darzins PJ, Quinn M, Heller RF. Modern methods of searching the medical literature. *Med J Aust* 1992;**157**:603–11.

Steinberg KK, Smith SJ, Stroup DF, *et al.* Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *Am J Epidemiol* 1997;**145**:917–25.

Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997;**315**:640–5.

Teagarden JR. Meta-analysis: whither narrative review? *Pharmacotherapy* 1997;**9**:274–84.

Tramer MR, Reynolds DJM, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997;**315**:635–40.

Tweedie RL. Assessing sensitivity to data problems in epidemiological meta-analyses. Technical Report. Department of Statistics, Colorado State University, Fort Collins CO 80523, USA, 1997.

Bibliography

Below is the report bibliography. This contains all the 967 references, in some way, concerned with meta-analysis/systematic review methodology considered in the review. Additionally, see appendix 2 for references found too late for inclusion in the review.

Aalen, O. (1992) Modelling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability*;2:951–972.

Abel, UR, Edler, L. (1988) A pitfall in the meta-analysis of hazard ratios. *Controlled Clin Trials*;**9**:149–151.

Abram, SE, Hopwood, M. (1996) Can metaanalysis rescue knowledge from a sea of unintelligible data. *Regional Anesthesia*;21:514–516.

Abrams, KR, Jones, DR. (1995) Meta-analysis and the synthesis of evidence. *IMA J Math Appl Med Biol;***12**:297–313.

Abrams, KR, Sanso, B. (1995) Discrimination in meta-analysis – a Bayesian perspective. Technical Report #95-03, Department of Epidemiology and Public Health, University of Leicester.

Abrams, KR, Sanso, B. (1997) Approximate Bayesian inference in random effects meta-analysis. *Stat Med* 1998;**17**:201–218.

Abrams, KR, Hellmich, M, Jones, DR. (1997) Bayesian approach to health care evidence. Technical Report #9701. Department of Epidemiology and Public Health, University of Leicester.

Abramson, JH. (1990) Meta-analysis: a review of pros and cons. *Public Health Review*;**9**:149–151.

Abramson, NS, Adams, HP, Algra, A, Asplund, K, Barer, D, Barnett, HJM, *et al.* (1995) How good are volunteers at searching for published randomized controlled trials? *Fundamental and Clinical Pharmacology;***9**:384–386.

Adams, CE, Lefebvre, C, Chalmers, I. (1992) Difficulty with MEDLINE searches for randomised controlled trials. *Lancet;***340**:915–916.

Adams, CE, Power, A, Frederick, K, Lefebvre, C. (1994) An investigation of the adequacy of medline searches for randomized controlled trials (RCTs) of the effects of mental health-care. *Psychological Medicine;***24**:741–748.

Adamson, GD, Pasta, DJ. (1994) Surgical treatment of endometriosis-associated infertility: meta-analysis compared with survival analysis. *American Journal of Obstetrics and Gynecology;***171**:1488–1505.

Allison, DB, Gorman, BS. (1993) Calculating effect sizes for meta-analysis: the case of the single case (review). *Behaviour Research and Therapy;***31**:621–631.

Altman, LK. (1990) New method of analyzing health data stirs debate. New York Times B5–B9.

Altman, D, Chalmers, I. (1995) Systematic reviews. London: BMJ Publishing.

Altman, DG, Dore, CJ. (1990) Randomisation and baseline comparisons in clinical trials. *Lancet;***335**:149–53.

Altman, DG, Elbourne, D. (1988) Commentary: combining results from several clinical trials. *Br J Obstet Gynaecol*;**95**:1–2.

Andersen, JW, Harrington, D. (1992) Meta-analyses need new publication standards. *J Clin Oncol*;**10**:878–880.

Andrew, E, Anis, A, Chalmers, T, Cho, M, Clarke, M, Felson, D, *et al.* (1994) A proposal for structured reporting of randomized controlled trials. *JAMA;***272**:1926–31.

Anello, C, Fleiss, JL. (1995) Exploratory or analytic meta-analysis: should we distinguish between them? *J Clin Epidemiol*;**48**:109–116.

Anonymous. (1996) The Cochrane Library. Issue 1 edn, BMJ Publishing Group.

A'Hern, RP, Ebbs, SR, Baum, MB. (1988) Does chemotherapy improve survival in advanced breast cancer? A statistical overview. *Br J Cancer;***57**:615–618.

Anonymous. (1986) Conferences on interpretation of environmental data: II statistical issues in combining environmental studies. Washington DC: EPA. Conference Proceeding.

Anonymous. (1992) Hypertension study stands tall as a model of meta-analysis. Science Watch. 5: August.

Anonymous. (1995) Magnesium, myocardial infarction, meta-analysis and megatrials (review). *Drug Ther Bull;***33**:25–27.

Antman, EM, Lau, J, Kupelnick, B, Mosteller, F, Chalmers, TC. (1992) A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA*;**268**:240–248.

Antman, EM, Lau, J, Kupelnick, B, Mosteller, F, Chalmers, TC. (1993) Cumulative metaanalyses and the problem of multiple-drug effects – reply. *JAMA*;**269**:214.

Austenfeld, MS, Thompson, IM Jr, Middleton, RG. (1994) Meta-analysis of the literature: guideline development for prostate cancer treatment. American Urological Association Prostate Cancer Guideline Panel. *J Urol*;**152**:1866–1869.

Avant, K. (1993) Using a systematic literature review to develop a theoretical framework for research. *Vard Nord Utveckl Forsk*:15–17.

Bailar, JC, Mosteller, F. (1988) Guidelines for statistical reporting in articles for medical journals. *Annals of Internal Medicine;***18**:266–273.

Bailar, JC 3rd. (1995) The practice of meta-analysis (review). *J Clin Epidemiol;***48**:149–157.

Bailey, KR. (1987) Inter-study differences – how should they influence the interpretation and analysis of results. *Stat Med;***6**:351–360.

Bailey, KR. (1994) Generalizing the results of randomized clinical trials. *Controlled Clin Trials;***15**:15–23.

Balas, EA, Austin, SM, Ewigman, BG, Brown, GD, Mitchell, JA. (1995) Methods of randomized controlled clinical trials in health services research. *Med Care;*33:687–699.

Bangert-Drowns, RL. (1986) Review of developments in meta-analytic method. *Psychol Bull;***99**:388–399.

Barnard, GA. (1992) Statistics and OR – some needed interactions. *Journal Of The Operational Research Society;***43**:787–795.

Barratt, A, Irwig, L. (1994) Is cholesterol testing/treatment really beneficial – reply. *Medical Journal of Australia;*160:451.

Bartolucci, AA, Katholi, CR, Singh, KP, Alarcon, GS. (1994) Issues in meta-analysis: an overview (review). *Arthritis Care Res;***7**:156–160.

Basinski, A, Naylor, CD. (1991) Asprin and fibrinolysis in acute myocardial infarction: meta-analytic evidence for synergy. *J Clin Epidemiol;***44**:1085–1096.

Bausell, RB. (1993) After the meta-analytic revolution. *Evaluation and the Health Professions;***16**:3–12.

Beck, CT. (1995) Meta-analysis: overview and application to clinical nursing practice (review). *J Obstet Gynecol Neonatal Nurs;***24**:131–135.

Becker, BJ. (1993) The generalizability of empirical research results. In: Benbow, C, Lubinsky, D. (eds) From psychometrics to giftedness: Papers in honor of Julian Stanley.

Becker, BJ. (1994) Combining significance levels. In: Cooper H, Hedges LV. (eds) The handbook of research synthesis, pp. 215–230. New York: Russell Sage Foundation].

Becker, BJ, Schram, CM. (1994) examining explanatory models through research synthesis. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 357–382. New York: Russell Sage Foundation.

Begg, CB. (1985) A measure to aid in the interpretation of published clinical trials. *Statist Med;***4**:1–9.

Begg, CB. (1994) Publication bias. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 399–310. New York: Russell Sage Foundation.

Begg, CB. (1996) The role of meta-analysis in monitoring clinical trials. *Stat Med;***15**:1299–1306.

Begg, CB, Berlin, JA. (1987) Publication bias: a problem in interpreting medical data. Technical Report. 490Z, Dana-Farber Cancer Institute.

Begg, CB, Berlin, JA. (1988) Publication bias: a problem in interpreting medical data (with discussion). *J R Statist Soc A*;**151**:419–463.

Begg, CB, Berlin, JA. (1989) Publication bias and dissemination of clinical research. *J Natl Cancer Inst;***81**:107–115.

Begg, CB, Mazumdar, M. (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics;***50**:1088–1101.

Begg, CB, Pilote, L. (1991) A model for incorporating historical controls into a meta-analysis. *Biometrics;***47**:899–906.

Behar, D. (1992) FastPro – software for metaanalysis by the confidence profile method. *JAMA*;**268**:2109.

Beral, V. (1995) ''The practice of meta-analysis'': discussion. Meta-analysis of observational studies: a case study of work in progress. *J Clin Epidemiol;***48**:165–166.

Berger, JO, Mortera, J. (1991) Interpreting the stars in precise hypothesis-testing. *International Statistical Review;***59**:337–353.

Bergmann, JF, Chassany, O, Segrestaa, JM, Caulin, C. (1994) Scoring systems for the methodological evaluation of therapeutic trials – an overview. *Therapie;***50**:181–184.

Bergstrom, R. (1994) The use and misuse of meta-analysis in clinical medicine. *J Intern Med;***236**:3–6.

Berkey, CS, Antczak-Bouckoms, A, Hoaglin, DC, Mosteller, F, Pihlstrom, BL. (1995) Multiple-outcomes meta-analysis of treatments for periodontal disease. *J Dent Res;*74:1030–1039.

Berkey, CS, Hoaglin, DC, Mosteller, F, Colditz, GA. (1995) A random-effects regression model for meta-analysis. *Stat Med;***14**:395–411.

Berkey, CS, Anderson, JJ, Hoaglin, DC. (1996) Multiple-outcome meta-analysis of clinical trials. *Stat Med;***15**:537–557.

Berkey, CS, Mosteller, F, Lau, J, Antman, EM. (1996) Uncertainty of the time of first significance in random effects cumulative metaanalysis. *Controlled Clin Trials;***17**:357–371.

Berlin, JA. (1992) Will publication bias vanish in an age of online journals. *Online Journal of Current Clinical Trials* doc # 12.

Berlin, JA. (1995) Invited commentary: benefits of heterogeneity in meta- analysis of data from epidemiologic studies. *Am J Epidemiol;***142**:383–387.

Berlin, JA. (1997) Commentary: summary statistics of poor quality studies must be treated cautiously. *BMJ;***314**:337.

Berlin, JA, Colditz, GA. (1990) A meta-analysis of physical activity in the prevention of coronary heart disease. *Am J Epidemiol*;**132**:612–628.

Berlin, JA, Begg, CB, Louis, TA. (1987) A method for assessing the magnitude of publication bias in a sample of published clinical trials. 518Z, Boston: Dana-Farber Cancer Institute.

Berlin, JA, Begg, CB, Louis, TA. (1989) An assessment of publication bias using a sample of published clinical trials. *J Am Statist Assoc*;**84**:381–392.

Berlin, JA, Laird, NM, Sacks, HS, Chalmers, TC. (1989) A comparison of statistical methods for combining event rates from clinical trials. *Stat Med;***8**:141–151.

Berlin, JA, Goodman, SN, Fletcher, SW. (1993). An instrument for assessing the quality of reporting of clinical research. Chicargo, Ill. Read before the Second International Congress on Peer Review in Biomedical Publication.

Berlin, JA, Longnecker, MP, Greenland, S. (1993) Meta-analysis of epidemiologic dose-response data. *Epidemiology*;**4**:218–228.

Bernstein, F. (1988) The retrieval of randomized clinical trials in liver diseases from the medical literature: manual versus MEDLARS searches. *Controlled Clin Trials*;**9**:23–31.

Bero, LA, Glantz, SA, Rennie, D. (1994) Publication bias and public health policy on environmental tobacco smoke. *JAMA*;**272**:133–136.

Berry, DA. (1991) Bayesian methods in Phase III trials. *Drug Information Journal;***25**:345–368.

Besag, J. (1994) Some applications of Markov-chain Monte-Carlo in Bayesian inference. In: Appleton, D.R. (ed.) Aspects of medical statistics, London: Chapman and Hall.

Besag, J, Higdon, D, Mengersen, K. (1996) Meta-analysis via MCMC. In preparation.

Bhansali, MS, Vaidya, JS, Bhatt, RG, Patil, PK, Badwe, RA, Desai, PB. (1996) Chemotherapy for carcinoma of the esophagus – a comparison of evidence from metaanalyses of randomized trials and of historical control studies. *Annals of Oncology;***7**:355–359.

Biggerstaff, BJ. (1997) Confidence intervals in the one-way random effects model for meta-analytic applications. Technical Report. Department of Statistics, Colorado State University.

Biggerstaff, BJ, Tweedie, RL. (1997) Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med*;**16**:753–768.

Biggerstaff, BJ, Tweedie, RL, Mengersen, KL. (1994) Passive smoking in the workplace: classical and Bayesian meta-analyses. *Int Arch Occup Environ Health*;**66**:269–277.

Birge, RT. (1932) The calculation of errors by the method of least squares. *Phys Rev*;**16**:1–32.

Black, N. (1994) Experimental and observational methods of evaluation. *BMJ*;**309**:540.

Blair, A, Burg, J, Foran, J, Gibb, H, Greenland, S, Morris, R, *et al.* (1995) Guidelines for application of meta-analysis in environmental epidemiology. ISLI Risk Science Institute. *Regul Toxicol Pharmacol*;**22**:189–197.

Bland, CJ, Meurer, LN, Maldonado, G. (1995) A systematic approach to conducting a non-statistical meta-analysis of research literature. *Acad Med*;**70**:642–653.

BoadaJuarez, J, delCampo, RBF. (1996) Meta-analysis v1.0: a computer program for calculating meta-analysis units. *Methods Find Exp Clin Pharmacol;*18:109.

Bobbio, M, Demichelis, B, Giustetto, G. (1994) Completeness of reporting trial results: effect on physicians' willingness to prescribe. *Lancet;***343**:1209–1211.

Bochmann, F, Petermann, F. (1994) [Psychiatric and clinico-psychologic epidemiology: critique of methods and systematization approach] (review) [German]. *Z Klin Psychol Psychopathol Psychother;***42**:1–15.

Boden, WE. (1992) Meta-analysis in clinical trials reporting: has a tool become a weapon? [editorial]. *Am J Cardiol;***69**:681–686.

Boers, M, Guyatt, GH, Oxman, AD, Felson, DT, Anderson, JJ, Meenan, RF. (1991) Combined effect size: comment on the metaanalysis of second-line drugs in rheumatoid arthritis (1). *Arthritis and Rheumatism;***34**:1342–1343.

Boissel, JP. (1993) Ad hoc working party of the international collaborative group on clinical trials registries. The position paper and consensus recommendation on clinical trial registries. *Clinical Trials and Meta-Analysis;***28**:255–266.

Boissel, JP. (1994) [Advantages of meta-analysis of clinical trials]. [French]. *Therapie;***49**:161–164.

Boissel, JP, Haugh, MC. (1993) The iceberg phenomenon and publication bias: the editors' fault? *Clinical Trials and Meta-Analysis;*28:309–315.

Boissel, JP, Blanchard, J, Panak, E, Peyrieux, JC, Sacks, H. (1989) Considerations for the meta-analysis of randomized clinical trials: summary of a panel discussion. *Controlled Clin Trials;*10:254–281.

Bollini, P, Garcia Rodriguez, LA, Perez Gutthann, S, Walker, AM. (1992) The impact of research quality and study design on epidemiologic estimates of the effect of nonsteroidal anti-inflammatory drugs on upper gastrointestinal tract disease. *Arch Intern Med;***152**:1289–1295.

Borzak, S, Ridker, PM. (1995) Discordance between meta-analyses and large-scale randomized, controlled trials - examples from the management of acute myocardial infarction. *Annals of Internal Medicine;***123**:873–877.

Borzak, S, Ridker, PM. (1995) Discordance between metaanalyses and large-scale randomized, controlled trials – examples from the management of acute myocardial– infarction. *Annals of Internal Medicine;***123**:873–877.

Borzak, S, Rosman, H, Antman, EM, Lau, J, Kupelnick, B, Mosteller, F, Chalmers, TC. (1993) Cumulative meta-analyses and the problem of multiple drug effects (4). *JAMA;*269:214.

Bracken, MB. (1992) Statistical methods for analysis of effects of treatment in overviews of randomized trials. In: Sinclair, JC, Bracken, MB. (eds) Effective care of the newborn infant, Oxford: Oxford University Press.

Brand, R. (1994) Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials – reply. *Stat Med;*11:295–296.

Brand, R, Kragt, H. (1992) Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med;*11:2077–2082.

Breipohl, AM, Albrecht, P, Allan, R, Asgarpoor, S, Bhavaraju, M, Billinton, R, *et al.* (1995) Pooling generating-unit data for improved estimates of performance indexes. *Ieee Transactions On Power Systems;*10:1912–1918.

Brennan, P, Croft, P. (1994) Interpreting the results of observational research: chance is not such a fine thing. *BMJ*;309:727–730.

Breslow, NE, Clayton, DG. (1993) Approximate inference in generalized linear mixed models. *J Am Statist Assoc;*88:9–25.

Breslow, NE, Day, NE. (1980) Classical methods of analysis of grouped data. In: Anonymous analysis of case-control studies. Statistical methods in cancer research (Vol 1), pp. 122–159. Lyon: International Agency for Research on Cancer.

Brown, BW Jr. (1993) Meta-analysis and patient-controlled analgesia (editorial; comment). *Journal of Clinical Anesthesia;*5:179–181.

Bruno, JE, Ellett, FS. (1988) A core-analysis of meta-analysis. *Quality and Quantity;*22:111–126.

Bryant, FB, Wortman, PM. (1984) Methodological issues in the meta-analysis of quasi-experiments. *New Directions for Program Evaluation;*24:5–24.

Buck, D, Sutton, M. (1994) Interpreting results of observational research – Problem with confounders can be tackled. *BMJ*;309:1439.

Buffler, PA. (1989) The evaluation of negative epidemiologic studies: the importance of all available evidence in risk characterization. *Regulatory Toxicology and Pharmacology;*9:34–43.

Bulpitt, CJ. (1988) Meta-analysis. *Lancet;*ii:93–94.

Bushman, BJ. (1994) Vote-counting procedures in meta-analysis. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 193–214. New York: Russell Sage Foundation.

Buyse, M. (1993) Meta-analyses, use and misuse (letter; comment). *J Clin Oncol;*11:382.

Buyse, M, Ryan, LM. (1987) Issues of efficiency in combining proportions of deaths from several clinical-trials. *Stat Med;*6:565–576.

Byar, DP. (1980) Why data bases should not replace randomized clinical trials. *Biometrics;*36:337–342.

Byar, DP. (1991) Problems with using observational databases to compare treatments. *Stat Med;*10:663–666.

Canadian Task Force for the Periodic Health Examination. (1979) The periodic health association. *Canadian Medical Association Journal;*121:1193–1197.

Canner, PL. (1987) An overview of 6 clinical-trials of aspirin in coronary heart-disease. *Stat Med;*6:255.

Cappelleri, JC, Ioannidis, JPA, Schmid, CH. (1997) Large trials vs meta-analysis of smaller trials – reply. *JAMA;*277:377–378.

Cappelleri, JC, Ioannidis, JPA, Schmid, CH, Deferranti, SD, Aubert, M, Chalmers, TC, *et al* (1996) Large trials vs metaanalysis of smaller trials – how do their results compare. *JAMA;*276:1332–1338.

Carlin, JB. (1992) Meta-analysis for 2 x 2 tables: a Bayesian approach. *Stat Med;*11:141–158.

Carlton, PL, Strawderman, W.E. (1996) Evaluating cumulated research .1. The inadequacy of traditional methods. *Biological Psychiatry;*39:65–72.

