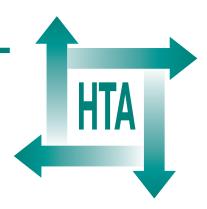# Review of guidelines for good practice in decision-analytic modelling in health technology assessment

Z Philips, L Ginnelly, M Sculpher, K Claxton,
S Golder, R Riemsma, N Woolacott and
J Glanville

September 2004

**Health Technology Assessment
NHS R&D HTA Programme**

HTA

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   Email: orders@hta.ac.uk
c/o Direct Mail Works Ltd   Tel: 02392 492 000
4 Oakwood Business Centre   Fax: 02392 478 555
Downley, HAVANT PO9 2NP, UK   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.

Review of guidelines for good practice in decision-analytic modelling in health technology assessment

Z Philips,[1] L Ginnelly,[1*] M Sculpher,[1] K Claxton,[1,2] S Golder,[3] R Riemsma,[3] N Woolacott[3] and J Glanville[3]

[1] Centre for Health Economics, University of York, UK
[2] Department of Economics, University of York, UK
[3] Centre for Reviews and Dissemination, University of York, UK

* Corresponding author

Published September 2004

NHS R&D HTA Programme

The research findings from the NHS R&D Health Technology Assessment (HTA) Programme directly influence key decision-making bodies such as the National Institute for Clinical Excellence (NICE) and the National Screening Committee (NSC) who rely on HTA outputs to help raise standards of care. HTA findings also help to improve the quality of the service in the NHS indirectly in that they form a key component of the 'National Knowledge Service' that is being developed to improve the evidence of clinical practice throughout the NHS.

The HTA Programme was set up in 1993. Its role 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 provide care in the NHS. 'Health technologies' are broadly defined to include all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care, rather than settings of care.

The HTA programme commissions research only on topics where it has identified key gaps in the evidence needed by the NHS. Suggestions for topics are actively sought from people working in the NHS, the public, consumer groups and professional bodies such as Royal Colleges and NHS Trusts.

Research suggestions are carefully considered by panels of independent experts (including consumers) whose advice results in a ranked list of recommended research priorities. The HTA Programme then commissions the research team best suited to undertake the work, in the manner most appropriate to find the relevant answers. Some projects may take only months, others need several years to answer the research questions adequately. They may involve synthesising existing evidence or designing a trial to produce new evidence where none currently exists.

Additionally, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme is able to commission bespoke reports, principally for NICE, but also for other policy customers, such as a National Clinical Director. TARs bring together evidence on key aspects of the use of specific technologies and usually have to be completed within a limited time period.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

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.

The research reported in this monograph was commissioned and funded by the HTA Programme on behalf of NICE as project number 02/32/01. The authors have been wholly responsible for all data collection, analysis and interpretation and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme, NICE or the Department of Health.

HTA Programme Director: Professor Tom Walley
Series Editors: Dr Peter Davidson, Professor John Gabbay, Dr Chris Hyde,
   Dr Ruairidh Milne, Dr Rob Riemsma and Dr Ken Stein
Managing Editors: Sally Bailey and Caroline Ciupek

T

Abstract

Review of guidelines for good practice in decision-analytic modelling in health technology assessment

Z Philips,[1] L Ginnelly,[1]* M Sculpher,[1] K Claxton,[1,2] S Golder,[3] R Riemsma,[3] N Woolacott[3] and J Glanville[3]

[1] Centre for Health Economics, University of York, UK

[2] Department of Economics, University of York, UK

[3] Centre for Reviews and Dissemination, University of York, UK

* Corresponding author

Objectives: To identify existing guidelines and develop a synthesised guideline plus accompanying checklist. In addition to provide guidance on key theoretical, methodological and practical issues and consider the implications of this research for what might be expected of future decision-analytic models.

Data sources: Electronic databases.

Review methods: A systematic review of existing good practice guidelines was undertaken to identify and summarise guidelines currently available for assessing the quality of decision-analytic models that have been undertaken for health technology assessment. A synthesised good practice guidance and accompanying checklist was developed. Two specific methods areas in decision modelling were considered. The first method's topic is the identification of parameter estimates from published literature. Parameter searches were developed and piloted using a case-study model. The second topic relates to bias in parameter estimates; that is, how to adjust estimates of treatment effect from observational studies where there are risks of selection bias. A systematic literature review was conducted to identify those studies looking at quantification of bias in parameter estimates and the implication of this bias.