Carroll, RJ, Stefanski, LA. (1994) Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Stat Med*;13:1265–1282.

Carroll, RJ, Simpson, DG, Zhou, H, Guth, D. (1994) Stratified ordinal regression: a tool for combining information from disparate toxicological studies. Technical Report #26, Research Triangle Park, NC: National Institute of Statistical Sciences.

Center, BA, Skiba, RJ, Casey, A. (1985) A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education;*19:387–400.

Chalmers, I. (1990) Underreporting research is scientific misconduct. *JAMA*;263:1405–1408.

Chalmers, I. (1991) Can meta-analyses be trusted? *Lancet*;338:1464–1465.

Chalmers, I. (1991) Improving the quality and dissemination of reviews of clinical research. In: Lock, S. (ed.) The future of medical journals. In commemoration of 150 years of the *British Medical Journal.* pp. 127–146. London: BMJ Publishing Group.

Chalmers, TC. (1991) Problems induced by meta-analyses. *Stat Med;*10:971–980.

Chalmers, TC. (1993) Clinical trial quality needs to be improved to facilitate meta-analyses (editorial). *Online J Curr Clin Trials;*Doc, No 89:.

Chalmers, I. (1993) Publication bias. *Lancet;*342:1116.

Chalmers, I, Haynes, B. (1994) Reporting, updating, and correcting systematic reviews of the effects of health care. *BMJ*;309:862–865.

Chalmers, TC, Lau, J. (1993) Meta-analytic stimulus for changes in clinical trials (review). *Stat Methods Med Res;***2**:161–172.

Chalmers, TC, Lau, J. (1996) Changes in clinical trials mandated by the advent of meta-analysis. *Stat Med;***15**:1263–1268.

Chalmers, TC, Matta, RJ, Smith, H, Kunzler, A. (1977) Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med;***297**:1091–1096.

Chalmers, TC, Smith H Jr, Blackburn, B, Silverman, B, Schroeder, B, Reitman, D, *et al* (1981) A method for assessing the quality of a randomized control trial. *Controlled Clin Trials;***2**:31–49.

Chalmers, TC, Celano, P, Sacks, HS, Smith, H. (1983) Bias in treatment assignment in controlled clinical-trials. *N Engl J Med;***309**:1358–1361.

Chalmers, I, Heatherington, J, Newdick, M, Mutch, L, Grant, A, Enkin, M, *et al.* (1986) The Oxford database of perinatal trials – developing a registrar of published reports of controlled trials. *Controlled Clin Trials;***7**:306–324.

Chalmers, TC, Levin, H, Sacks, HS, Reitman, D, Berrier, J, Nagalingam, R. (1987) Meta-analysis of clinical trials as a scientific discipline. I: control of bias and comparison with large co-operative trials. *Stat Med;***6**:315–325.

Chalmers, TC, Berrier, J, Sacks, HS, Levin, H, Reitman, D, Nagalingam, R. (1987) Meta-analysis of clinical trials as a scientific discipline. II: replicate variability and comparison of studies that agree and disagree. *Stat Med;***6**:733–744.

Chalmers, TC, Berrier, J, Hewitt, P, Reitman, D, Nagalingam, R, Sacks, H. (1988) Logistics of determining rates of rare side effects by meta-analysis. Paper presented at the 9th Annual Meeting of the Society for Clinical Trials, 22–25 May, San Diego, 1988.

Chalmers, I, Heatherington, J, Elbourne, D, Keirse, MJNC, Enkin, M. (1989) Materials and methods used in synthesising evidence to evaluate the effects of care during pregnancy and childbirth. In: Chalmers, I, Enkin, M, Keirse, M. (eds) Effective care in pregnancy and childbirth, pp. 39–65. Oxford: Oxford University Press.

Chalmers, TC, Frank, CS, Reitman, D. (1990) Minimizing the three stages of publication bias. *JAMA;***263**:1392–1395.

Chalmers, I, Dickersin, K, Chalmers, TC. (1992) Getting to grips with Archie Cochrane's agenda: a register of all randomized controlled trials. *BMJ;***305**:786–688.

Chalmers, I, Sandercock, P, Wennberg, J. (1993) The Cochrane collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences;***703**:156–165.

Champney, TF. (1983) Adjustments for selection: publication bias in quantitative research synthesis. University of Chicargo.

Chan, SS, Sacks, HS, Chalmers, TC. (1982) The epidemiology of unpublished randomized control trials. *Clinical Research;***30**:234A.

Charlton, BG. (1994) Practice guidelines and practical judgement – the role of mega trials, meta analysis and consensus. *Br J Gen Pract;***44**:290–291.

Charlton, BG. (1996) The uses and abuses of meta-analysis. *Family Practice;*13:397–401.

Chassany, O, Bergmann, JF, Segrestaa, JM, Caulin, C. (1994) The uncertainty of the prescriber faced with meta-analysis: the gastro-intestinal toxicity of steroids. *Therapie;***50**:169–170.

Checkoway, H. (1991) Data pooling in occupational studies. *J Occup Med;***33**:1257–1260.

Chelimsky, E. (1994) Politics, policy and research synthesis. Keynote address before national conference on Research Synthesis. In: Anonymous Washington DC: Russell Sage Foundation.

Chelimsky, E, Silberman, G, Droitcour, J. (1993) Cross design synthesis (letter; comment). *Lancet;***341**:498.

Chêne, G, Thompson, SG. (1996) Methods for summarizing the risk associations of quantitative variables in epidemiologic studies in a consistent form. *Am J Epidemiol;***144**:610–621.

Chernoff, MC, Wang, MZ, Anderson, JJ, Felson, DT. (1995) Problems and suggested solutions in creating an archive of clinical trials data to permit later meta-analysis: An example of methotrexate trials in rheumatoid arthritis. *Controlled Clin Trials;***16**:42–355.

Chernoff, MC, Wang, MZ, Anderson, JJ, Felson, DT. (1995) Problems and suggested solutions in creating an archive of clinical-trials data to permit later metaanalysis – an example of methotrexate trials in rheumatoid-arthritis. *Controlled Clin Trials;***16**:342–355.

Chiou, P. (1995) A pooling procedure for the exponential-distribution. *Communications In Statistics – Theory And Methods;***24**:445–458.

Cho, MK, Bero, LA. (1994) Instruments for assessing the quality of drug studies published in the medical literature. *JAMA;***272**:101–104.

Cho, MK, Bero, LA. (1996) The quality of drug studies published in symposium proceedings. *Annals of Internal Medicine;***124**:485.

Clarke, M. (1993) Searching Medline for randomised trials. *BMJ;***307**:565.

Clarke, MJ, Stewart, LA. (1994) Systematic reviews – obtaining data from randomized controlled trials – how much do we need for reliable and informative meta-analyses. *BMJ;***309**:1007–1010.

Cochran, WG. (1937) Problems arising in the analysis of a series of similar experiments. *J R Statist Soc;***4**:102–118.

Cochran, WG. (1954) The combination of estimates from different experiments. *Biometrics;*10:101–129.

Colditz, G.A, Burdick, E, Mosteller, F. (1995) Heterogeneity in meta-analysis of data from epidemiologic studies: commentary. *Am J Epidemiol;***142**:371–382.

Colditz, GA, Miller, JN, Mosteller, F. (1988) Measuring gain in the evaluation of medical technology: the probability of a better outcome. *International Journal of Technology;***4**:637–642.

Colditz, GA, Miller, JN, Mosteller, F. (1989) How study design affects outcomes in comparisons of therapy. I: Medical. *Stat Med;***8**:441–454.

Cole, BF, Gelber, RD, Goldhirsch, A. (1995) A quality-adjusted survival meta-analysis of adjuvant chemotherapy for premenopausal breast cancer. International Breast Cancer Study Group. *Stat Med;***14**:1771–1784.

Collette, L, Suciu, S, Bijnens, L, Sylvester, R. (1997) Including literature data in individual patient data meta-analyses for time-to-event endpoints. *Controlled Clin Trials;***18**:188S.

Collins, R, Gray, R, Godwin, J, Peto, R. (1987) Avoidance of large biases and large random errors in the assessment of moderate treatment effects – the need for systematic overviews. *Stat Med;***6**:245–254.

Colton, T, Freedman, LS, Johnson, AL. (1987) Proceedings of the workshop of methodologic issues in overviews of randomized clinical trials, May 1986. *Stat Med;***6**:217–410.

Consonni, G, Veronese, P. (1993) A Bayesian method for combining results from several binomial experiments. Technical Report: Studi Statistici 40, L. Bocconi University, Milan, Italy.

Conti, CR. (1993) Clinical decision making using cumulative meta-analysis [editorial]. *Clin Cardiol;***16**:167–168.

Cook, D. (1995) "Cumulative meta-analysis of clinical trials: Builds evidence for exemplary medical care": Discussion. *J Clin Epidemiol;***48**:59–60.

Cook, D. (1997) Systematic reviews: the case for rigorous methods and rigorous reporting [editorial; comment]. *Can J Anaesth;***44**:350–353.

Cook, DJ, Guyatt, GH, Laupacis, A, Sackett, DL. (1992) Rules of evidence and clinical recommendations in the use of antithrombotic agents. Antithrombotic Therapy Consensus Conference. *Chest;***102**:305S–311S.

Cook, DJ, Guyatt, GH, Ryan, G, Clifton, J, Buckingham, L, Willan, A, Mcllroy, W, Oxman, AD. (1993) Should unpublished data be included in metaanalyses – current convictions and controversies. *JAMA;*269:2749–2753.

Cook, DJ, Mulrow, CD, Haynes, RB. (1997) Systematic reviews: synthesis of best evidence for clinical decisions. *Annals of Internal Medicine;***126**:376–380.

Cook, DJ, Reeve, BK, Guyatt, GH, Heyland, DK, Griffith, LE, Buckingham, L, Tryba, M. (1996) Stress-ulcer prophylaxis in critically ill patients – resolving discordant meta-analyses. *JAMA;***275**:308–314.

Cook, DJ, Sackett, DL, Spitzer, WO. (1995) Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-analysis (review). *J Clin Epidemiol;***48**:167–171.

Cook, RJ, Sackett, DL. (1995) The number needed to treat: a clinically useful measure of treatment effect. *BMJ;***310**:452–454.

Cook, TD, Campbell, DT. (1979) Quasi-experimentation: design & analysis issues for field settings. Boston: Houghton Mifflin.

Cook, TD, Leviton, LC. (1980) Reviewing the literature: a comparison of traditional methods with meta-analysis. *Journal of Personality;***48**:448–471.

Cooley, PC, Myers, LE, Hamill, DN. (1996) A meta-analysis of estimates of the AIDS incubation distribution. *European Journal Of Epidemiology;***12**:229–235.

Cooper, GS, Zangwill, L. (1992) An analysis of the quality of research reports in the *Journal of General Internal Medicine Journal of General Internal Medicine;***45**:255–265.

Cooper, H, Hedges, LV. (1994) Research synthesis as a scientific enterprise. In: Cooper, H, Hedges, L.V. (eds) The handbook of research synthesis, pp. 3–14. New York: Russell Sage Foundation.

Cooper, H, Hedges, LV. (1994) Potentials and limitations of research synthesis. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 521–530. New York: Russell Sage Foundation.

Cooper, H, Ribble, RG. (1989) Influences on the outcome of literature searches for integrative reviews. *Knowledge;***10**:179–201.

Cooper, HM. (1982) Scientific guidelines for conducting intergrative research reviews. *Review of Educational Research;***52**:291–302.

Cooper, HM. (1985) Literature searching strategies of integrative research reviewers. *American Psychologist;***40**:1267–1269.

Cooper, HM. (1987) Literature searching strategies of integrative research reviews: a first survey. *Knowledge: Creation, Diffusion, Utilization;***8**:372–383.

Cooper, HM. (1989) Integrating research: A guide for literature reviews, 2nd edn. Newbury Park, CA: Sage.

Cooper, HM. (1996) Organizing knowledge synthesis: a taxonomy of literature reviews. Knowledge in Society 1, 104–126.

Cooper, HM, Rosenthal, R. (1980) Statistical versus traditional procedures for summarizing research findings. *Psychol Bull;***87**:442–449.

Coplen, SE, Antman, EM, Berlin, JA, Hewitt, P, Chalmers, TC. (1990) Efficacy and safety of quinidine therapy for maintenance of sinus rhythm after cardioversion: a meta-analysis of randomized control trials. *Circulation;***82**:1106–1116.

Corcoran, KJ. (1985) Aggregating the idiographic data of single subject research. *Social Work Research and Abstracts;***21**:9–12.

Couchoud, C, Laville, M, Boissel, JP. (1994) Treatnent of membranous nephropathy by immunosuppressive agents of the difficulty of doing a meta-analysis in nephrology. *Therapie;***50**:171–173.

Counsell, C, Fraser, H. (1995) Identifying relevant studies for systematic reviews. *BMJ;***310**:126.

Counsell, CE, Clarke, MJ, Slattery, J, Sandercock, PAG. (1994) The miracle of DICE therapy for acute stroke: Fact or fictional product of subgroup analysis? *BMJ;***309**:1677–1681.

Counsell, CE, Fraser, H, Sandercock, PAG. (1994) Archie Cochrane's challenge: can periodically updated reviews of all randomised controlled trials relevant to neurology and neurosurgery be produced? *J Neurol Neurosurg Psychiatry;***57**:529–533.

Coursol, A, Wagner, EE. (1986) Effect of positive findings on submission and acceptance rates: a note on meta-analysis bias. *Professional Psychology;***17**:136–137.

Cox, DR, Snell, EJ. (1989) Analysis of binary data, London: Chapman and Hall.

Cox, LH, Piegorsch, WW. (1994) Combining environmental information: Environmetric research in ecological monitoring, epidemiology, toxicology, and environmental data reporting. #12, Research Triangle Park, NC: National Institute of Statistical Sciences.

Cox, LH, Piegorsch, WW. (1996) Combining environmental information I: Environmental monitoring, measurement and assessment. *Environmetrics;***7**:299–308.

Craun, G, Calderon, R, Frost, F. (1997) Article sparks debate over meta-analysis – reply. Journal of the American Water Works Association 89, 4.

Csada, RD, James, PC, Espie, RHM. (1996) The file drawer problem of nonsignificant results – does it apply to biological-research. *Oikos;***76**:591–593.

Curlette, WL, Canella, KS. (1985) Going beyond the narrative summarization of research findings: the meta-analysis approach. *Res Nurs Health;***8**:293–301.

D'Agostino, RB, Weintraub, M. (1995) Meta-analysis: a method for synthesizing research (review). *Clin Pharmacol Ther;***58**:605–616.

Daniels, MJ, Hughes, MD. (1997) Meta-analysis for the evaluation of potential surrogate markers. *Stat Med*;In press.

Davidson, D. (1977) The effect of individual differences of cognitive style on judgments of document relevance. *Journal of the American Society for Information Science;***28**:273–284.

Davidson, RA. (1986) Source of funding and outcome of clinical trials. *J Gen Intern Med;***1**:155–158.

Dawes, RM, Landman, J, Williams, M. (1984) Discussion on meta-analysis and selective publication bias. *Am Psychol;***39**:75–78.

Dawid, AP, Dickey, JM. (1977) Likelihood and Bayesian inference from selectively reported data. *J Am Statist Assoc;***72**:845–850.

De Oliveira, IR, Dardennes, RM, Amorim, ES, Diquet, B, de Sena, EP, Moreira, EC, *et al.* (1995) Is there a relationship between antipsychotic blood levels and their clinical efficacy? An analysis of studies design and methodology. *Fundam Clin Pharmacol;***9**:488–502.

Dear, KBG. (1994) Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics;***50**:989–1002.

Dear, KBG, Begg, CB. (1992) An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science;*7:237–245.

Deeks, J, Glanville, J, Sheldon, T. (1996) Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. #4, Centre for Reviews and Dissemination, York: York Publishing Services Ltd.

Deeks, JJ, Altman, DG, Dooley, G, Sackett, DLS. (1997) Choosing an appropriate dichotomous effect measure for meta-analysis: empirical evidence of the appropriateness of the odds ratio and relative risk. *Controlled Clin Trials;***18**:84s–85s.

Del Junco, DJ, Annegers, JF. (1992) Efficacy of immunotherapy for unexplained recurrent spontaneous abortion: the role of guidelines and meta analysis [editorial; comment]. *Am J Reprod Immunol;***27**:94–96.

Delahaye, F. (1994) A new column: meta-analysis on Medline. *Clin Trial Meta-Anal;***29**:191–246.

Delahaye, F, Landrivon, G, Ecochard, R, Colin, C. (1991) Meta-analysis. *Health Policy;***19**:185–196.

Delgado Rodriguez, M, Sillero Arenas, M, Galvez Vargas, R. (1992) [Meta-analysis in epidemiology (2): quantitative methods] (review) [Spanish]. *Gac Sanit;***6**:30–39.

Delgado-Rodriguez, M, Sillero Arenas, M. (1995) [Inclusion of research quality in meta-analyses] (review) [Spanish]. *Gac Sanit;***9**:265–272.

Demets, DL. (1987) Methods for combining randomised clinical trials: strengths and limitations. *Stat Med;***6**:341–348.

Department of Clinical Epidemiology and Biostatistics, MM, Health Sciences Centre (1981) How to read clinical journals, IV: to determine etiology or causation. *Canadian Medical Association;***124**:985–990.

Dersimonian, R. (1996) Meta-analysis in the design and monitoring of clinical trials. *Stat Med;***15**:1237–1252.

Dersimonian, R, Charette, LJ, McPeek, B, Mosteller, F. (1982) Reporting on methods in clinical trials. *N Engl J Med;***306**:1332–1337.

Dersimonian, R, Laird, N. (1986) Meta-analysis in clinical trials. *Controlled Clin Trials;*7:177–188.

Detsky, AS, Naylor, CD, ORourke, K, McGeer, AJ, LAbbe, KA, O'Rourke, K, *et al.* (1992) Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol;***45**:255–265.

Devries, SO, Hunink, MGM, Polak, JF. (1996) Summary receiver operating characteristic curves as a technique for metaanalysis of the diagnostic performance of duplex ultrasonography in peripheral arterial-disease. *Academic Radiology;***3**:361–369.

Dickersin, K. (1987) Reference bias in reports of drug trials. *BMJ;***295**:1066–1067.

Dickersin, K. (1988) Report from the panel on the case for registers of clinical trials at the Eighth Annual Meeting of the Society for Clinical Trials. *Controlled Clin Trials;***9**:76–81.

Dickersin, K. (1990) The existence of publication bias and risk factors for its occurrence. *JAMA;***263**:1385–1389.

Dickersin, K. (1992) Keeping posted. Why register clinical trials? –revisited. *Controlled Clin Trials;***13**:170–177.

Dickersin, K. (1994) Research registers. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 71–84. New York: Russell Sage Foundation.

Dickersin, K, Berlin, JA. (1992) Meta-analysis: state-of-the-science (review). *Epidemiol Rev;***14**:154–176.

Dickersin, K, Chan, S, Chalmers, TC, Sacks, HS, Smith, HJ. (1987) Publication bias and clinical trials. *Controlled Clin Trials;***8**:343–353.

Dickersin, K, Hewitt, P, Mutch, L, Chalmers, I, Chalmers, TC. (1985) Pursuing the literature: comparison of MEDLINE searching with a perinatal trials database. *Controlled Clin Trials;***6**:306–317.

Dickersin, K, Higgins, K, Meinert, CL. (1990) Identification of meta-analyses: The need for standard terminology. *Controlled Clin Trials;***11**:52–66.

Dickersin, K, Min, YI. (1993) NIH clinical trials and publication bias. *Online J Curr Clin Trials;*Doc, No 50:.

Dickersin, K, Min, YI, Meinert, CL. (1992) Factors influencing publication of research results: follow-up of applications submitted to two ionstitutional review boards. *JAMA;***263**:374–378.

Dickersin, K, Min, YI, Orza, M, Lucey, J, Johnson, K, Clarke, J, *et al.* (1993) Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences;***703**:135–148.

Dickersin, K, Scherer, R, Lefebvre, C. (1994) Systematic reviews – identifying relevant studies for systematic reviews. *BMJ;***309**:1286–1291.

Doll, R. (1994) The use of meta-analysis in epidemiology: Diet and cancers of the breast and colon. *Nutr Rev;***52**:233–237.

Domanski, MJ, Friedman, LM. (1994) Relative role of meta-analysis and randomized controlled trials in the assessment of medical therapies. *Am J Cardiol;***74**:395–396.

Dopfmer, S, Guggenmoos-Holzmann, I. (1997) [Meta-analysis] (review) [German]. *Dtsch Med Wochenschr;***122**:589–593.

Droitcour, J, Silberman, G, Chelimsky, E. (1993) Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care;***9**:440–449.

Dubey, SD. (1988) Regulatory considerations on meta-analysis, dentifrice studies and multicenter trials. In: Anonymous Proceedings of the Biopharmaceutical Section, pp. 18–27. Alexandria, Virginia: American Statistical Association.

Dubois, D, Prade, H. (1992) On the relevance of nonstandard theories of uncertainty in modeling and pooling expert opinions. *Reliability Engineering & System Safety;***36**:95–107.

Duffy, SW, Rohan, TE, Altman, DG. (1989) A method for combining matched and unmatched binary data. Application to randomized, controlled trials of photocoagulation in the treatment of diabetic retinopathy. *Am J Epidemiol;***130**:371–378.

Duffy, SW, Rohan, TE, Altman, DG. (1990) Re: 'A method for combining matched and unmatched binary data: application to randomized, controlled trials of photocoagulation in the treatment of diabetic retinopathy' (I: Reply). *Am J Epidemiol;***132**:198–199.

Duley, L. (1996) Systematic reviews – what can they do for you. *J R Soc Med;***89**:242–244.

DuMouchel, W. (1989) Bayesian Metaanalysis. In: Berry, D.A. (Ed) Statistical Methodology in the Pharmaceutical Sciences, pp. 509–529. New York: Marcel Dekker.

DuMouchel, W. (1994) Predictive cross-validation in Bayesian meta-analysis. Draft Form.

DuMouchel, W. (1994) Hierarchical Bayes linear models for meta-analysis. 27, Research Triangle Park, NC: National Institute of Statistical Sciences.

DuMouchel, W. (1995) Meta-analysis for dose-response models [comment]. *Stat Med;***14**:679–685.

DuMouchel, W, Waternaux, C. (1992) Comment on 'Hierarchical models for combining information and for meta-analyses'. In: Bernardo, J.M, Berger, J.O, Dawid, AP, Smith, AFM. (eds) pp. 338–339. Oxford University Press.

Sall, J. (Ed.) Bayesian meta-analyses with general-purpose software. Research Triangle Park: NC: Interface Foundation. (1994).

DuMouchel, WH, Harris, JE. (1983) Bayes methods for combining the results of cancer studies in humans and other species (with comment). *J Am Statist Assoc;***78**:293–308.

Dunn, G, Sham, P, Hand, D. (1993) Statistics and the nature of depression. Psychological Medicine 156, 871–889.

Durlak, JA, Lipsey, MW. (1991) A practitioner's guide to meta-analysis. *American Journal of Community Psychology;***19**:291–332.

Dyer, AR. (1986) A method for combining results from several prospective epidemiological studies. *Stat Med;***5**:307–317.

Eagly, AH, Wood, W. (1994) Using research synthesis to plan future research. In: Cooper, H, Hedges, L.V. (eds) The handbook of research synthesis, pp. 485–502. New York: Russell Sage Foundation.

Earleywine, M. (1993) The file drawer problem in the meta-analysis of the subjective responses to alcohol (letter). *Am J Psychiatry;***150**:1435–1436.

Early Breast Cancer Trialists' Collaborative Group (1990) Statistical Methods. In: Anonymous Treatment of early breast cancer. Vol 1. Worldwide evidence 1985–1990. pp. 13–18. Oxford: Oxford University Press.

Easterbrook, PJ. (1992) Directory of registries of clinical trials. *Stat Med;*11:345–423.

Easterbrook, PJ, Berlin, JA, Gopalan, R, Matthews, DR. (1991) Publication bias in clinical research. *Lancet;***337**:867–872.

Eberly, LE, Casella, G. (1996) Estimating the number of unseen studies. BUM #1308–MA.

Eddy, DM. (1989) The confidence profile method: a Bayesian method for assessing health technologies. *Operations Research;***37**:210–228.

Eddy, DM. (1990) Anatomy of a decision. *JAMA;***263**:441–443.

FastPro: Software for MetaAnalysis by the Confidence Profile Method. Eddy, D.M, Hasselblad, V. (1992) San Diego, California: Academic Press Inc. 0–12–230621–X. IBM-PC. 3.5-inch disk.

Eddy, DM, Hasselblad, V, Shachter, R. (1990) An introduction to a Bayesian method for meta-analysis: the confidence profile method. *Med Decis Making;***10**:15–23.

Eddy, D.M, Hasselblad, V, Shachter, R. (1990) A Bayesian method for synthesizing evidence: The confidence profile method. *International Journal of Technology Assessment in Health Care;***6**:31–55.

Eddy, DM, Hasselblad, V, Shachter, R. (1992) Meta-analysis by the Confidence Profile Method, San Diego: Academic Press.

Editorial (1987) Whither meta-analysis. *Lancet;***1***: 897–898. (*Source clearly gave the wrong volume).

Effective Health Care (1992) Effective Health Care: Reviewing the Evidence. In: Anonymous Screening for Osteoporosis to Prevent Fractures, Leeds: University of Leeds.

Efron, B. (1994) Missing data, imputation, and the bootstrap. *J Am Statist Assoc;*89:463–479.

Efron, B. (1996) Empirical bayes methods for combining likelihoods. *J Am Statist Assoc;***91**:538–550.

Egger, M, Davey Smith, G, Song, F, Sheldon, TA. (1993) Making sense of meta-analysis. *Pharmacoepidemiology Drug Safety;***2**:65.

Egger, M, Smith, GD. (1995) Misleading meta-analysis [editorial]. *BMJ;***310**:752–754.

Egger, M, Smith, GD. (1995) Risks and benefits of treating mild hypertension: a misleading meta-analysis? *J Hypertension;***13**:813–815.

Einarson, TR, Arikian, SR, Shear, NH. (1994) Cost-effectiveness analysis for onychomycosis therapy in canada from a government perspective. *British Journal Of Dermatology;***130**:32–34.

Einarson, TR, McGhan, WF, Sabers, DL. (1985) Quantitative integration of independent research results. *Am J Hosp Pharm;***42**:1957–1964.

Eisenberg, M, Barry, C. (1988) Order effects: a study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science;***39**:293–300.

Elliott, AT. (1994) Meta-analysis [editorial]. Nucl. Med. Commun. 15, 218–220.

Elston, RC. (1991) On Fisher's method of combining *p*-values. *Biometrical Journal;***33**:339–345.

Emerson, JD. (1994) Combining estimates of the odds ratio: the state of the art (review). *Stat Methods Med Res;***3**:157–178.

Emerson, JD, Burdick, E, Hoaglin, DC, Mosteller, F, Chalmers, TC. (1990) An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clin Trials;***11**:339–352.

Emerson, JD, Hoaglin, DC, Mosteller, F. (1993) A modified random-effect procedure for combining risk difference in sets of 2x2 tables from clinical trials. *Journal of the Italian Statistical Society;***2**:269–290.

Emerson, JD, Hoaglin, DC, Mosteller, F. (1996) Simple robust procedures for combining risk differences in sets of 2x2 tables. *Stat Med;***15**:1465–1488.

Enas, GG, Goldstein, DJ. (1995) Defining, monitoring and combining safety information in clinical trials. *Stat Med;***14**:1099–1111.

Evans, S. (1996) Misleading meta-analysis – Statistician's comment. *BMJ;***312**:125.

Everitt, BS. (1993) Meta-analysis. *Statistical Methods in Medical Research;***2**:117–192.

Eysenck, HJ. (1994) Systematic reviews – Meta-analysis and its problems. *BMJ;***309**:789–792.

Fackler, ML, Huth, EJ, Pitkin, RM, Rennie, D, Begg, C, Greenland, S, *et al.* (1996) Checklist of information for inclusion in reports of clinical-trials. *Annals of Internal Medicine;***124**:741–743.

Fagard, RH, Staessen, JA, Thijs, L. (1996) Advantages and disadvantages of the metaanalysis approach. *Journal of Hypertension;***14**:S9–S12.

Fahey, T, Griffiths, S, Peters, TJ. (1995) Evidence based purchasing: understanding results of clinical trials and systematic reviews. *BMJ;***311**:1056–1060.

Fanning, J, Bennett, TZ, Hilgers, RD. (1992) Meta-analysis of cisplatin, doxorubicin, and cyclophosphamide versus cisplatin and cyclophosphamide chemotherapy of ovarian carcinoma. *Obstet Gynecol;*80:954–960.

Fanning, J, Hilgers, RD. (1993) Meta-analysis (letter). *American Journal of Obstetrics & Gynecology;***169**:236–237.

Farbey, R. (1993) Searching the literature. *BMJ;***307**:66.

Feder, PI, MorgensteinWagner, TB, Chou, YL, Lordo, RA. (1992) Statistical methods for the metaanalysis of multiple clinical studies. Abstracts Of Papers Of The American Chemical Society 204, 119–ENVR.

Feinstein, AR. (1985) Clinical Epidemiology: the architecture of clinical research, Philadelphia: Saunders.

Feinstein, AR. (1995) Meta-analysis: statistical alchemy for the 21st century (review). *J Clin Epidemiol;***48**:71–79.

Feinstein, AR. (1996) Meta-analysis and meta-analytic monitoring of clinical trials. *Stat Med;***15**:1273–1280.

Feldman, KA. (1971) Using the work of others: some observations on reviewing and integrating. *Sociology of Education;***4**:86–102.

Felson, DT. (1992) Bias in metaanalytic research. *J Clin Epidemiol;***45**:885–892.

Fidel, R. (1991) Searchers' selection of search keys: III. Searching styles. *Journal of the American Society for Information Science;***42**:515–527.

Finney, DJ. (1995) A statistician looks at meta-analysis (review). *J Clin Epidemiol;***48**:87–103.

Fisher, RA. (1932) Statistical Methods for Research Workers, 4th edn. London: Oliver and Boyd.

Fiske, DW. (1983) The meta-analytic revolution in outcome research. *J Consult Clin Psychol;***51**:65–70.

Flather, MD, Farkouh, ME, Yusuf, S. (1994) Meta-analysis in the evaluation of therapies. In: Julian, D, Braunwald, E. (eds) Management of Acute Myocardial Infarction, pp. 393–406. London: WB Saunders.

Fleiss, JL. (1981) Statistical methods for rates and proportions, 2nd edn. New York: Wiley.

Fleiss, JL. (1993) The statistical basis of meta-analysis (review). *Stat Methods Med Res;***2**:121–145.

Fleiss, JL. (1994) Measures of effect size for categorical data. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 245–260. New York: Russell Sage Foundation.

Fleiss, JL, Gross, AJ. (1991) Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol;***44**:127–139.

Follmann, D, Elliott, P, Suh, I, Cutler, J. (1992) Variance imputation for overviews of clinical trials with continuous response. *J Clin Epidemiol;***45**:769–773.

Fortin, PR, Lew, RA, Liang, MH, Wright, EA, Beckett, LA, Chalmers, TC, *et al.* (1995) Validation of a meta-analysis: the effects of fish oil in rheumatoid arthritis. *J Clin Epidemiol;***48**:1379–1390.

Fouque, D, Laville, M, Haugh, M, Boissel, JP. (1996) Systematic reviews and their roles in promoting evidence-based medicine in renal disease. *Nephrology Dialysis Transplantation;***11**:2398–2401.

Freedman, LS. (1994) Meta-analysis of animal experiments on dietary fat intake and mammary tumours. *Stat Med;***13**:709–718.

Freeman, CD, Strayer, AH. (1996) Mega-analysis of meta-analysis – an examination of metaanalysis with an emphasis on once-daily aminoglycoside comparative trials. *Pharmacotherapy;***16**:1093–1102.

Fricke, R, Treines, G. (1985) Einfuhrung in die Meta-analyse, Bern, Switzerland: Hans Huber.

Friedenreich, CM. (1993) Methods for pooled analyses of epidemiologic studies (review). *Epidemiology;***4**:295–302.

Friedenreich, CM, Brant, RF, Riboli, E. (1994) Influence of methodologic factors in a pooled analysis of 13 case-control studies of colorectal cancer and dietary fiber [published erratum appears in *Epidemiology* 1994 May;**5**(3):385] (review). *Epidemiology;***5**:66–79.

Friedman, HP, Goldberg, JD. (1996) Meta-analysis: an introduction and point of view. *Hepatology;***23**:917–928.

Friedman, MA. (1987) Potential pooling opportunities: cancer. *Stat Med;***6**:307–312.

Gail, M, Simon, R. (1985) Testing for qualitative interaction between treatment effects and patient subsets. *Biometrics;***41**:361–372.

Galbraith, RF. (1988) A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med;***7**:889–894.

Galbraith, RF. (1994) Some applications of radial plots. *J Am Statist Assoc;***89**:1232–1242.

Gart, JJ. (1992) Pooling 2x2 tables – asymptotic moments of estimators. *Journal Of The Royal Statistical Society Series B;***54**:531–539.

Gelber, RD, Coates, AS, Goldhirsch, A. (1992) Meta-analysis: the fashion of summing-up evidence. Part II: Interpretations and uses. *Ann Oncol;***3**:683–691.

Gelber, RD, Cole, BF, Gelber, S, Goldhirsch, A. (1995) Comparing treatments using quality-adjusted survival – the Q-TWIST method. *American Statistician;***49**:161–169.

Gelber, RD, Cole, BF, Goldhirsch, A, Rose, C, Fisher, B, Osborne, C.K, *et al.* (1996) Adjuvant chemotherapy plus tamoxifen compared with tamoxifen alone for postmenopausal breast cancer: meta-analysis of quality-adjusted survival. *Lancet;***347**:1066–1071.

Gelber, RD, Goldhirsch, A. (1986) The concept of an overview of cancer clinical trials with special emphasis on early breast cancer. *J Clin Oncol;***4**:1696–1703.

Gelber, RD, Goldhirsch, A. (1987) Interpretation of results from subset analyses within overviews of randomized clinical trials. *Stat Med;***6**:371–378.

Gelber, RD, Goldhirsch, A. (1987) The evaluation of subsets in meta-analysis. *Stat Med;***6**:371–388.

Gelber, RD, Goldhirsch, A. (1991) Meta-analysis: The fashion of summing-up evidence. Part I. Rationale and conduct. *Ann Oncol;***2**:461–468.

Gelber, RD, Goldhirsch, A. (1993) From the overview to the patient: how to interpret meta-analysis data. *Recent Results in Cancer Research;***127**:167–176.

Geller, NL, Proschan, M. (1996) Meta-analysis of clinical trials: a consumer's guide. *J Biopharmaceutical Stats;***6**:377–394.

Geller, NL, Scher, HI, Parmar, MKB, Dalesio, O, Kaye, S. (1990) Can we combine available data to evaluate the effects of neoadjuvant chemotherapy for invasive bladder cancer? *Seminars in Oncology;***17**:628–634.

General Accounting Office. (1992) Cross design synthesis: a new strategy for medical effectiveness research. Washington, DC: GAO.

George, EO. (1977) Combining independent one-sided and two-sided statistical tests – some theory and applications. Unpublished Thesis, University of Rochester.

Gerbarg, ZB, Horwitz, RI. (1988) Resolving conflicting clinical trials: guidelines for meta-analysis. *J Clin Epidemiol;***41**:503–509.

Gervasio, AL, Pfeffer, MA, Hamm, P, Wertheimer, J, Rouleau, JL, Braunwald, E. (1992) Do the results of randomised clinical trials of cardiovascular drugs influence medical practice? *N Engl J Med;***327**:241–247.

Gibaldi, M. (1993) Meta-analysis. A review of its place in therapeutic decision making (review). *Drugs;***46**:805–818.

Gilbody, S, House, A. (1995) Publication bias and meta-analysis (letter; comment). *Br J Psychiatry;***167**:266.

Gilbody, S, House, A, Song, F, Sheldon, T. (1995) Misleading meta-analysis. subject to many potential biases (letter). *BMJ;***311**:1303–1304.

Gingerich, WJ. (1984) Meta-analysis of applied time series data. *Journal of Applied Behavioral Science;***20**:71–79.

Givens, GH, Smith, DD, Tweedie, RL. (1995) Estimating and adjusting for publication bias using data augmentation in Bayesian meta-analysis. Technical Report # 95/31, Department of Statistics, Colorado State University.

Givens, GH, Smith, DD, Tweedie, RL. (1997) Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. Anonymous.

Glass, GV. (1976) Primary, secondary and meta-analysis of research. *Educ Res;*5:3–8.

Glass, GV. (1996) Reply to Mansfield and Busse. *Educ Res;***7**:3.

Glass, GV, McGraw, B, Smith, ML. (1981) Meta-analysis in social research, Newbury Park, CA: Sage.

Glasziou, P, Irwig, L. (1995) Generalizing randomized trial results using additional epidemiologic information. *Am J Epidemiol;***141**:S47.

Glasziou, PP, Irwig, LM. (1995) An evidence based approach to individualizing treatment. *BMJ;***311**:1356–1359.

Gleser, LJ, Olkin, I. (1994) Stochastically dependent effect sizes. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 339–356. New York: Russell Sage Foundation.

Gleser, LJ, Olkin, I. (1995) Models for estimating the number of unpublished studies. Technical Report #313, Department of Statistics, Stanford University.

Godlee, F. (1994) The Cochrane collaboration: deserves the support of doctors and governments. *BMJ;***309**:969–970.

Goldman, L, Feinstein, AR. (1979) Anticoagulants and myocardial infarction. The problems of pooling, drowning and floating. *Ann Intern Med;***90**:92–94.

Goldsmith, JR, Beeser, S. (1984) Strategies for pooling data in occupational epidemiological studies. *Ann Acad Med;***13**:297–307.

Gonser, M, Vetter, H, Noack, F, Schulz, KF. (1995) Meta-analyses of interventional trials done in populations with different risks. *Lancet;***345**:1304–1305.

Goodkin, K, Mcgrath, P.A. (1996) Beyond metaanalysis – the case of pediatric migraine headache. *Pain;***65**:119–121.

Goodman, C. (1993) Literature searching and evidence interpretation for assessing health care practices, Stockholm: SBU: The Sweedish Council of Technology Assessment in Health Care.

Goodman, SN. (1989) Meta-analysis and evidence. *Controlled Clin Trials;***10**:188–204.

Goodman, SN. (1991) Have you ever meta-analysis you didn't like? *Annals of Internal Medicine;***114**:244–246.

Goodman, SN. (1993) Calculation errors in metaanalyses – reply. *Annals of Internal Medicine;***118**:78.

Gotzsche, PC. (1987) Reference bias in reports of drug trials. *BMJ;***295**:654–656.

Gotzsche, PC. (1989) Patients' preference in indomethacin trials: an overview. *Lancet;***1**:88–91.

Gotzsche, PC. (1989) Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled Clin Trials;***10**:31–56.

Gotzsche, PC. (1989) Multiple publication of reports of drug trials. *European Journal of Clinical Pharmacology;***36**:429–432.

Gotzsche, PC. (1990) Sensitivity of effect variables in rheumatoid arthritis: a meta-analysis of 130 placebo controlled NSAID trials. *J Clin Epidemiol;***43**:1313–1318.