Results: Fifteen studies met the inclusion criteria and were reviewed and consolidated into a single set of brief statements of good practice. From this, a checklist was developed and applied to three independent decision-analytic models. Although the checklist provided excellent guidance on some key issues for model evaluation, it was too general to pick up on the specific nuances of each model. The searches that were developed helped to identify important data for inclusion in the model. However, the quality of life searches proved to be problematic: the published search filters did not focus on those measures specific to cost-effectiveness analysis and although the strategies developed as part of this project were more successful few data were found. Of the 11 studies meeting the criteria on the effect of selection bias, five concluded that a non-randomised trial design is associated with bias and six studies found 'similar' estimates of treatment effects from observational studies or non-randomised clinical trials and randomised controlled trials (RCTs). One purpose of developing the synthesised guideline and checklist was to provide a framework for critical appraisal by the various parties involved in the health technology assessment process. First, the guideline and checklist can be used by groups that are reviewing other analysts' models and, secondly, the guideline and checklist could be used by the various analysts as they develop their models (to use it as a check on how they are developing and reporting their analyses). The Expert Advisory Group (EAG) that was convened to discuss the potential role of the guidance and checklist felt that, in general, the guidance and checklist would be a useful tool, although the checklist is not meant to be used exclusively to determine a model's quality, and so should not be used as a substitute for critical appraisal.

Conclusions: The review of current guidelines showed that although authors may provide a consistent message regarding some aspects of modelling, in other areas conflicting attributes are presented in different guidelines. In general, the checklist appears to perform well, in terms of identifying those aspects of the model that should be of particular concern to the reader. The checklist cannot, however, provide answers to the appropriateness of the model structure and structural

assumptions, as these may be seen as a general problem with generic checklists and do not reflect any shortcoming with the synthesised guidance and checklist developed here. The assessment of the checklist, as well as feedback from the EAG, indicated the importance of its use in conjunction with a more general checklist or guidelines on economic evaluation. Further methods research into the following areas would be valuable: the quantification of selection bias in non-controlled studies and in controlled observational studies; the level of bias in the different non-RCT study designs; a comparison of results from RCTs with those from other non-randomised studies; assessment of the strengths and weaknesses of alternative ways to adjust for bias in a decision model; and how to prioritise searching for parameter estimates.

Contents

Glossary and list of abbreviations

Technical terms and abbreviations are used throughout this report. The meaning is usually clear from the context, but a glossary is provided for the non-specialist reader. In some cases, usage differs in the literature, but the term has a constant meaning throughout this review.

Glossary

Assessment team Independent teams, based in academic centres, that produce technology assessment reviews for NICE, commissioned by the NHS HTA programme.

Bias Deviation of results or inferences from the truth, processes leading to systematic deviation. Any trend in the collection, analysis, interpretation, publication or review of data that can lead to conclusions that are systematically different from the truth.

Cost-effectiveness The consequences of the alternatives are measured in natural units, such as years of life gained. The consequences are not given a monetary value.

Cost-effectiveness acceptability curve A Bayesian approach to the presentation of cost-effectiveness. The curve illustrates the probability of intervention A being more cost-effective than intervention B given a range of values that a decision-maker may attach to an additional quality-adjusted life-year.

Decision analysis A structured way of thinking about how an action taken in a current decision would lead to a result. Will usually involve the construction of a logical model, which is a mathematical representation of the relationships between inputs and results.

Expected value of perfect information The difference between the expected value of a model with perfect information and the expected value with current information allowing for whether perfect information would change the optimal decision.

Health technology assessment Health technology covers any method used by those working in health services to promote health, prevent and treat disease, and improve rehabilitation and long-term care. 'Technologies' in this context are not confined to new drugs or pieces of sophisticated equipment.

Health Technology Assessment (HTA) programme A national programme of research established and funded by the Department of Health's research and development programme.

Incremental cost-effectiveness ratio The ratio of the difference in costs between two alternatives to the difference in effectiveness between the same two alternatives.

National Coordinating Centre for Health Technology Assessment Coordinates the Health Technology Assessment Programme under contract from the Department of Health's research and development division.

National Institute for Clinical Excellence Part of the NHS. It is the independent organisation responsible for providing national guidance on treatments and care for those using the NHS in England and Wales. Its guidance is for healthcare professionals and patients and their carers, to help them to make decisions about treatment and healthcare.

Quality-adjusted life-years An index of survival that is weighted or adjusted by a value associated with patients' quality of life during the survival period.

Randomised controlled trial In healthcare evaluation, these are designed for particular

continued

Glossary *continued*

measurements; in particular relative treatment effects that are potentially subject to selection bias, such as a hazard ratio. Selection bias is minimised by randomly assigning people to one or two or more treatment groups and, where possible, blinding them and the investigators to the treatment that they are receiving. The outcome of interest is then compared between the treatment groups. Such studies are designed to minimise the possibility of an association due to confounding and remove sources of bias present in other study designs.

Technology Assessment Review The reports produced by the assessment teams as an input to a NICE technology appraisal.