Gotzsche, PC. (1993) Meta-analysis of NSAIDs: contribution of drugs, doses, trial designs, and meta-analytic techniques. *Scand J Rheumatol;***22**:255–260.

Gotzsche, PC, Lange, B. (1991) Comparison of search strategies for recalling doubleblind trials from MEDLINE. *Dan Med Bull;***38**:47–68.

Gotzsche, PC, Podenphant, J, Olesen, M, Halberg, P. (1992) Meta-analysis of second-line antirheumatic drugs: Sample size bias and uncertain benefit. *J Clin Epidemiol;***45**:587–594.

Gotzsche, PD. (1989) Meta-analysis of grip strength: most common, but superfluous variable in comparative NSAID trials. *Danish Medical Bulletin;***36**:493–495.

Gould, AL, Rossouw, JE, Santanello, NC, Heyse, JF, Furberg, CD. (1995) Cholesterol reduction yields clinical benefit: a new look at old data. *Circulation;***91**:2274–2282.

Goutis, C, Casella, G, Wells, MT. (1996) Assessing evidence in multiple hypotheses. *J Am Statist Assoc;***91**:1268–1277.

Gray, A, Berlin, JA, McKinlay, JB, Longcope, C. (1991) An examination of research design effects on the association of testosterone and male aging: results of a meta-analysis. *J Clin Epidemiol;***44**:671–684.

Green, BF, Hall, JA. (1984) Quantitative methods for literature reviews. *Annual Review of Psychology;***35**:37–53.

Green, SJ, Fleming, TR, Emerson, S. (1987) Effects on overviews of early stopping rules for clinical-trials. *Stat Med;***6**:361.

Greene, RJ, Maklan, CW. (1994) Research into outcomes and effectiveness. *BMJ;***309**:879–880.

Greenhalgh, T. (1997) How to read a paper: Papers that summarise other papers (systematic reviews and meta-analyses). *BMJ;***315**:672–675.

Greenhouse, JB, Fromm, D, Iyengar, S, Dew, MA, Holland, A, Kass, R. (1986) The making of a meta-analysis: a case study of a quantitative review of the aphasia treatment literature. Technical Report. 379, Dept. Statistics, Carnegie Mellon Univ.

Greenhouse, JB, Iyengar, S. (1994) Sensitivity analysis and diagnostics. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 383–398. New York: Russell Sage Foundation.

Greenland, S. (1982) Interpretation of summary measures when interaction is present. *Am J Epidemiol;***116**:587.

Greenland, S. (1982) Interpretation and estimation of summary odds ratios under heterogeneity. *Stat Med;***1**:217–227.

Greenland, S. (1987) Quantitative methods in the review of epidemiological literature. *Epidemiol Rev;***9**:1–30.

Greenland, S. (1994) Can meta-analysis be salvaged? *Am J Epidemiol;***140**:783–787.

Greenland, S. (1994) Invited commentary: a critical look at some popular meta- analytic methods. *Am J Epidemiol;***140**:290–296.

Greenland, S. (1994) Quality scores are useless and potentially misleading. *Am J Epidemiol;***140**:300–301.

Greenland, S. (1995) Point Counterpoint: meta-analysis of observational studies – reply. *Am J Epidemiol;***142**:781–782.

Greenland, S, Longnecker, MP. (1992) Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am J Epidemiol;***135**:1301–1309.

Greenland, S, Salvan, A. (1990) Bias in the one-step method for pooling study results. *Stat Med;***9**:247–252.

Gregoire, G, Derderian, F, Lelorier, J, Le Lorier, J. (1995) Selecting the language of the publications included in a meta-analysis – is there a Tower-of-Babel bias? *J Clin Epidemiol;***48**:159–163.

Gueyffier, F, Boutitie, F, Boissel, JP, Coope, J, Cutler, J, Ekbom, T, *et al.* (1995) INDANA: a meta-analysis on individual patient data in hypertension. Protocol and preliminary results. *Therapie;***50**:353–362.

Guggenmoosholzmann, I. (1995) ''Exploratory or analytic meta-analysis: should we distinguish between them": discussion. *J Clin Epidemiol;***48**:117–118.

Haines, M. (1994) Current issues in evidenced-based practice. *Health Libraries Review;***11**:221–286.

Hale, S, Myerson, J. (1995) Fifty years older, fifty percent slower? Meta-analytic regression models and semantic context effects. *Aging and Cognition;***2**:132–145.

Hall, JA, Rosenthal, R. (1995) Interpreting and evaluating meta-analysis. *Evaluation and the Health Professions;***18**:393–407.

Hall, JA, Tickle-Degnen, L, Rosenthal, R, Mosteller, F. (1994) Hypotheses and problems in research synthesis. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 17–28. New York: Russell Sage Foundation.

Halvorsen, KT. (1986) Combining results from independent investigations: Meta-analysis in medical research. In: Bailar, JCI, Mosteller, F. (eds) Medical Uses of Statistics, pp. 392–416. Waltham, Mass: NEJM Books.

Halvorsen, KT. (1994) The reporting format. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 425–438. New York: Russell Sage Foundation.

Hamrick, HJ, Garfunkel, JM. (1994) Therapy for acute otitis media: applicability of metaanalysis to the individual patient [editorial; comment]. *J Pediatr;***124**:431.

Hansson, L, Fagard, R, Mancia, G, Sleight, P, Macmahon, S. (1996) Advantages and disadvantages of the metaanalysis approach – discussion. *Journal of Hypertension;***14**:S13.

Haque, KN. (1995) Pitfalls of meta-analysis. *Arch Dis Child;***73**:F196.

Hardy, RJ, Thompson, SG. (1996) A likelihood approach to meta-analysis with random effects. *Stat Med;***15**:619–629.

Harlan, WR. (1994) Creating an NIH clinical trials registry: a user-friendly approach to health care. *JAMA*;**271**:1729.

Hartmann, A, Herzog, T. (1995) Varianten der effektstarkenberechnung in meta-analysen: kommt es zu variablen ergebnissen? Calculating effect size by varying formulas: are there varying results? *Zeitschrift fur Klinische Psychologie;***24**:337–343.

Hartzema, AG. (1992) Guide to interpreting and evaluating the pharmacoepidemiologic literature. *Ann Pharmacother;***26**:96–98.

Harville, DA. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J Am Statist Assoc;***72**:320–338.

Haselkorn, JK, Turner, JA, Diehr, PK, Ciol, MA, Deyo, RA. (1994) Meta-analysis. A useful tool for the spine researcher (review). *Spine;***19**:2076S–2082S.

Hasselblad, V. (1994) Meta-analysis in environmental statistics. In: Patil, GP, Rao, CR. (eds) Handbook of Statistics, Volume 12: Environmental Statistics, pp. 691–716. New York: North Holland/Elsevier.

Hasselblad, V. (1995) Meta-analysis of environmental health data. *Science of the Total Environment;*160–161, 545–558.

Hasselblad, V, Hedges, LV. (1995) Meta-analysis of screening and diagnostic tests. *Psychol Bull;***117**:167–178.

Hasselblad, V, Mosteller, F, Littenberg, B, Chalmers, TC, Hunink, MGM, Turner, JA, *et al.* (1995) A survey of current problems in metaanalysis – discussion from the agency for health-care policy and research inter-port work group on literature-review metaanalysis. *Medical Care;***33**:202–220.

Hasselblad, VIC, Mccrory, DC. (1995) Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making;***15**:81–96.

Hauck, WW. (1984) A comparative study of conditional maximum likelihood estimation of a common odds ratio. *Biometrics;***40**:1117–1123.

Haynes, RB. (1992) Clinical review articles. *BMJ;***304**:330–331.

Haynes, RB, Mulrow, CD, Huth, EJ, Altman, DG, Gardner, MJ. (1990) More informative abstracts revisited. *Ann Intern Med;***113**:69–76.

Heatherington, J, Dickersin, K, Chalmers, I, Meinert, CL. (1989) Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. *Pediatrics;***84**:374–380.

Hedges, L. (1981) Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Stat;***6**:107–128.

Hedger, L. (1986) Statistical issues in the meta-analysis of environmental studies. In: ASA/EPA Conferences on interpretation of environmental data. II. Statistical issues in combining environmental studies. 2: 30–44 Washington DC: EPA. (1987).

Hedges, LV. (1982) Fitting categorical models to effect sizes from a series of experiments. *J Educ Statist;***7**:119–137.

Hedges, LV. (1982) Estimating effect size from a series of independent experiments. *Psychol Bull;***92**:490–499.

Hedges, LV. (1982) Fitting continuous models to effect size data. *J Educ Statist;***7**:245–270.

Hedges, LV. (1983) Combining independent estimators in research synthesis. *British Journal of Mathematical and Statistical Psychology;***36**:123–131.

Hedges, LV. (1984) Estimation of effect size under nonrandom sampling: the effects of censoring studies yielding statistically insignificant mean differences. *J Educ Stat;***9**:61–85.

Hedges, LV. (1986) Estimating effect sizes from vote counts or box score data. Conference proceeding. American Educational Research Association, San Fransisco.

Hedges, LV. (1987) Commentary. *Stat Med;***6**:381–385.

Hedges, LV. (1989) Estimating the normal mean and variance under a publication selection model. In: Gleser, LJ, Perlman, MD, Press, SJ, Sampson, AR. (eds) Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin, pp. 447–458. New York: Springer.

Hedges, LV. (1992) Modeling publication selection effects in meta-analysis. *Statistical Science;***7**:246–255.

Hedges, LV. (1994) Fixed effects models. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 285–300. New York: Russell Sage Foundation.

Hedges, LV. (1994) Statistical considerations. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 29–40. New York: Russell Sage Foundation.

Hedges, LV, Olkin, I. (1980) Vote-counting methods in research synthesis. *Psychol Bull;***88**:359–369.

Hedges, LV, Olkin, I. (1984) Nonparametric estimators of effect size in meta-analysis. *Psychol Bull;***96**:573–580.

Hedges, LV, Olkin, I. (1985) Statistical Methods for Meta-analysis, London: Academic Press.

Helfand, M. (1993) Meta-analysis in deriving summary estimates of test performance [editorial]. *Med Decis Making;***13**:182–183.

Hellmich, M, Sutton, AJ. (1997) A Bayesian approach to meta-analysis of areas under ROC curves. Technical Report, Department of Epidemiology and Public Health: University of Leicester, England.

Hemminki, E. (1980) Study of information submitted by drug companies to licensing authorities. *BMJ;***280**:833–836.

Henderson, WG, Moritz, T, Goldman, S, Copeland, J, Sethi, G. (1995) Use of cumulative meta-analysis in the design, monitoring, and final analysis of a clinical trial: a case study. *Controlled Clin Trials;***16**:331–341.

Hennekens, CH, Buring, JE, Hebert, PR. (1987) Implications of overviews of randomized trials. *Stat Med;***6**:397–409.

Henry, DA, Wilson, A. (1992) Meta-analysis: part 1: an assessment of its aims, validity and reliability. *Med J Aust;***156**:31–38.

Herbold, M. (1993) Meta-analysis of environmental and occupational epidemiological studies: a method demonstrated using the carcinogenicity of PCBs as an example. *Soz Praventivmed;***38**:185–189.

Hertz-Picciotto, I, Neutra, RR. (1994) Resolving discrepancies among studies: the influence of dose on effect size. *Epidemiology;***5**:156–163.

Higgins, JPT, Whitehead, A. (1996) Borrowing strength from external trials in a metaanalysis. *Stat Med;***15**:2733–2749.

Higgins, JPT, Whitehead, A. (1997) Inclusion of both patient level and study-level covariates in a meta-analysis. *Controlled Clin Trials;*18:84S.

Higgins, MS, Stiff, JL. (1993) Pitfalls in performing meta-analysis: I (letter). *Anesthesiology;***79**:405.

Hlatky, MA. (1991) Using databases to evaluate therapy. *Stat Med;***10**:647–652.

Hofmans, EA. (1990) Publikatiebias – realiteit of mythe? De toegankelijkheid van onderzoek naar de effectiviteit van acupunctuur. Publication bias – reality or myth? The accessibility of research on the effectivity of acupuncture. *Huisarts en Wetenschap;***33**:14–15.

Hofmans, EA. (1990) The results of a MEDLINE search. The accessibility of research on the effectiveness of acupuncture II. *Huisarts Wet;***33**:103–106.

Holme, I. (1996) Relationship between total mortality and cholesterol reduction as found by meta-regression analysis of randomized cholesterol-lowering trials. *Controlled Clin Trials;***17**:13–22.

Holme, I, Ekelund, LG, Hjermann, I, Leren, P. (1994) Quality-adjusted meta-analysis of the hypertension/coronary dilemma. *Amer J Hypertens;***7**:703–712.

Horwitz, RI. (1995) ‘Large-scale randomized evidence: large, simple trials and overviews of trials’: discussion. A clinician’s perspective on meta-analyses [comment]. *J Clin Epidemiol;***48**:41–44.

Howard, S, Mugford, M, Normand, C. (1996) A cost-effectiveness analysis of neonatal ECMO using existing evidence. *International Journal of Technology Assessment in Health Care*

Hughes, EG. (1996) Systematic literature-review and metaanalysis. *Seminars In Reproductive Endocrinology;***14**:161–169.

Hughes, EG, Fedorkow, DM, Collins, JA. (1993) A quantitative overview of controlled trials in endometriosis-associated infertility. *Fertil Steril;***59**:963–970.

Hughes, MD, Degruttola, V, Welles, SL. (1995) Evaluating surrogate markers. [review]. *J Acquir Immune Defic Syndr Hum Retrovirol;***10**:S1–8.

Hughes, MD, Freedman, LS, Pocock, SJ. (1992) The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics;***48**:41–53.

Hunink, MGM, Wong, JB. (1994) Meta-analysis of failure-time data with adjustment for covariates. *Med Decis Making;***14**:59–70.

Hunter, JE, Schmidt, FL. (1994) Correcting for sources of artificial variation across studies. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 323–338. New York: Russell Sage Foundation.

Hunter, JE, Schmit, FL. (1990) Methods of Meta-analysis: Correcting error and bias in research findings. SAGE Publications.

Huque, MF. (1988) Experiences with meta-analysis in NDA submissions. *Proc Biopharmaceutical Section of the American Statistical Association*:28–33.

Huque, MF, Dubey, SD. (1994) A metaanalysis methodology for utilizing study-level covariate-information from clinical trials. *Communications In Statistics – Theory And Methods;***23**:377–394.

Hutchison, BG. (1993) Critical appraisal of review articles (review). *Can Fam Physician;***39**:1097–1102.

Huth, EJ. (1987) Needed: review article with more scientific rigor. *Ann Intern Med;***106**:470–471.

Irwig, L. (1995) 'A statistician looks at met-analysis': Discussion. *J Clin Epidemiol;***48**:105–108.

Irwig, L, Glasziou, P. (1996) The Cochrane methods working group on systematic review of screening and diagnostic tests: recommended methods. Online: http://som. flinders. edu. au/cochrane/.

Irwig, L, Macaskill, P, Glasziou, P, Fahey, M. (1995) Meta-analytic methods for diagnostic test accuracy (review). *J Clin Epidemiol;***48**:119–130.

Irwig, L, Tosteson, AN, Gatsonis, C, Lau, J, Colditz, G, Chalmers, TC, Mosteller, F. (1994) Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med;***120**:667–676.

Irwig, L, Tosteson, ANA, Gatsonis, C. (1994) Metaanalyses evaluating diagnostic-tests – response. *Ann Intern Med;***121**:817–818.

Iyengar, S. (1991) Much ado about meta-analysis. *Chance: New Directions for Statistics and Computing;***4**:33–40.

Iyengar, S, Greenhouse, JB. (1988) Selection models and the file drawer problem. *Statistical Science;***3**:109–135.

Jackson, GB. (1980) Methods for integrative reviews. Review of Educational Research 50, 438–460.

Jadad, AR, McQuay, HJ. (1993) Searching the literature: Be systematic in your searching. *BMJ;***307**:66.

Jadad, AR, McQuay, HJ. (1993) A high-yield strategy to identify randomized controlled trials for systematic reviews. *Online Journal of Current Clinical Trials;*Doc No 33.

Jadad, AR, McQuay, HJ. (1996) Meta-analyses to evaluate analgesic interventions: a systematic qualitative review of their methodology (review). *J Clin Epidemiol;***49**:235–243.

Jadad, AR, Moore, RA, Carroll, D, Jenkinson, C, Reynolds, DJM, Gavaghan, DJ, McQuay, H. (1996) Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clin Trials;***17**:1–12.

Jayaratne, S, Tripodi, T, Talsma, E. (1988) The comparative analysis and aggregation of single case data. *Journal of Applied Behavioral Science;***24**:119–128.

Jefferson, T, DeMicheli, V, Mugford, M. (1996) Current issues. In: Anonymous Elementary economic evaluation in health care, London: BMJ Publishing Group.

Jefferson, T, Mugford, M, Gray, A, DeMicheli, V. (1996) An exercise in the feasibility of carrying out secondary economic analysis. *Health Economics;*5:155–165.

Jeng, GT, Scott, JR, Burmeister, LF. (1995) A comparison of metaanalytic results using literature vs individual patient data – paternal cell immunization for recurrent miscarriage. *JAMA;***274**:830–836.

Jenicek, M. (1989) Meta-analysis in medicine: where we are and where we want to go. *J Clin Epidemiol;***42**:35–44.

Jensen, LA, Allen, MN. (1996) Meta-synthesis of qualitative findings. *Qualitative Health Research;***6**:553–560.

Johnson, BT. (1989) DSTAT: Software for the meta-analytic review of research literatures, Hillsdale, NJ: Erlbaum.

Johnstone, BM, Leino, EV, Motoyoshi, MM, Temple, MT, Fillmore, KM, Hartka, E. (1991) An integrated approach to metaanalysis in alcohol studies. *British Journal of Addiction;***86**:1211–1220.

Jones, DR. (1992) Meta-analysis of observational epidemiological studies: a review. *J R Soc Med;***85**:165–168.

Jones, DR. (1993) El metanalisis en los estudos epidemiologicos observacionales. *Bol of Sanit Panam;***115**:438–445.

Jones, DR. (1995) Meta-analysis: weighing the evidence. *Stat Med;***14**:137–149.

Jones, DR, Lewis, JA. (1992) Meta-analysis in the regulation of medicines. *Pharm Med;***6**:195–205.

Jones, MP, O'Gorman, TW, Lemke, JH, Woolson, RF. (1989) A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics;***45**:171–181.

Kaaks, R, Plummer, M, Riboli, E, Esteve, J, Vanstaveren, W. (1994) Adjustment for bias due to errors in exposure assessments in multicenter cohort studies on diet and cancer – a calibration approach. *American Journal of Clinical Nutrition;*59:S245–S250.

Kardaun, JWPF, Kardaun, OWJF. (1990) Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Meth Inform Med;***29**:12–22.

Kass, RE, Steffey, D. (1989) Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *J Am Statist Assoc;***84**:717–726.

Kassirer, JP. (1992) Clinical trials and meta-analysis. What do they do for us? *N Engl J Med;***327**:273–274.

Katerndahl, DA. (1993) Techniques of meta-analysis (letter). *Journal of the American Board of Family Practice;***6**:433–434.

Katsouyanni, K, Zmirou, D, Spix, C, Sunyer, J, Schouten, JP, Ponka, A, *et al.* (1995) Short-term effects of air pollution on health: a European approach using epidemiological time-series data. The APHEA project: background, objectives, design. *Eur Respir J;***8**:1030–1038.

Kattan, MW, Inoue, Y, Giles, FJ, Talpaz, M, Ozer, H, Guilhot, F, *et al.* (1996) Cost-effectiveness of interferon-alpha and conventional chemotherapy in chronic myelogenous leukemia. *Ann Intern Med;***125**:541.

Katz, RT, Campagnolo, DI, Goldberg, G, Parker, JC, Pine, ZM, Whyte, J. (1995) Critical evaluation of clinical research. [review]. *Arch Phys Med Rehabil;***76**:82–93.

Kaufman, DW, Shapiro, S. (1997) Meta-analysis of risk of gastrointestinal complications with NSAIDs. Narrative review should have been used (letter; comment). *BMJ;***314**:445–446.

Keller, F, Erdmann, K, Giehl, M, Buettner, P. (1993) Nonparametric meta-analysis of published data on kidney-function dependence of pharmacokinetic parameters for the aminoglycoside netilmicin (review). *Clinical Pharmacokinetics;***25**:71–79.

Keselman, JC, Lix, LM, Keselman, HJ. (1996) The analysis of repeated measurements – a quantitative research synthesis. *British Journal Of Mathematical & Statistical Psychology;***49**:275–298.

Khan, KS, Daya, S, Collins, JA, Walter, SD. (1996) Empirical evidence of bias in infertility research – overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertility and Sterility;***65**:939–945.

Khan, KS, Daya, S, Jadad, AR. (1996) The importance of quality of primary studies in producing unbiased systematic reviews. *Archives of Internal Medicine;***156**:661–666.

Kilpatrick, SJ. (1992) The epidemiology of environmental tobacco smoke (ETS) and the weight of evidence argument. *Int Surg;***77**:131–133.

Kirpalani, H, Schmidt, B, McKibbon, KA, Haynes, RB, Sinclair, JC. (1989) Searching MEDLINE for randomized clinical trials involving care of the newborn. *Pediatrics;***83**:543–546.

Kleijnen, J, Knipschild, P. (1992) Review articles and publication bias. *Arzneimittel-Forschung/Drug Research;*42–1, 587–591.

Klein, S, Simes, J, Blackburn, GL. (1986) Total parenteral nutrition and cancer clinical trials. *Cancer;***58**:1378–1386.

Kleinjen, J, Knipschild, P. (1992) The comprehensiveness of Medline and Embase computer searches. *Pharm Weekbl[Sci];***14**:316–320.

Knipschild, P. (1994) Systematic reviews – some examples. *BMJ;***309**:719–721.

Koch, GG, Schmid, JE, Begun, JM, Maier, WC. (1993) Meta-analysis of drug safety data. In: Anonymous Drug Safety Assessment in Clinical Trials, pp. 279–304. New York: Marcel Dekker.

Kong, A, Liu, JS, Wong, WH. (1994) Sequential imputations and Bayesian missing data problems. *J Am Statist Assoc;***89**:278–288.

Kopcke, W. (1996) Metaanalysis. *Geburtshilfe und Frauenheilkunde;***56**:M9–M13.

Koren, G, Shear, H, Graham, K, Einarson, T. (1989) Bias against the null hypothesis – the reproductive hazards of cocaine. *Lancet;***2**:1440–1442.

Kotchmar, DJ, Hasselblad, V. (1993) The influence of covariates in a metaanalysis of respiratory effects associated with nitrogen-dioxide exposure. *American Review of Respiratory Disease;***147**:A630.

Kraemer, HC, Andrews, G. (1982) A non-parametric technique for meta-analysis effect size calculation. *Psychol Bull;***91**:404–412.

Kriebel, D, Wegman, DH, Moure-Eraso, R, Punnett, L. (1990) Limitations of meta-analysis: cancer in the petroleum industry. *Am J Ind Med;***17**:269–271.

Kurosawa, K. (1984) Discussion on meta-analysis and selective publication bias. *Am Psychol;***39**:73–74.

Kuss, O, Koch, A. (1996) Metaanalysis macros for sas. *Computational Statistics & Data Analysis;***22**:325–333.

L'Abbe, KA, Detsky, AS, O'Rourke, K. (1987) Meta-analysis in clinical research. *Ann Intern Med*;**107**:224–233.

Lacy, JB, Ohlsson, A. (1995) Pitfalls of meta-analysis – Comment. *Arch Dis Child;***73**:F196.

Lacy, JB, Ohlsson, A. (1996) Consistent definition of outcomes – a prerequisite for metaanalysis. *American Journal of Obstetrics and Gynecology;***175**:1394–1395.

LaFleur, B, Taylor, S, Smith, DD, Tweedie, RL. (1996) Bayesian assessment of publication bias in meta-analyses of cervical cancer and oral contraceptives. Anonymous.

Laird, N, Louis, TA. (1989) Empirical Bayes confidence intervals for a series of related experiments. *Biometrics;***45**:481–495.

Laird, NM, Mosteller, F. (1990) Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care;***6**:5–30.

Lambert, PC, Abrams, KR. (1996) Meta-analysis using multilevel models. *Multilevel Modelling Newsletter;***7**:17–19.

Lancet. (1991) Making clinical trialists register. *Lancet;***338**:244–245.

Lancet (1992) Cross design synthesis: a new strategy for studying medical outcomes? *Lancet;***340**:944–946.

Lane, DM, Dunlap, WP. (1978) Estimating effect-size bias resulting from significance criterion in editorial decisions. *Br J Math Stat Psyc;***31**:107–112.

Larholt, KM. (1989) Statistical Methods and Heterogeneity in Meta-analysis. Harvard School of Public Health, Boston.

Larose, DT, Dey, DK. (1997) Grouped random effects models for Bayesian meta-analysis. *Stat Med;***16**:1817–1829.

Larose, DT, Dey, DK. (1997) Modeling dependent covariate subclass effects in Bayesian meta-analysis. Technical Report #96–22, University of Connecticut.

Lau, J, Antman, EM. (1994) Lessons from metaanalyses of acute MI trials. *Choices in Cardiology;***8**:2–5.

Lau, J, Antman, EM, Jimenez-Silva, J, Kupelink, B, Mosteller, SF, Chalmers, TC, *et al.* (1992) Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med;***327**:248–254.

Lau, J, Chalmers, TC. (1995) The rational use of therapeutic drugs in the 21st century: Important lessons from cumulative meta-analyses of randomized control trials. *International Journal of Technology Assessment in Health Care;***11**:509–522.

Lau, J, Schmid, CH, Chalmers, TC. (1995) Cumulative meta-analysis of clinical trials: builds evidence for exemplary medical care. *J Clin Epidemiol;***48**:45–57.

Laupacis, A, Sackett, DL, Roberts, RS. (1988) An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med;***318**:1728–1733.

Law, MR, Frost, CD, Wald, NJ. (1991) By how much does dietary salt reduction lower blood pressure? I – Analysis of observational data among populations. *BMJ;***302**:811–815.

Law, MR, Wald, NJ. (1995) Misleading meta-analysis. one incorrect meta-analysis does not invalidate them all (letter). *BMJ;***311**:1303.

Le Fanu, J. (1995) Misleading meta-analysis. public policy is based on results of epidemiological meta-analyses that contradict common sense (letter). *BMJ;***310**:1603–1604.

Lecky, FE, Little, RA, Brennan, P. (1996) The use and misuse of metaanalysis. *Journal Of Accident & Emergency Medicine;***13**:373–378.

Lee, YJ, Ellenberg, JH, Hirtz, DG, Nelson, KB. (1991) Analysis of clinical-trials by treatment actually recieved – is it really an option? *Stat Med;***10**:1595–1605.

Leidl, RM. (1994) Some factors to consider when using the results of economic avaluation studies at the population level. *International Journal of Technology Assessment in Health Care;***10**:467–478.

Leitich, H, Egarter, C. (1996) Consistent definition of outcomes – a prerequisite for metaanalysis – reply. *American Journal of Obstetrics and Gynecology;***175**:1395.

Leizorovicz, A, Haugh, MC, Boissel, JP. (1992) Metaanalysis and multiple publication of clinical-trial reports. *Lancet;***340**:1102–1103.

Lent, V, Langenbach, A. (1996) A retrospective quality analysis of 102 randomized trials in 4 leading urological journals from 1984–1989. *Urological Research;***24**:119–122.

Letzel, H. (1995) 'Best-evidence synthesis: an intelligent alternative to meta-analysis': discussion. A case of 'either-or' or 'as well' [comment]. *J Clin Epidemiol;***48**:19–21.

Levine, M, Walters, S, Lee, H, Haines, T, Holbrook, A, Moyer, V. (1994) IV: How to use an article about harm. *JAMA;*1615–1619.

LeVois, ME, Layard, MW. (1995) Publication bias in the environmental tobacco smoke/coronary heart disease epidemiologic literature (review). *Regul Toxicol Pharmacol;***21**:184–191.

Levy, G. (1991) Publication bias: its implications for clinical pharmacology. *Clinical Pharmacology and Therapeutics;***52**:115–119.

Li, YZ, Powers, TE, Roth, HD. (1994) Random-effects linear-regression metaanalysis models with application to the nitrogen-dioxide health-effects studies. *Journal Of The Air & Waste Management Association;***44**:261–270.

Li, YZ, Shi, L, Roth, HD. (1994) The bias of the commonly-used estimate of variance in metaanalysis. *Communications In Statistics – Theory And Methods;***23**:1063–1085.

Li, Z. (1995) A multiplicative random effects model for meta-analysis with application to estimation of admixture component. *Biometrics;***51**:864–873.

Li, ZH, Begg, CB. (1994) Random effects models for combining results from controlled and uncontrolled studies in a metaanalysis. *J Am Statist Assoc;***89**:1523–1527.

Li, ZH, Rao, DC. (1996) Random effects model for metaanalysis of multiple quantitative sibpair linkage studies. *Genetic Epidemiology;***13**:377–383.

Liang, K, Self, SG. (1985) Tests for homogeneity of odds ratio when the data are sparse. *Biometrica;***72**:353–358.

Liberati, A. (1995) 'Meta-analysis: statistical alchemy for the 21st century': discussion. A plea for a more balanced view of meta-analysis and systematic overviews of the effect of health care interventions [comment]. *J Clin Epidemiol;***48**:81–86.

Light, RJ. (1987) Accumulating evidence from independent studies – what we can win and what we can lose. *Stat Med;***6**:221–231.

Light, RJ, Pillemar, DB. (1984) Summing Up: The science of Reviewing Research. Cambridge, Mass: Harvard University Press.

Light, RJ, Singer, JD, Willett, JB. (1994) The visual presentation and interpretation of meta-analyses. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 439–454. New York: Russell Sage Foundation.

Light, RJ, Smith, PV. (1971) Accumulating evidence: procedures for resolving contradictions among different research studies. Harvard Educational Review 41, 429–471.

Linde, K, Melchart, D, Jonas, W.B. (1994) Durchfuhrung und interpretation systematischer ubersichtsarbeiten kontrollierter studien in der komplementarmedizin. *Separata aus Forschende Komplementarmedizin;***1**:8–16.

Link, WA, Sauer, JR. (1995) Estimation and confidence intervals for empirical mixing distributions. *Biometrics;***51**:810–821.

Lipsey, MW. (1994) Identifying potentially interesting variables and analysis opportunities. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 111–124. New York: Russell Sage Foundation.

Littenberg, B, Moses, LE. (1993) Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making;***13**:313–321.

Little, RJA. (1992) Regression with missing x's: a review. *J Am Statist Assoc;***87**:1227–1237.

Little, RJA, Rubin, DB. (1987) Statistical analysis with missing data, New York: Wiley.

Longnecker, MP. (1995) Re: 'point/counterpoint: meta-analysis of observational studies' (letter). *Am J Epidemiol;***142**:779–782.

Longnecker, MP, Berlin, JA, Orza, MJ, Chalmers, TC. (1988) A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA;***260**:652–656.

Looney, MA, Feltz, CJ, VanVleet, CN. (1994) The reporting and analysis of research findings for within-subject designs: methodological issues for meta-analysis. *Res Q Exerc Sport;***65**:363–366.

Louis, TA. (1984) Estimating a population of parameter values using bayes and empirical bayes methods. *J Am Statist Assoc;***79**:393–398.

Louis, TA. (1991) Assessing, accommodating, and interpreting the influences of heterogeneity. *Environmental Health Perspectives;***90**:215–222.

Louis, TA. (1993) Meta-analysis of clinical studies: the whole is greater than the sum of its parts [editorial]. Transfusion 33, 698–700.

Louis, TA, Fineberg, HV, Mosteller, F. (1985) Findings for public health from meta-analyses. *Annual Review of Public Health;***6**:1–20.

Louis, TA, Robins, J, Dockery, DW, *et al.* (1986) Explaining discrepancies between longitudinal and cross-sectional models. *Journal of Chronic Diseases;***39**:831–839.

Louis, TA, Zelterman, D. (1994) Bayesian approaches to research synthesis. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 411–422. New York: Russell Sage Foundation.

Lowe, HJ, Barnett, GO. (1994) Understanding and using the Medical Subject Headings (MeSH) Vocabulary to perform literature searches. *JAMA;***271**:1103–1108.

Lubsen, J. (1996) Mega-trials: is meta-analysis an alternative? *Eur J Clin Pharmacol;***49**:S29–S33.

Lund, T. (1988) Some metrical issues with meta-analysis of therapy effects. *Scandinavian Journal of Psychology;***29**:1–8.

Macarthur, C, Foran, PJ, Bailar, JC. (1995) Qualitative assessment of studies included in a meta- analysis: DES and the risk of pregnancy loss. *J Clin Epidemiol;***48**:739–747.

Maclure, M. (1993) Demonstration of deductive meta-analysis: ethanol intake and risk of myocardial infarction. *Epidemiol Rev;***15**:328–351.

Macmahon, S, Peto, R. (1990) Blood pressure,stroke and coronary heart disease, part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet;***335**:765–774.

Macmahon, S, Zanchetti, A, Menard, J, Swales, JD, Luscher, TF, Sleight, P. (1996) Integration of trial, metaanalysis and cohort results with treatment guidelines – discussion. *Journal of Hypertension;***14**:S133–S134.

Mahoney, MJ. (1990) Bias, controversy, and abuse in the study of the scientific publication system. *Science, Thechnology, & Human Values;***15**:50–55.

Malaguarnera, M, Restuccia, S, Motta, M, Pistone, G, Trovato, B. (1995) Meta-analysis: limits and benefits. *Clin Drug Invest;***10**:188–189.

Malec, D, Sedransk, J. (1992) Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika;***79**:593–601.

Mann, C. (1990) Meta-analysis in the breech. *Science;***249**:476–480.

Mann, CC. (1994) Can meta-analysis make policy? *Science;***266**:960–962.

Mansfield, RS, Busse, TV. (1977) Meta-analysis of research: a rejoinder to Glass. *Education Research;***6**:3.

Mant, J, Hicks, N, Rosenberg, W, Sackett, D. (1996) How to use overviews of prevention trials to treat individual patients. *Cerebrovascular Diseases;*6:34–39.

Mantel, N. (1963) Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure. *J Am Statist Assoc;***58**:690–700.

Mantel, N, Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst;***22**:719–748.

Marchioli, R. (1993) Easy meta-analysis? *Journal of Nephrology;*6:193–201.

Marchioli, R, Marfisi, RM, Carinci, F, Tognoni, G. (1996) Metaanalysis, clinical-trials, and transferability of research results into practice – the case of cholesterol-lowering interventions in the secondary prevention of coronary heart disease. *Archives of Internal Medicine;*156:1158–1172.

Marcus, RJ. (1997) Son of metaanalysis. *Chemtech;***27**:50.

Margitic, SE, Morgan, TM, Inouye, SK, Landefeld, CS, Durham, NC, Sager, MA, *et al.* (1992) Planning, designing and implementing a prospective meta-analysis study. *Stat Med* (Abstract)* (* We could not find this – additional notes are: As delivered at ISCB Brussels July 1992 (Paper a35)).

Margitic, SE, Morgan, TM, Sager, MA, Furberg, CD. (1995) Lessons learned from a prospective meta-analysis [see comments]. *J Am Geriatr Soc;***43**:435–439.

Maritz, JS, Lwin, T. (1989) Empirical Bayes Methods, 2nd edn. London: Chapman and Hall.

Matt, GE, Cook, TD. (1994) Threats to the validity of research synthesis. In: Cooper, H, Hedges, L.V. (eds) The handbook of research synthesis, pp. 503–520. New York: Russell Sage Foundation.

Matuzzi, M, Hills, M. (1995) Estimating the degree of heterogeneity between event rates using likelihood. *Am J Epidemiol;***141**:369–374.

Mcclish, DK. (1992) Combining and comparing area estimates across studies or strata. *Med Decis Making;***12**:274–279.

McGeer, AJ, Naylor, CD, O'Rourke, K, Detsky, AS. (1989) Study quality as a factor in meta-analysis: inconsistent approaches in the literature. *Clin Res;*37:320.

McIntosh, MW. (1996) The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med;*15:1713–1728.

McKenzie, PJ. (1993) Pitfalls in performing meta-analysis: II (letter). *Anesthesiology;***79**:406–408.

McKibbon, KA, Walker, CJ. (1994) Beyond ACP Journal Club: how to harness Medline for therapy problems [Editorial]. *Annals of Internal Medicine;***121**:A10.

McKibbon, KA, Walker, CJ. (1994) Beyond ACP Journal Club: how to harness Medline for diagnostic problems [Editorial]. *Annals of Internal Medicine;***121**:A10.

McKibbon, KA, Walker, CJ. (1994) Beyond ACP Journal Club: how to harness Medline for etiology problems [Editorial]. *Annals of Internal Medicine;***121**:A10–A11.

McKinlay, SM. (1978) The effect of nonzero second-order interaction on combined estimators of the odds ratio. *Biometrika;***65**:191–202.

Mehta, CR, Patel, NR, Grey, R. (1997) Computing an exact confidence interval for the common odds ratio in several 2 x 2 contingency tables. *J Am Statist Assoc;***80**:969–973.

Meier, P. (1987) Proceedings of methodologic issues in overviews of randomized clinical trials – commentary. *Stat Med;***6**:329–331.

Meinert, CL. (1988) Toward prospective registration of clinical trials. *Controlled Clin Trials;***9**:1–5.

Meinert, CL. (1989) Meta-analysis: science or religion? *Controlled Clin Trials;***10**:257S–263S.

Meitzel, MT, Trull, TJ. (1988) Meta-analytic approaches to social comparisons: a method for measuring clinical significance. *Behavioral Assessment;***10**:159–169.

Menard, J, Chatellier, G. (1996) Integration of trial, metaanalysis and cohort results with treatment guidelines. *Journal of Hypertension;***14**:S129–S133.

Mengersen, K, Besag, J. (1993) Ranking and selection using MCMC. Proceedings of the 3rd Schwerin Conference in Mathematical Statistics. Germany.

Mengersen, KL, Tweedie, RL, Biggerstaff, BJ. (1995) The impact of method choice in meta-analysis. *Aust J Stats;***37**:19–44.

Messori, A, Rampazzo, R. (1993) Metaanalysis of clinical trials based on censored end-points – simplified theory and implementation of the statistical algorithms on a microcomputer. *Computer Methods And Programs In Biomedicine;***40**:261–267.

Messori, A, Scroccaro, G, Martini, N. (1993) Calculation errors in meta-analysis (letter; comment). *Ann Intern Med;***118**:77–78.

Meyer, TJ, Mark, MM. (1996) Statistical power and implications of meta-analysis for clinical research in psychosocial oncology. *Journal of Psychosomatic Research;***41**:409–413.

Mi, J. (1990) Notes on the mle of correlation coefficient in meta analysis. *Communications in Statistics, Theory and Methods;***19**:2035–2052.

Michels, KB. (1992) Quo vadis meta-analysis? A potentially dangerous tool if used without adequate rules (review). Important. *Adv Oncol* *:243–248. (* Difficulty ascertaining volume – not even printed on paper).

Midgette, AS, Stukel, TA, Littenberg, B. (1993) A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. *Med Decis Making;***13**:253–257.

Midgette, AS, Wong, JB, Beshansky, JR, Porath, A, Fleming, C, Pauker, SG. (1994) Cost-effectiveness of streptokinase for acute myocardial infarction – a combined metaanalysis and decision-analysis of the effects of infarct location and of likelihood of infarction. *Med Decis Making;***14**:108–117.

Miller, JN, Colditz, GA, Mosteller, F. (1989) How study design affects outcomes in comparisons of therapy. II: Surgical. *Stat Med;***8**:455–466.

Miller, N, Pollock, VE. (1994) Meta-analytic synthesis for theory development. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 457–484. New York: Russell Sage Foundation.

Mintz, J. (1993) Integrating research evidence: a commentary on meta-analysis. *J Consult Clin Psychol;***51**:71–75.

Moher, D, Fortin, P, Jadad, AR, Juni, P, Klassen, T, Le Lorier, J, *et al.* (1996) Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet;***347**:363–366.

Moher, D, Jadad, AR, Nichol, G, Penman, M, Tugwell, P, Walsh, S. (1995) Assessing the quality of randomized controlled trials – an annotated– bibliography of scales and checklists. *Controlled Clin Trials;***12**:62–73.

Moher, D, Jadad, AR, Tugwell, P. (1996) Assessing the quality of randomised controlled trials: current issues and future directions. *International Journal of Technology Assessment in Health Care;***12**:195–208.

Moher, D, Olkin, I. (1995) Meta-analysis of randomized controlled trials: a concern for standards. *JAMA;***274**:1962–1964.

Moore, A, McQuay, H, Gavaghan, D. (1996) Deriving dichotomous outcome measures from continuous data in randomized controlled trials of analgesics. *Pain;***66**:229–237.

Moreno, V, Martin, ML, Bosch, FX, De Sanjose, S, Torres, F, Munoz, N, *et al.* (1996) Combined analysis of matched and unmatched case-control studies: comparison of risk estimates from different studies. *Am J Epidemiol;***143**:293–300.

Morris, CN. (1983) Parametric empirical Bayes inference: theory and applications. *J Am Statist Assoc;***78**:47–65.

Morris, CN. (1992) Hierachical models for combining information and for meta-analysis. *Bayesian Statistics;***4**:321–344.

Morris, RD. (1994) Meta-analysis in cancer epidemiology. *Environmental Health Perspectives;***102** Suppl 8: 61–66.

Moses, LE, Shapiro, D, Littenberg, B. (1993) Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med;***12**:1293–1316.

Mossman, D, Somoza, E. (1989) Maximizing diagnostic information from the dexamethasone suppression test: an approach to criterion selection using receiver operating characteristic analysis. *Archives of General Psychiatry;***46**:653–660.

Mosteller, F. (1994) Using meta-analysis for research synthesis: pooling data from several studies. In: Ingelfinger, JA, Mosteller, F, Thibodeau, LA, Ware, JH. (eds) Biostatistics in clinical medicine, 3rd edn. pp. 332–360. New York: McGraw-Hill.

Mosteller, F, Chalmers, TC. (1992) Some progress and problems in meta-analysis of clinical trials. *Statistical Science;***7**:227–236.

Mosteller, F, Colditz, GA. (1996) Understanding research synthesis (meta-analysis). *Annual Review of Public Health;*1–23.

Mosteller, F, Gilbert, JP, McPeek, B. (1980) Reporting standards and research strategies for controlled trials: agenda for the editor. *Controlled Clin Trials;*37–58.

Mugford, M, Piercy, J, Chalmers, I. (1991) Cost implications of different approaches to the prevention of respiratory distress syndrome. *Archives of Disease in Childhood;***66**:757–764.

Mullen, B, Rosenthal, R. (1985) Basic Meta-analysis: Procedures and Programs, Hillsdale, NJ. Lawrence Erlbaum Associates.

Mulrow, CD. (1987) The medical review article: state of the science. *Ann Intern Med*;**106**:485–488.

Mulrow, CD. (1994) Systematic reviews – rationale for systematic reviews .1. *BMJ*;**309**:597–599.

Mulrow, CD, Thacker, SB, Pugh, JA. (1988) A proposal for more informative abstracts of review articles. *Ann Intern Med*;**108**:613–615.

Munoz, A, Rosner, B. (1984) Power and sample size for a collection of 2x2 tables. *Biometrics;***40**:995–1004.

National Research Council (1992) Combining Information: Statistical Issues and Opportunities for Research, Washington DC. National Academy Press.

Naylor, CD. (1988) Two cheers for meta-analysis: problems and opportunities in aggregating results of clinical trials. *Can Med Assoc J;***138**:891–895.

Naylor, CD. (1989) Meta-analysis of controlled clinical trials. *Journal of Rheumatology;***16**:424–426.

Naylor, CD. (1997) Meta-analysis and the meta-epidemiology of clinical research. *BMJ*;**315**:617–619.

Needleman, HL. (1995) Environmental lead and children's intelligence. studies included in the meta-analysis are not representative (letter). *BMJ*;**310**:1408; discussion 1409.

Newcombe, RG. (1987) Towards a reduction in publication bias. *BMJ*;**295**:656–659.

Newnham, JP, Evans, SF. (1995) Clinical evaluation of ultrasound technology. *Ultrasound Quarterly;***13**:103–110.

NHS Executive 1995. The National research register [database available on disk]. Welwyn Garden City: VEGA Group Plc.

Nierenberg, AA, Feinstein, AR. (1988) How to evaluate a diagnostic marker test. Lessons from the rise and fall of dexamethasone suppression test. *JAMA;***259**:1699–1702.

Nony, P, Boissel, JP, Lievre, M, Cucherat, M, Haugh, MC, Dayoub, G. (1995) Introduction a la Methodologie metaanalytique. Introduction to meta-analysis methods. *Revue de Medecine Interne;***16**:536–546.

Nony, P, Cucherat, M, Haugh, MC, Boissel, JP. (1995) Critical reading of the meta-analysis of clinical trials. [review]. *Therapie;***50**:339–351.

Normand, SLT. (1995) Metaanalysis software – a comparative review – DSTAT, version 1.10. *American Statistician;***49**:298–309.

Normand, SLT. (1995) Metaanalysis software – a comparative review – Fast*Pro, version 1.0. *American Statistician;***49**:298–309.

Normand, SLT. (1995) Metaanalysis software – a comparative review – true EPISTAT, version 4.0. *American Statistician;***49**:298–309.

Norton, L. (1987) Proceedings of methodologic issues in overviews of randomized clinical trials – commentary. *Stat Med;***6**:333.

Nylenna, M, Riis, P, Karlsson, Y. (1994) Multiple blinded reviews of the same two manuscripts: effects of referee characteristics and publication language. *JAMA;***272**:149–151.

O'Flynn, AI. (1982) Meta-analysis. *Nurs Res;***31**:314–316.

O'Rourke, K, Detsky, AS. (1989) Second thaughts: meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. *J Clin Epidemiol;***42**:1021–1026.

Oakes, M. (1993) The logic and role of meta-analysis in clinical research. *Statistical Methods in Medical Research;***2**:147–160.

Ohlsson, A. (1994) Systematic reviews – theory and practice. *Scandinavian Journal Of Clinical & Laboratory Investigation;***54**:25–32.

Ohlsson, A. (1994) Systematic reviews – theory and practice. [review]. *Scand J Clin Lab Invest Suppl;***54**:25–32.

Ohlsson, A. (1996) Randomized controlled trials and systematic reviews – a foundation for evidence-based perinatal medicine. *Acta Paediatrica;***85**:647–655.

Olkin, I. (1990) History and goals. In: Wachter, K.W, Straf, ML. (eds) The future of meta-analysis, pp. 3–10. New York: Russell Sage Foundation.

Olkin, I. (1992) Reconcilable differences: gleaning insight from conflicting scientific studies. *The Sciences;*July/August: 30–36.

Olkin, I. (1994) Re: 'A critical look at some popular meta-analytic methods' [comment]. *Am J Epidemiol;***140**:297–299.

Olkin, I. (1995) Statistical and theoretical considerations in meta-analysis. *J Clin Epidemio;***48**:133–146.

Olkin, I. (1995) Meta-analysis: reconciling the results of independent studies. *Stat Med;***14**:457–472.

Olkin, I. (1996) Meta-analysis: current issues in research synthesis. *Stat Med;***15**:1253–1257.

Olkin, I, Shaw, DV. (1995) Metaanalysis and its applications in horticultural science. *Hortscience;***30**:1343–1348.

Olsen, J. (1995) Meta-analyses or collaborative studies. *J Occup Environ Med;***37**:897–902.

Origasa, H. (1987) Combining results from several independent studies – a meta analysis (In Japanese). *Oyotokeigaku;***16**:105–114.

Orwin, R. (1983) A fail-safe N for effect size in meta-analysis. *J Ed Statist;***8**:157–159.

Orwin, RG. (1994) Evaluating coding decisions. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 139–162. New York: Russell Sage Foundation.

Osiewalski, J, Steel, MFJ. (1996) A Bayesian analysis of exogeneity in models pooling time-series and cross-sectional data. *Journal Of Statistical Planning And Inference;***50**:187–206.

Ottenbacher, K. (1983) Quantitative reviewing: the literature review as scientific enquiry (research, statistical procedures, methodology). *Am J Occup Ther;***37**:313–319.

Ottenbacher, K. (1992) Impact of random assignment on study outcome: an empirical examination. *Controlled Clin Trials;***13**:50–61.

Ottenbacher, K, York, J. (1984) Strategies for evaluating clinical change: implications for practice and research. *Am J Occup Ther;***38**:647–659.

Oxman, AD. (1994) Checklists for review articles. *BMJ;***309**:648–651.

Oxman, AD, Clarke, MJ, Stewart, LA. (1995) From science to practice. Meta-analyses using individual patient data are needed [editorial; comment]. *JAMA;***274**:845–846.

Oxman, AD, Cook, DJ, Guyatt, GH. (1994) Users guides to the medical literature .6. how to use an overview. *JAMA;***272**:1367–1371.

Oxman, AD, Guyatt, GH. (1988) Guidelines for reading literature reviews. *Can Med Assoc J;***138**:697–703.

Oxman, AD, Guyatt, GH. (1991) Validation of an index of the quality of review articles. *J Clin Epidemiol;***44**:1271–1278.

Oxman, AD, Guyatt, GH. (1992) A consumers guide to subgroup analyses. *Ann Intern Med;***116**:78–84.

Oxman, AD, Guyatt, GH, Singer, J, Goldsmith, CH, Hutchison, BG, Milner, RA, *et al.* (1991) Agreement among reviewers of review articles. *J Clin Epidemiol;***44**:91–98.

Oxman, AD. (1996) The Cochrane Collaboration handbook: preparing and maintaining systematic reviews, Second edn. Oxford: Cochrane Collaboration.

Pao, ML, Worthen, DB. (1989) Retrieval effectiveness by semantic and citation searching. *Journal of the American Society for Information Science;***40**:226–235.

Parmar, MKB, Stewart, LA, Altman, DG. (1996) Metaanalyses of randomized trials – when the whole is more than just the sum of the parts. *Br J Cancer;***74**:496–501.

Parmar, MKB, Torri, V, Stewart, L. (1997) Meta-analysis of the published literature for survival endpoints: making a silk purse out of a pig's ear. Anonymous.

Parmley, WW. (1994) Publication bias [editorial]. *J Am Coll Cardiol;***24**:1424–1425.

Patil, GP, Taillie, C. (1989) Probing encountered data, meta-analysis and weighted distribution methods. In: Dodge, Y. (Ed) Statistical Data Analysis and Inference, B.V, North Holland: Elsevier Science Publishers.

Paul, NL. (1995) Hierarchical Selection Models with Applications in Meta-analysis. Technical Report #621, Department of Statistics, Carnegie Mellon University.

Paul, NL. (1995) Non-parametric classes of weight functions to model publication bias. Technical Report #622, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.

Paul, SR, Donner, A. (1989) A comparison of tests of homogeneity of odds ratios in k 2x2 tables. *Stat Med;***8**:1455–1468.

Pearson, K. (1904) Report on certain enteric fever inoculation statistics. *BMJ;***3**:1243–1246.

Perry, A, Persaud, R. (1995) Misleading meta-analysis. variability among studies should be investigated (letter). *BMJ;***310**:1604.

Persaud, R. (1996) Misleading meta-analysis – ''Fail safe N" is a useful mathematical measure of the stability of results. *BMJ;***312**:125.

Petitti, DB. (1994) Meta-analysis, Decision Analysis and Cost-Effectiveness Analysis. New York: Oxford University Press.

Petitti, DB. (1994) Of babies and bathwater. *Am J Epidemiol;***140**:770–782.

Peto, R. (1987) Why do we need systematic overviews of randomised trials? *Stat Med;***6**:233–240.

Peto, R, Collins, R, Gray, R. (1995) Large-scale randomized evidence: large, simple trials and overviews of trials (review). *J Clin Epidemiol;***48**:23–40.

Peto, R, Pike, MC, Armitage, P, Breslow, NE, Cox, DR, Howard, SV, *et al.* (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II: Analysis and examples. *Br J Cancer;***35**:1–39.

Phillips, A, Holland, PW. (1987) Estimators of the variance of the Mantel–Haenszel log-odds-ratio estimate. *Biometrics;*43:425–431.

Phillips, KA. (1991) The use of meta-analysis in technology assessment: a meta-analysis of the enzyme immunosorbent assay human immunodeficiency virus antibody test. *J Clin Epidemiol;***44**:925–931.

Piedbois, P, Buyse, M. (1993) What can we learn from a meta-analysis of trials testing the modulation of 5–FU by leucovorin? Advanced Colorectal Meta-analysis Project. *Ann Oncol;***4**:15–19.

Piegorsch, WW, Cox, LH. (1996) Combining environmental information .2. environmental epidemiology and toxicology. *Environmetrics;***7**:309–324.

Pignon, JP. (1996) Interest of the cochrane collaboration – the point-of-view of a metaanalysis practitioner. *Therapie;***51**:257–260.

Pignon, JP, Arriagada, R. (1993) Meta-analysis. *Lancet;***341**:418–422.

Pignon, JP, Arriagada, R. (1994) Meta-analyses of randomized clinical trials: how to improve their quality? (review). *Lung Cancer;***10** Suppl 1:S135–S141.

Pignon, JP, Arriagada, R, Ihde, DC, Johnson, DH, Perry, MC, Souhami, RL, *et al.* (1992) A meta-analysis of thoracic radiotherapy for small-cell lung cancer. *N Engl J Med;***327**:1618–1624.

Pignon, JP, Bourhis, J. (1995) Meta-analysis of chemotherapy in head and neck cancer: individual patient data vs literature data (letter; comment). *Br J Cancer;***72**:1062–1063.

Pignon, JP, Poynard, T. (1991) Metaanalysis of randomized clinical-trials. *Gastroenterologie Clinique Et Biologique;***15**:229–238.

Pignon, JP, Poynard, T. (1993) [Meta-analysis of therapeutic trials. Principles, methods and critical reading]. [French]. *Rev Prat;***43**:2383–2386.

Pigott, TD. (1994) Methods for handling missing data in research synthesis. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 163–176. New York: Russell Sage Foundation.

Pillemar, DB, Light, RJ. (1980) Synthesizing outcomes: how to use research evidence from many studies. *Harvard Educ Rev;***50**:176–195.

PladevallVila, M, Delclos, GL, Varas, C, Guyer, H, BruguesTarradellas, J, AngladaArisa, A. (1996) Controversy of oral contraceptives and risk of rheumatoid arthritis: meta-analysis of conflicting studies and review of conflicting meta- analyses with special emphasis on analysis of heterogeneity. *Am J Epidemiol;***144**:1–14.

Pocock, S. (1993) Meta-analysis [editorial]. *Stat Methods Med Res;***2**:117–119.

Pocock, SJ. (1976) The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases;***29**:175–188.

Pocock, SJ, Hughes, MD. (1990) Estimation issues in clinical trials and overviews. *Stat Med;***9**:657–671.

Powe, NR, Turner, JA, Maklan, CW, Ersek, M. (1994) Alternative methods for formal literature review and metaanalysis in ahcpr patient outcomes research teams. *Medical Care;***32**:JS22–JS37.

Powe, NR, Turner, JA, Maklan, CW, Ersek, M. (1994) Alternative methods for formal literature review and meta-analysis in ahcpr patient outcomes research teams. agency for health care policy and research. *Med Care;***32**:S22–S37.

Poynard, T, Conn, H.O. (1985) The retrieval of randomized clinical trials in liver disease from the medical literature. A comparison of MEDLARS and manual methods. *Controlled Clin Trials;***6**:271–279.

Prioleau, L, Murdock, M, Broody, N. (1983) An analysis of psychotherapy versus placebo studies. *Behavioral and Brain Sciences;***6**:275–310.

Probstfield, J, Applegate, WB. (1995) Prospective meta-analysis: ahoy: a clinical trial? [editorial; comment]. *J Am Geriatr Soc;***43**:452–453.

Proskin, HM. (1993) Statistical considerations related to a meta-analytic evaluation of published caries clinical studies comparing the anticaries efficacy of dentifrices containing sodium fluoride and sodium monofluorophosphate. *Am J Dent;***6** Spec No:S43–S49.

Putzrath, RM, Ginevan, ME. (1991) Meta-analysis: methods for combining data to improve quantitative risk assessment. *Regulatory Toxicology and Pharmacology;***14**:178–188.

Raghunathan, TE. (1991) Pooling controls from different studies. *Stat Med;***10**:1417–1426.

Raghunathan, TE, Ii, YC. (1993) Analysis of binary data from a multicenter clinical trial. *Biometrika;***80**:127–139.

Rahman, MI, Chagoury, ME. (1994) Selections from current literature: magnesium, myocardial infarction and meta-analysis. *Fam Pract;***11**:96–101.

Rao, PSRS. (1984) Variance components models for combining estimates. In: Rao, PSRS, Sedransk, J. (eds) WG Cochran's Impact on Statistics, pp. 203–221. New York: Wiley.

Raudenbush, SW. (1983) Utilizing controversy as a source of hypotheses for meta-analysis: The case of teacher expectancy on pupil IQ. In: Light, R.J. (Ed.) Evaluation Studies Review Annual, Beverly Hills: Sage.

Raudenbush, SW. (1994) Random effects models. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 301–322. New York: Russell Sage Foundation.

Raudenbush, SW, Becker, BJ, Kalaian, H. (1988) Modeling multivariate effect sizes. *Psychol Bull;***103**:111–120.

Raudenbush, SW, Bryk, AS. (1985) Empirical Bayes meta-analysis. *J Educ Statist;***10**:75–98.

Raudenbush, SW, Bryk, AS. (1987) Examining correlates of diversity. *J Educ Statist;***12**:241–269.

Ravnskov, U. (1992) Cholesterol lowering trials in coronary heart disease – frequency of citation and outcome. *BMJ;***305**:15–19.

Reed, JG, Baxter, PM. (1994) Using reference databases. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 57–70. New York: Russell Sage Foundation.

Reitman, D, Chalmers, TC, Nagalingam, R, Sacks, H. (1988) Can efficacy of blinding be documented by meta-analysis? Paper presented to the Society for Clinical Trials, San Diego, 23–26 May, 1988.

Rennie, D, Flanagin, A. (1992) Publication bias: The triumph of hope over experience. *JAMA;***267**:411–412.

Richards, SM. (1995) Meta-analyses and overviews of randomised trials (review). *Blood Rev;***9**:85–91.

Rifat, SL. (1990) Graphic representations of effect estimates: an example from a meta- analytic review. *J Clin Epidemiol;***43**:1267–1269.

Risch, HA. (1988) A unified framework for meta-analysis by maximum likelihood. *Am J Epidemiol;***128**:906.

Robins, J, Breslow, N, Greenland, S. (1986) Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics;***42**:311–323.

Robins, J, Greenland, S, Breslow, NE. (1986) A general estimator for the variance of the Mantel–Haenszel odds ratio. *Am J Epidemiol;***124**:719–723.

Rochon, PA, Gurwitz, JH, Cheung, CM, Hayes, JA, Chalmers, TC. (1994) Evaluating the quality of articles published in journal supplements compared with the quality of those published in the parent journal. *JAMA;***272**:108–113.

Rockette, HE. (1991) Statistical issues in carcinogenic risk assessment. *Environmental Health Perspectives;***90**:223–227.

Rogatko, A. (1992) Bayesian approach for metaanalysis of controlled clinical trials. *Communications In Statistics – Theory And Methods;*21:1441–1462.

Rogers, WJ. (1994) What is the optimal tool to define appropriate therapy – the randomized clinical trial, meta-analysis, or outcomes research – commentary. *Curr Opin Cardiol;***9**:401–403.

Rosendaal, FR, Van Everdingen, JJE. (1993) Cumulative metanalyse als ultieme waarheid. Cumulative meta analysis as the ultimate *truth Nederlands Tijdschrift voor Geneeskunde;***137**:1591–1594.

Rosenfeld, RM. (1996) How to systematically review the medical literature. *Otolaryngology – Head And Neck Surgery;***115**:53–63.

Rosenthal, MC. (1994) The fugitive literature. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 85–96. New York: Russell Sage Foundation.

Rosenthal, R. (1978) Combining the results to independent studies. *Profess Psychol;***17**:136–137.

Rosenthal, R. (1979) The file drawer problem and tolerance for null results. *Psychol Bull;***86**:638–641.

Rosenthal, R. (1979) Combining the results of independent studies. *Psychol Bull;***85**:185–193.

Rosenthal, R. (1991) Quality-weighting of studies in meta-analytic research. *Psychotherapy Research;***1**:25–28.

Rosenthal, R. (1991) Meta-analysis: a review. *Psychosomatic Medicine;***53**:247–271.

Rosenthal, R. (1991) Meta-analytic procedures for social research, Revised edition edn. California: Sage.

Rosenthal, R. (1994) Parametric measures of effect size. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 231–244. New York: Russell Sage Foundation.

Rosenthal, R, Rubin, DB. (1979) Comparing significance levels of independent studies. *Psychol Bull;***86**:1165–1168.

Rosenthal, R, Rubin, DB. (1982) Comparing effect sizes of independent studies. *Psychol Bull;***92**:500–504.

Rosenthal, R, Rubin, DB. (1986) Meta-analytic procedures for combining studies with multiple effect sizes. *Psychol Bull;***99**:400–406.

Roth, L, Sackett, PR. (1991) Development and Monte-Carlo evaluation of meta-analytic estimators for correlated data. *Psychol Bull;***110**:318–327.

Rothstein, HR, McDaniel, MA. (1989) Guidelines for conducting and reporting meta-analyses. *Psychological Reports;***65**:759–770.

Rubin, D. (1992) A new perspective. In: Wachter, KW, Straf, ML. (eds) The future of meta-analysis, pp. 155–165. New York: Russell Sage Foundation.

Rubin, DB. (1987) Multiple imputation for non response in surveys, New York: Wiley.

Rutter, CM, Gatsonis, CA. (1996) Regression methods for meta-analysis of diagnostic test data. *Acad Radiol;***2**:S48–S56.

Sackett, DL. (1994) The Cochrane Collaboration. *Ann Intern Med*;**120**:A11.

Sackett, DL. (1994) Cochrane collaboration. *BMJ;***309**:1514–1515.

Sackett, DL. (1996) Applying overviews and meta-analyses at the bedside. *J Clin Epidemiol*

Sackett, DL, Cook, RJ. (1994) Understanding clinical trials. What measures of efficacy should journal articles provide busy clinicians? *BMJ;***309**:755–756.

Sackett, DL, Spitzer, WO. (1994) Guidelines for improving meta-analysis. *Lancet;***343**:910.

Sackett, PR, Harris, MM, Orr, JM. (1986) On seeking moderator variables in the meta-analysis of correlational data: a Monte Carlo investigation of statistical power and resistance to type I error. *Journal of Applied Psychology;***71**:302–310.

Sacks, HS, Berrier, J, Reitman, D, Ancona-Berk, VA, Chalmers, TC. (1987) Meta-analysis of randomized controlled trials. *N Engl J Med;***316**:450–455.

Sacks, HS, Berrier, J, Reitman, D, Pagano, D, Chalmers, TC. (1992) Meta-analyses of randomised control trials: an update of the quality and methodology. In: Bailar, JC, Mosteller, F. (eds) Medical uses of statistics, 2nd ed edn. pp. 427–442. Boston: NEJM Books.

Sacks, HS, Chalmers, TC. (1983) Randomized v historical control trials – reply. *Archives of Internal Medicine;***143**:2342.

Sacks, HS, Chalmers, TC, Smith H Jr. (1982) Randomized versus historical controls for clinical trials. *American Journal of Medicine;***72**:233–240.

Sacks, HS, Reitman, D, Pagano, D, Kupelnick, B. (1996) Meta-analysis – an update. *Mount Sinai Journal Of Medicine;***63**:216–224.

Sagebiel, RW. (1994) The pathology of melanoma as a basis for prognostic models: the UCSF experience. *Pigment Cell Res;***7**:101–103.

Sandercock, P. (1993) Collaborative worldwide overviews of randomized trials (review). *Ann N Y Acad Sci;*149–155.

Sankey, SS, Weissfeld, LA, Fine, MJ, Kapoor, W. (1996) An assessment of the use of the continuity correction for sparse data in metaanalysis. *Communications In Statistics Simulation And Computation;***25**:1031–1056.

Sato, T. (1990) Confidence limits for the common odds ratio based on the asymptotic distribution of the Mantel–Haenszel estimator. *Biometrics;***46**:71–80.

Sawka, CA, Pritchard, KI. (1992) Metaanalysis – its application to clinical medicine. *Diagnostic Oncology;***2**:80–89.

Scherer, RW, Dickersin, K, Kaplan, E. (1994) The accessible biomedical literature represents a fraction of all studies in a field. In: Weeks, RA, Kinser, DL. (eds) Editing the refereed scientific journal: practical, political, and ethical issues, pp. 120–125. New York: IEEE Press.

Scherer, RW, Dickersin, K, Langenberg, P. (1994) Full publication of results initially presented in abstracts: A meta- analysis. *JAMA;***272**:158–162.

Schmid, CH, Cappelleri, JC, Lau, J. (1994) Applying bayesian meta-regression to the study of thrombolytic therapy. *Clinical Research;***42**:A290.

Schmid, HC, McIntosh, MW, Cappelleri, JC, Lau, J, Chalmers, TC. (1995) Measuring the impact of the control rate in meta-analysis of clinical trials. *Controlled Clin Trials;***16**:665* (* We could not find this reference, at least part of the citation is incorrect – it was picked up because McIntosh (1996) cited it).

Schmid, JE, Koch, GG, LaVange, LM. (1991) An overview of statistical issues and methods of meta-analysis. *J Biopharm Stat;***1**:103–120.

Schmidt, F, Hunter, JE. (1995) The impact of data analysis methods on cumulative research knowledge. Statistical significance testing, confidence intervals, and meta-analysis. *Evaluation and the Health Professions;***18**:408–427.

Schmidt, FL, Hunter, JE, Pearlman, K, Hirsh, HR. (1985) Forty questions about validity generalization and meta-analysis. *Personnel Psychology;***38**:697–798.

Schneider, B. (1989) Analysis of clinical trial outcomes: alternative approaches to subgroup analysis. *Controlled Clin Trials;***10**:176S–186S.

Schneider, B. (1992) Ginkgo biloba extract in peripheral arterial disease: meta-analysis of controlled clinical trials. *Arzneimittel-Forschung/Drug Research;*42–1, 428–436.

Schulz, KF, Altman, DG. Statistical Methods for Data Synthesis: Cochrane Collaboration Workshop Report. p. 1–13 (1993).

Schulz, KF, Chalmers, I, Grimes, DA, Altman, DG. (1994) Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA;***272**:125–128.

Schulz, KF, Chalmers, I, Hayes, RJ, Altman, DG. (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA;***273**:408–412.

Schwartz, J. (1993) Beyond LOEL's, *p* values, and vote counting: methods for looking at the shapes and strengths of associations. *NeuroToxicology;***14**:237–246.

Scotti, JR, Evans, IM, Meyer, LH, Walker, P. (1991) A meta-analysis of intervention research with problem behavior: treatment validity and standards of practice. *American Journal on Mental Retardation;***96**:233–256.

Scruggs, TE, Mastropieri, MA. (1994) The utility of the PND statistic: a reply to Allison and Gorman [comment]. *Behav Res Ther;***32**:879–883.

Scruggs, TE, Mastropieri, MA, Casto, G. (1987) The quantitative synthesis of single subject research: methodology and validation. *Remedial and Special Education;***8**:24–33.

Sechrest, L, Hannah, M. (1996) The critical importance of non-experimental data. In: Sechrest, L, Perrin, J, Bunker, J. (eds) Research methodology: strengthening causal interpretations of non-experimental data. AHCBR Conference Proceedings. Washington DC: Agency for Health.

Seltzer, M. (1991) The use of data augmentation in fitting hierarchical models to education data. Anonymous. Unpublished doctorial dissertation.

Senn, S. (1994) Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med;***13**:293–296.

Senn, S. (1996) Meta-analysis with Mathcad. *ISCB News;*(20):4–5.

Senn, S. (1996) Relation between treatment benefit and underlying risk in metaanalysis – standard of label invariance should not abandoned. *BMJ;***313**:1550.

Senn, S, Harrel, F. (1997) On wisdom after the event. *J Clin Epidemiol*

Shadish, WR, Haddock, CK. (1994) Combining estimates of effect size. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 261–284. New York: Russell Sage Foundation.

Shapiro, DA, Shapiro, D. (1982) Meta-analysis of comparative therapy outcome studies: a replication and refinement. *Psychol Bull;***92**:581–604.

Shapiro, DA, Shapiro, D. (1983) Comparative therapy outcome research: methodological implications of meta-analysis. *J Consult Clin Psychol;***51**:42–53.

Shapiro, DE. (1995) Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol;***2**:S37–S47.

Shapiro, S. (1995) Point Counterpoint: meta-analysis of observational studies – reply. *Am J Epidemiol;***142**:780–781.

Sharp, S, Thompson, S, Altman, D. (1996) Relation between treatment benefit and underlying risk in metaanalysis – standard of label invariance should not abandoned – reply. *BMJ;***313**:1550–1551.

Sharp, SJ, Thompson, SG, Altman, DG. (1996) The relation between treatment benefit and underlying risk in metaanalysis. *BMJ;***313**:735–738.

Sheehe, PR. (1966) Combination of log relative risk in retrospective studies of disease. *Am J Public Health;***56**:1745–1750.

Sheldon, T, Chalmers, I. (1994) The UK Cochrane Centre and the NHS. *Health Economics;***3**:201–203.

Sheldon, TA. (1994) Please bypass the PORT: observational studies of effectiveness run a poor second to randomised controlled trials. *BMJ;***309**:142–143.

Silagy, C. (1993) Developing a register of randomised controlled trials in primary care. *BMJ;***306**:897–900.

Silagy, CA, Jewell, D. (1994) Review of 39 years of randomized controlled trials in the British Journal of General Practice. *British Journal of General Practice;***44**:359–363.

Silagy, CA, Jewell, D, Mant, D. (1994) An analysis of randomized controlled trials published in the US Family Medicine Literature, 1987–1991. *Journal of Family Practice;***39**:236–242.

Silberberg, JS. (1994) After the meta-analyses: a commentary on treatment of dyslipidaemia in the primary prevention of coronary heart disease [see comments]. *Aust N Z J Med;***24**:717–721.

Silcocks, P. (1994) Experiment and observation. *Lancet;***344**:1775–1776.

Sim, I, Hlatky, MA. (1996) Growing pains of metaanalysis – advances in methodology will not remove the need for well designed trials. *BMJ;***313**:702–703.

Simes, RJ. (1986) Publication bias: the case for an international registry of clinical trials. *J Clin Oncol;***4**:1529–1541.

Simes, RJ. (1987) Confronting publication bias: a cohort design for meta-analysis. *Stat Med;*__6__:11–29.

Simes, RJ. (1992) Meta-analysis and quality of evidence in the economic evaluation of drug trials. *Pharmacoeconomics;*__1__:282–292.

Simon, R. (1987) Overviews of randomised clinical trials. *Cancer Treat Rev;*__71__:3–5.

Simon, R. (1994) Randomized clinical trials in oncology. Principles and obstacles (review). *Cancer;*__74__:2614–2619.

Sinclair, JC, Bracken, MB. (1994) Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol;*__47__:881–889.

Skene, AM, Wakefield, JC. (1990) Hierarchical models for multicentre binary response studies. *Stat Med;*__9__:919–929.

Slavin, RE. (1986) Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educ Res;*__15__:5–11.

Slavin, RE. (1995) Best evidence synthesis: an intelligent alternative to meta- analysis (review). *J Clin Epidemiol;*__48__:9–18.

Smith, AFM. (1973) Bayes estimates in one-way and two-way models. *Biometrika;*__60__:319–329.

Smith, DD, Givens, GH, Tweedie, RL. (1997) Adjustment for publication and quality bias in Bayesian meta-analysis.

Smith, GD, Song, F, Sheldon, TA, Song, FJ. (1993) Cholesterol lowering and mortality: the importance of considering initial level of risk. *BMJ;*__306__:1367–1373.

Smith, ML. (1980) Publication bias and meta-analysis. *Evaluation in Education;*__4__:22–24.

Smith, ML, Glass, GV. (1977) Meta-analysis of psychotherapy outcome studies. *American Psychologist;*__32__:752–760.

Smith, ML, Glass, GV, Miller, TI. (1980) The benefits of psychotherapy, Baltimore, MD: John Hopkins University Press.

Smith, SJ, Caudill, SP, Steinberg, KK, Thacker, SB. (1995) On combining dose-response data from epidemiological studies by meta-analysis. *Stat Med;*__14__:531–544.

Smith, TC. (1995) Interpreting evidence from multiple randomised and non-randomised studies. Unpublished Thesis, University of Cambridge.

Smith, TC, Abrams, KR, Jones, DR. (1995) Using hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. Department of Epidemiology and Public Health Technical Report. University of Leicester.

Smith, TC, Spiegelhalter, D, Parmar, MKB. (1995) Bayesian meta-analysis of randomized triles using graphical models and BUGS. In: Anonymous Bayesian Biostatistics, pp. 411–427.

Smith, TC, Spiegelhalter, DJ, Thomas, A. (1995) Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med;*2685–2699.

Smith, TC, Abrams, KR, Jones, DR. (1996) Assessment of prior distributions and model parameterisation in hierarchical models for the generalised synthesis of evidence. Technical Report #96–01, Department of Epidemiology and Public Health: University of Leicester, England.

Solari, ME, Wheatley, D. (1966) A method of combining the results of several clinical trials. *Clinical Trials Journal;*537–545.

Solomon, MJ, Laxamana, A, Devore, L, McLeod, RS. (1994) Randomized controlled trials in surgery. *Surgery;*__115__:707–712.

Sommer, B. (1987) The file drawer effect and publication rates in menstrual cycle research. *Psychology of Women Quarterly;*__11__:233–242.

Soreide, E, Steen, PA. (1993) Dangers of review articles. *BMJ;*__306__:66–67.

Soreide, E, Stromskag, KE, Steen, PA. (1995) Statistical aspects in studies of preoperative fluid intake and gastric content. *Acta Anaesthesiologica Scandinavica;*__39__:738–743.

Sorofman, BA, Milavetz, G. (1994) Meta-analysis and the scientific research that precedes it [editorial; comment]. *Ann Pharmacother;*__28__:1290–1291.

Soto, J, Galende, I, Sacristan, JA. (1994) [The quality of clinical trials published in Spain: an evaluation by an analysis of 3 journals during the 1985–1991 period]. [Spanish]. *Med Clin (Barc);*__102__:241–245.

Spector, PE, Levine, EL. (1987) Meta-analysis for integrating study outcomes: a Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology;*__72__:3–9.

Spector, TD, Thompson, SG. (1991) Research methods in epidemiology .5. the potential and limitations of metaanalysis. *J Epidemiol Comm Hlth;*__45__:89–92.

Spiegelhalter, D, Thomas, A, Gilks, W. (1994) BUGS Examples 0.30.1, Cambridge: MRC Biostatistics Unit.

Spitzer, WO. (1991) Meta-meta-analysis: unanswered questions about aggregating data. *J Clin Epidemiol;*__44__:103–107.

Spitzer, WO. (1995) The challenge of meta-analysis. *J Clin Epidemiol;*__48__:1–4.

Spitzer, WO. (1995) The Potsdam international consultation on meta-analysis. *J Clin Epidemiol;*__48__:1–171.

Spitzer, WO, Lawrence, V, Dales, R. (1990) Links between passive smoking and disease: a best evidence synthesis: a report of the working group on passive smoking. *Clin Invest Med;*__13__:17–42.

Spoor, P, Airey, M, Bennet, C, Greensill, J, Williams, R. (1996) Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. *BMJ;*__313__:342–343.

Srinivasan, C, Zhou, M. (1993) A note on pooling Kaplan–Meier estimators. *Biometrics;***49**:861–864.

Standards of Reporting Trials Group (1994) A proposal for structured reportion of randomized controlled trials. *JAMA;***272**:1926–1931.

Stangl, DK. (1995) Prediction and decision making using Bayesian hierarchical models. *Stat Med;***14**:2173–2190.

Stein, RA. (1988) Meta-analysis from one FDA reviewer's perspective. *Proc Biopharmaceutical Section of the American Statistical Association;*34–38.

Steinberg, KK, Smith, SJ, Thacker, SB, Stroup, DF. (1994) Breast cancer risk and duration of estrogen use: the role of study design in meta-analysis. *Epidemiology;***5**:415–421.

Sterling, TD, Rosenbaum, WL, Weinkam, JJ. (1995) Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician;***49**:108–112.

Stewart, LA, Clarke, MJ. (1995) Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Stat Med;***14**:2057–2079.

Stewart, LA, Parmar, MK. (1993) Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet;***341**:418–422.

Stewart, LA, Parmar, MKB. (1996) Bias in the analysis and reporting of randomized controlled trials. *International Journal of Technology Assessment in Health Care;***12**:264–275.

Stijnen, T, Van Houwelingen, JC. (1990) Empirical Bayes methods in clinical trials meta-analysis. *Biometrical Journal;***32**:335–346.

Stjernsward, J. (1974) Decreased survival related to irradiation postoperatively in early operable breast cancer. *Lancet;*1285–1286.

Stock, WA. (1994) Systematic coding for research synthesis. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 125–138. New York: Russell Sage Foundation.

Stock, WA, Okun, MA, Haring, MJ, Miller, W, Ceurvost, RW. (1992) Rigor in data synthesis: a case study of reliability in meta-analysis. *Educ Res;*June–July:10–14.

Stockler, M, Coates, A. (1993) What have we learned from meta-analysis? *Med J Aust;***159**:291–293.

Stouffer, SA, Suchman, EA, DeVinney, LC, Star, SA, Williams, RM Jr. (1949) The American soldier: Adjustment during army life (Vol.1), Princeton, NJ: Princeton University Press.

Stram, DO. (1996) Meta-analysis of published data using a linear mixed-effects model. *Biometrics;***52**:536–544.

Strube, JJ, Hartmann, DP. (1983) Meta-analysis: Techniques, applications, and functions. J Consult Clin Psychol 51, 14–27.

Strube, MJ. (1985) Combining and comparing significance levels from nonindependent hypothesis tests. *Psychol Bull;***97**:334–341.

Strube, MJ, Gardner, W, Hartmann, DP. (1985) Limitations, liabilities, and obsticles in reviews of the literature: the current status of meta-analysis. *Clinical Psychology Review;***5**:63–78.

Strube, MJ, Hartman, DP. (1982) A critical appraisal of meta-analysis. *Br J Clin Psychol;***21**:129–139.

Su, XY, Po, ALW. (1996) Combining event rates from clinical trials: comparison of Bayesian and classical methods. *Annals of Pharmacotherapy;***30**:460–465.

Subtil, D, Truffert, P, Vinatier, D, Puech, F, Querleu, D, Crepin, G. (1994) Interets et limites des meta-analyses – Classiques ou cumulatives – en gynecologie-obstetrique. Usefulness and limits of conventional and comulative meta-analysis in gynaecology– obstetrics. *Therapie;***49**:175–179.

Sugita, M, Kanamori, M, Izuno, T, Miyakawa, M. (1992) Estimating a summarized odds ratio whilst eliminating publication bias in meta-analysis. *Jpn J Clin Oncol;***22**:354–358.

Sugita, M, Yamaguchi, N, Izuno, T, Kanamori, M, Kasuga, H. (1994) Publication probability of a study on odds ratio value circumstantial evidence for publication bias in medical study areas. *Tokai J Exp Clin Med;***19**:29–37.

Swales, JD. (1993) Meta-analysis as a guide to clinical practice. [review]. *J Hypertens Suppl;***11**:S59–S63.

Swales, JD. (1994) Meta-analyses: how seriously should cardiologigsts take them? *Journal of Myocardial Ischemia;***6**:10–14.

Sylvester, R. (1995) EORTC Data Centre receives grants under the EU's 4th framework RTD program (BIOMED 2) for meta-analysis and health economic studies. *Eur J Cancer;***31A**:1914.

Tabak, ER, Mullen, PD, SimonsMorton, DG, Green, LW, Mains, DA, EilatGreenberg, S, *et al.* (1991) Definition and yield of inclusion criteria for a meta-analysis of patient education studies in clinical preventive services. *Evaluation and the Health Professions;***14**:388–411.

Talwalker, S. (1996) Analysis of repeated measurements with dropouts among Alzheimer's disease patients using summary measures and meta-analysis. *J Biopharm Stat;***6**:49–58.

Tarone, RE. (1981) On summary estimators of relative risk. *Journal of Chronic Diseases;***34**:463–468.

Taylor, JM. (1989) Models for the HIV infection and AIDS epidemic in the United States. *Stat Med;***8**:450–458.

Thacker, SB. (1988) Meta-analysis. A quantitative approach to research integration. *JAMA;***259**:1685–1689.

Thacker, SB, Hoffman, DA, Smith, J, Steinberg, K, Zack, M. (1992) Effect of low-level body burdens of lead on the mental development of children: limitations of meta-analysis in a review of longitudinal data. *Arch Environ Health;***47**:336–346.

Thacker, SB, Peterson, HB, Stroup, DF. (1996) Metaanalysis for the obstetrician-gynecologist. American *Journal of Obstetrics and Gynecology;***174**:1403–1407.

The Cochrane Collaboration 1997. The Cochrane Database of Systematic Reviews [database on disk and CD ROM]. London: BMJ Publishing Group.

The NHS Centre for Reviews and Dissemination 1997. The Database of Abstracts of Reviews of Effectiveness [database on line]. York: University of York.

Thompson, SG. (1993) Controversies in meta-analysis: the case of the trials of serum cholesterol reduction (review). *Stat Methods Med Res;*2:173–192.

Thompson, SG. (1994) Why sources of heterogeneity in meta-analysis should be investigated. [review]. *BMJ;***309**:1351–1355.

Thompson, SG, Pocock, SJ. (1991) Can meta-analyses be trusted? *Lancet;***338**:1127–1130.

Thompson, SG, Smith, TC, Sharp, SJ. (1997) Investigation underlying risk as a source of heterogeneity in meta-analysis. *Stat Med:* in press.

Tippett, LHC. (1931) The methods of statistics, 1st edn. London: Williams & Norgate.

Tori, V, Simon, R, Russek-Cohen, E, Midthune, D, Friedman, M. (1992) Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. *J Natl Cancer Inst;***84**:407–414.

Tucker, K. (1996) The use of epidemiologic approaches and metaanalysis to determine mineral element requirements. *Journal Of Nutrition;***126**:S2365–S2372.

Tweedie, RL, Mengersen, KL. (1992) Lung cancer and passive smoking: reconciling the biochemical and epidemiological approaches. *Br J Cancer;***66**:700–705.

Tweedie, RL, Mengersen, KL. (1995) Meta-analytic approaches to dose-response relationships, with application in studies of lung cancer and exposure to environmental tobacco smoke. *Stat Med;***14**:545–569.

Tweedie, RL, Scott, DJ, Biggerstaff, BJ, Mengersen, KL. (1996) Bayesian meta-analysis, with application to studies of ETS and lung cancer. *Lung Cancer;***14**:S171–S194.

US Preventive Services Task Force (1996) Guide to clinical preventive services: an assessment of the affectiveness of 169 interventions. Baltimore, Md: Williams & Wilkins.

Van der Linden, S, Goldsmith, CH, Woodcock, J, Nassonova, V. (1994) Can observational studies replace or complement experiment? *Journal of Rheumatology;***21**:57–61.

Van der Wijden, CL, Overbeke, AJPM. (1993) Gerandomiseerde klinische trials in het. *Nederlands Tijdschrift voor Geneeskund;***137**:1607–1610.

Van Houwelingen, HC, Zwinderman, KH, Stijnen, T. (1993) A bivariate approach to meta-analysis. *Stat Med;***12**:2273–2284.

Vandenbroucke, JP. (1988) Passive smoking and lung cancer: a publication bias? *BMJ;***296**:391–392.

Vandenbroucke, JP. (1995) Re: 'Invited commentary: a critical look at some popular meta-analytic methods' (letter; comment). *Am J Epidemiol;***142**:1007–1009.

Vanhonacker, WR. (1996) Meta-analysis and response surface extrapolation: a least squares approach. *American Statistician;***50**:294–299.

Vanhouwelingen, HC. (1995) Meta-analysis; methods, limitations. *Byocybernetics and Biomedical Engineering;***15**:53–61.

Vanhouwelingen, HC, Stijnen, T. (1993) Monotone empirical bayes estimators based on more informative samples. *J Am Statist Assoc;***88**:1438–1443.

Velanovich, V. (1991) Meta-analysis for combining Bayesian probabilities. *Medical Hypotheses;***35**:192–195.

Veldhuyzen van Zanten, SJ, Boers, M. (1993) Metanalyse: de kunst van het systematisch oversicht. *Ned Tijdschr Geneeskd;***137**:1594–1599.

Verdinelli, I, Andrews, K, Detre, K, Peduzzi, P. (1995) The Bayesian approach to meta-analysis: a case study. Departmet of Statistics, Carnegie Mellon University.

Veronese, P. (1994) Mutually compatible hierarchical priors for combining information. Draft Form.

Victor, N. (1995) ''The challenge of meta-analysis": Discussion. Indications and contraindications for meta-analysis. *J Clin Epidemiol;***48**:5–8.

Villar, J, Carroli, G, Belizan, JM. (1995) Predictive ability of meta-analyses of randomised controlled trials. *Lancet;***345**:772–776.

Vital Durand, D. (1994) [Problems raised by the conducting and interpretation of meta-analysis]. [French]. *Therapie;***49**:165–168.

Voest, EE, Van Houwelingen, JC, Neijt, JP. (1989) A meta-analysis of prognostic factors in advanced ovarian cancer with median survival and overall survival (measured with the log (relative risk)) as main objectives. *European Journal of Cancer and Clinical Oncology;***25**:711–720.

Wachter, KW. (1988) Disturbed by meta-analysis? *Science;***241**:1407–1408.

Waclawiw, MA, Liang, KY. (1994) Empirical bayes estimation and inference for the random effects model with binary response. *Stat Med;***13**:541–551.

Wakeford, R, Roberts, W. (1993) Using Medline for comprehensive searches. *BMJ;***306**:1415.

Wald, NJ, Nanchahal, K, Thompson, SG, Cuckle, HS. (1986) Does breathing other people's tobacco smoke cause lung cancer? *BMJ;***293**:1217–1222.

Walker, AM, Martin-Moreno, JM, Artalejo, FR. (1988) Odd man out: a graphical approach to meta-analysis. *Am J Public Health;***78**:961–966.

Ward, MM, Leigh, JP. (1993) Pooled time series regression analysis in longitudinal studies. *J Clin Epidemiol;***46**:645–659.

Waternaux, C, DuMouchel, W. (1993) Combining information across sites with hierarchical Bayesian linear models. Proceedings of the Section on Bayesian Statistics. San Francisco.

West, RR. (1993) A look at the statistical overview (or meta-analysis) [published erratum appears in J *R Coll Physicians Lond* 1993 Jul;**27**(3):223]. *J R Coll Physicians Lond;***27**:111–115.

White, DM, Rusch, FR, Kazdin, AE, Hartmann, DP. (1989) Applications of meta analysis inindividual subject research. *Behavioral Assessment;***11**:281–296.

White, HD. (1994) Scientific communication and literature retrieval. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 41–56. New York: Russell Sage Foundation.

White, HD, Bates, MJ, Wilson, P. (1992) For information specialists: interpretations of reference and bibliographic work, Norwood, NJ: Ablex.

White, OR. (1987) Some comments concerning 'The quantitative synthesis of single-subject research'. *Remedial and Special Education;***8**:34–39.

Whitehead, A. (1994) Multicentre trials versus meta-analysis. *European Journal of Clinical Research;***6**:37–42.

Whitehead, A. (1997) A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Stat Med;***16** (fall 1997).

Whitehead, A. (1997) Book Review: The handbook of research synthesis. *Stat Med;***16**:713–714.

Whitehead, A, Bailey, AJ, Elbourne, D. (1997) Combining summaries of binary outcomes with those of continuous outcomes in a meta-analysis. Anonymous.

Whitehead, A, Jones, NMB. (1994) A meta-analysis of clinical trials involving different classifications of response into ordered categories. *Stat Med;***13**:2503–2515.

Whitehead, A, Whitehead, J. (1991) A general parametric approach to the meta-analysis of randomised clinical trials. *Stat Med;***10**:1665–1677.

Whiting, GW, Lau, J, Kupelnick, B, Chalmers, TC. (1995) Trends in inflammatory bowel disease therapy: a meta-analytic approach. *Canadian Journal of Gastroenterology;***9**:405–411.

Wieland, D, Stuck, AE, Siu, AL, Adams, J, Rubenstein, LZ. (1995) Meta-analytic methods for health services research: an example from geriatrics. *Evaluation and the Health Professions;***18**:252–282.

Williams, DH, Davis, CE. (1994) Reporting of assignment methods in clinical trials. *Controlled Clin Trials;***15**:294–298.

Wilson, A, Henry, DA. (1992) Meta-analysis. Part 2: Assessing the quality of published meta-analyses (review). *Med J Aust;***156**:173–184.

Wilson, TG, Rachman, SJ. (1983) Meta-analysis and the evaluation of psycho-therapy outcome: limitations and liabilities. *J Consult Clin Psychol;***51**:54–64.

Wittes, RE. (1987) Problems in the medical interpretation of overviews. *Stat Med;***6**:269–280.

Wolf, FM. (1986) Meta-analysis. Quantitative methods for research synthesis, Beverly Hills: Sage Publications.

Wolf, FM. (1993) Problem-based learning and meta-analysis: can we see the forest through the trees? [comment]. *Academic Medicine;***68**:542–544.

Wolpert, RL, Warren-Hicks, WJ. (1991) Bayesian hierarchical logistic models for combining field and laboratory survival data. In conference form.

Wong, O, Raabe, GK. (1990) Proper interpretation of meta-analysis in occupational epidemiologic studies: a response. *Am J Ind Med;***17**:273–276.

Wong, O, Raabe, GK. (1996) Application of metaanalysis in reviewing occupational cohort studies. *Occupational and Environmental Medicine;***53**:793–800.

Woodworth, G. (1994) Managing meta-analytic databases. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 177–190. New York: Russell Sage Foundation.

Woolf, SH, Battista, RN, Anderson, GM, Logan, AG, Wang, E, Canadian Task Force on the Periodic Health Examination. (1990) Assessing the clinical effectiveness of preventive maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol;***43**:891–905.

Wortman, PM. (1983) Evaluation research: a methodological perspective. *Annual Review of Psychology;***34**:223–260.

Wortman, PM. (1987) Meta-analysis (letter to editor). *N Engl J Med;***317**:575.

Wortman, PM. (1992) Lessons from meta-analysis of quasi-experiments. In: Bryant, FB. (ed) Methodological issues in applied social psychology, pp. 65–81. New York: Plenum.

Wortman, PM. (1994) Judging research quality. In: Cooper, H, Hedges, LV. (eds) The handbook of research synthesis, pp. 97–110. New York: Russell Sage Foundation.

Wortman, PM, Yeaton, WH. (1987) Using research synthesis in medical technology assessment. *International Journal of Technology Assessment in Health Care;***3**:509–522.

Yach, D. (1990) Meta-analysis in epidemiology. *S Afr Med J;***78**:94–97.

Yates, F, Cochran, WG. (1938) The analysis of groups of experiments. *Journal of Agricultural Sciences;***28**:556–580.

Yeaton, WH, Langenbrunner, JC, Smyth, JM, Wortman, PM. (1995) Exploratory research synthesis: methodological considerations for addressing limitations in data quality. *Evaluation And The Health Professions;***18**:283–303.

Yeaton, WH, Wortman, PM. (1994) On the reliability of meta-analytic reviews: the role of intercoder agreement. Evaluation Review.

Yeaton, WH, Wortman, PM, Langberg, N. (1983) Differential attrition: estimating the effect of crossovers on the evaluation of a medical technology. *Evaluation Review;***7**:831–840.

Yusuf, S. (1987) Obtaining medically meaningful answers from an overview of randomized clinical trials. *Stat Med;***6**:281–294.

Yusuf, S, Peto, R, Lewis, J, Collins, R, Sleight, P, *et al* (1985) Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Progress in Cardiovascular Diseases;***27**:335–371.

Yusuf, S, Simon, R, Ellenberg, SS. (1987) Proceedings of methodologic issues in overviews of randomized clinical trials – preface. *Stat Med;***6**:217–218.

Yusuf, S, Wittes, J, Probstfield, J, Tyroler, HA. (1991) Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA;***266**:93–98.

Zelen, M. (1971) The analysis of several 2 x 2 contingency tables. *Biometrika;***58**:129–137.

Zhang, WY. (1997) Pooling RCTs: outline of procedures and statistical issues. Anonymous.

Zhang, ZY. (1994) On improving omnibus tests in meta-analysis using vote-counts. *Communications In Statistics – Simulation And Computation;*23:803–812.

Zhou, XH. (1996) Empirical Bayes combination of estimated areas under ROC curves using estimating equations. *Med Decis Making;***16**:24–28.

Zimmerman, S. (1992) Considerations for the interpretation of findings from single and multiple studies of periodontal conditions. *Journal Of Periodontal Research;***27**:425–432.

Zwarenstein, M, Volmink, J, Irwig, L, Chalmers, I. (1995) Systematic review – state of the science health-care decision-making. *S Afr Med J;***85**:1266–1267.

Health Technology Assessment panel membership

This report was identified as a priority by the Methodology Panel.

Acute Sector Panel

Chair: Professor John Farndon, University of Bristol †

Diagnostics and Imaging Panel

Chair: Professor Mike Smith, University of Leeds †

Methodology Panel

Chair: Professor Martin Buxton, Brunel University †

* Previous Chair
† Current members

continued

continued

Pharmaceutical Panel

Chair: Professor Tom Walley, University of Liverpool †

Professor Michael Rawlins, University of Newcastle-upon-Tyne*

Dr Colin Bradley, University of Birmingham

Professor Alasdair Breckenridge, RDRD, Northwest Regional Health Authority

Ms Christine Clark, Hope Hospital, Salford †

Mrs Julie Dent, Ealing, Hammersmith & Hounslow Health Authority, London

Mr Barrie Dowdeswell, Royal Victoria Infirmary, Newcastle-upon-Tyne

Dr Tim Elliott, Department of Health †

Dr Desmond Fitzgerald, Mere, Bucklow Hill, Cheshire

Dr Felicity Gabbay, Transcrip Ltd †

Dr Alistair Gray, Health Economics Research Unit, University of Oxford †

Professor Keith Gull, University of Manchester

Dr Keith Jones, Medicines Control Agency

Professor Trevor Jones, ABPI, London †

Ms Sally Knight, Lister Hospital, Stevenage †

Dr Andrew Mortimore, Southampton & SW Hants Health Authority †

Mr Nigel Offen, Essex Rivers Healthcare, Colchester †

Dr John Posnett, University of York

Mrs Marianne Rigge, The College of Health, London †

Mr Simon Robbins, Camden & Islington Health Authority, London †

Dr Frances Rotblat, Medicines Control Agency †

Mrs Katrina Simister, Liverpool Health Authority †

Dr Ross Taylor, University of Aberdeen †

Dr Tim van Zwanenberg, Northern Regional Health Authority

Dr Kent Woods, RDRD, Trent RO, Sheffield †

Population Screening Panel

Chair: Professor Sir John Grimley Evans, Radcliffe Infirmary, Oxford †

Dr Sheila Adam, Department of Health*

Ms Stella Burnside, Altnagelvin Hospitals Trust, Londonderry †

Dr Carol Dezateux, Institute of Child Health, London †

Dr Anne Dixon Brown, NHS Executive, Anglia & Oxford †

Professor Dian Donnai, St Mary's Hospital, Manchester †

Dr Tom Fahey, University of Bristol †

Mrs Gillian Fletcher, National Childbirth Trust †

Professor George Freeman, Charing Cross & Westminster Medical School, London

Dr Mike Gill, Brent & Harrow Health Authority †

Dr JA Muir Gray, RDRD, Anglia & Oxford RO †

Dr Anne Ludbrook, University of Aberdeen †

Professor Alexander Markham, St James's University Hospital, Leeds †

Professor Theresa Marteau, UMDS, London

Dr Ann McPherson, General Practitioner, Oxford †

Professor Catherine Peckham, Institute of Child Health, London

Dr Connie Smith, Parkside NHS Trust, London

Dr Sarah Stewart-Brown, University of Oxford †

Ms Polly Toynbee, Journalist †

Professor Nick Wald, University of London †

Professor Ciaran Woodman, Centre for Cancer Epidemiology, Manchester

Primary and Community Care Panel

Chair: Dr John Tripp, Royal Devon & Exeter Healthcare NHS Trust †

Professor Angela Coulter, King's Fund, London *

Professor Martin Roland, University of Manchester *

Dr Simon Allison, University of Nottingham

Mr Kevin Barton, East London & City Health Authority †

Professor John Bond, University of Newcastle-upon-Tyne †

Ms Judith Brodie, Age Concern, London †

Dr Nicky Cullum, University of York †

Professor Shah Ebrahim, Royal Free Hospital, London

Mr Andrew Farmer, Institute of Health Sciences, Oxford †

Ms Cathy Gritzner, The King's Fund †

Professor Andrew Haines, RDRD, North Thames Regional Health Authority

Dr Nicholas Hicks, Oxfordshire Health Authority †

Professor Richard Hobbs, University of Birmingham †

Professor Allen Hutchinson, University of Sheffield †

Mr Edward Jones, Rochdale FHSA

Professor Roger Jones, UMDS, London

Mr Lionel Joyce, Chief Executive, Newcastle City Health NHS Trust

Professor Martin Knapp, London School of Economics & Political Science

Dr Phillip Leech, Department of Health †

Professor Karen Luker, University of Liverpool

Professor David Mant, NHS Executive South & West †

Dr Fiona Moss, Thames Postgraduate Medical and Dental Education †

Professor Dianne Newham, King's College London

Professor Gillian Parker, University of Leicester †

Dr Robert Peveler, University of Southampton †

Dr Mary Renfrew, University of Oxford

Ms Hilary Scott, Tower Hamlets Healthcare NHS Trust, London †

* Previous Chair
† Current members

National Coordinating Centre for Health Technology Assessment, Advisory Group

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

ISSN 1366-5278