List of abbreviations

BNF	British National Formulary
CDSR	Cochrane Database of Systematic Reviews
CEA	cost-effectiveness analysis
CENTRAL	bibliography of controlled trials
CI	confidence interval
CIN	cervical intraepithelial neoplasia
CIPFA	Chartered Institute of Public Finance and Accountancy
CMR	Cochrane Methodology Register
CRD	Centre for Reviews and Dissemination, University of York
CUA	cost-utility analysis
DARE	Database of Abstracts of Reviews of Effectiveness
EAG	Expert Advisory Group
EDSS	Expanded Disability Status Scale
HCT	historically controlled trial
HEED	Health Economic Evaluation Database
IFN	interferon
IPD	Incidence and Prevalence Database
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
ITT	intention to treat
LBC	liquid-based cytology
MIMS	Monthly Index to Medical Specialities
MSCRG	Multiple Sclerosis Collaborative Research Group
NHS EED	National Health Service Economic Evaluation Database
NICE	National Institute for Clinical Excellence
QALY	quality-adjusted life years
QEO	quasi-experimental and observational
RCT	randomised controlled trial
ScHARR	School of Health and Related Research, Sheffield
SCI	Science Citation Index
SSCI	Social Science Citation Index
TAR	Technology Assessment Review
UTI	urinary tract infections
VUR	vesicoureteral reflux

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices in which case the abbreviation is defined in the figure legend or at the end of the table.

Executive summary

Background

Decision-analytic models represent an explicit way to synthesise evidence currently available on the outcomes and costs of alternative (mutually exclusive) healthcare interventions. Usually their objective is to obtain a clear understanding of the relationship between incremental cost and effect in order to assess relative cost-effectiveness and to determine which interventions should be adopted given existing information. Given that the use of decision-analytic modelling for health technology assessment has increased exponentially in recent years, there is a need to consider how good practice in the field has been defined. Since the 1980s, several published guidelines have been available for those developing and evaluating decision-analytic models for health technology assessment. However, given the speed at which economic evaluation methodology has progressed, it is timely to review, critically appraise and consolidate those existing guidelines on the use of decision-analytic modelling in health technology assessment, and to identify key issues where guidance is lacking.

Objectives

- To identify and describe published guidelines for assessing the quality of decision-analytic models in health technology assessment.
- To develop a synthesised guideline and accompanying checklist using available good practice guidelines.
- To provide guidance on key theoretical, methodological and practical issues not yet covered in published guidelines. Two areas were identified in advance as priorities: literature searching for parameter estimation in decision models, and adjusting for bias in treatment effect estimates from observational studies used in decision models.
- To consider the implications of this research for what might be expected of future decision-analytic models relating to the National Institute for Clinical Excellence (NICE) technology appraisal process and health technology assessment in general.

Methods

The project consisted of four key elements:

- A systematic review of existing good practice guidelines was undertaken to identify and summarise guidelines currently available for assessing the quality of decision-analytic models that have been undertaken for health technology assessment. Areas of guidance that were relevant to economic evaluation in general, rather than decision-analytic modelling specifically, were omitted.
- A synthesised good practice guidance and accompanying checklist was developed. Each theme and subtheme from the review of guidelines was taken in turn, and its relevance was discussed in relation to the development of general guidelines for decision-analytic modelling in health technology assessment. Where previous guidelines were contradictory, a consensus decision was taken by the research team regarding the most appropriate item for the synthesised guideline. A checklist was constructed from the synthesised guideline, using the suggested headings and statements. This checklist was applied to three recent decision-analytic models undertaken as part of the NICE appraisal process.
- Two specific methods areas in decision modelling which have received relatively little consideration in the literature, were considered. They were selected on the basis of the team's experience, rather than any systematic review of the methods literature. The first method's topic is the identification of parameter estimates from published literature. Parameter searches were developed and piloted using a case-study model. The second topic relates to bias in parameter estimates; that is, how to adjust estimates of treatment effect from observational studies where there are risks of selection bias. A systematic literature review was conducted to identify those studies looking at quantification of bias in parameter estimates and the implication of this bias.
- The use of the guidance and checklist for future decision-analytic models developed as part of the NICE Technology Assessment Review (TAR) process was considered. Decision modelling is

central to the NICE technology assessment process and it is essential to assess the quality of those models that are developed to inform the Appraisal Committee.

Results

Synthesised guidance
Systematic searches identified 26 papers offering general guidance on good quality decision-analytic modelling. Of these, 15 met the inclusion criteria and were reviewed and consolidated into a single set of brief statements of good practice. Based on this review, a checklist was developed and applied to three independent decision-analytic models.

Elements were summarised under the headings of Structure, Data and Consistency. Within the published literature, the process of developing a framework for good practice has been iterative. Although the checklist provided excellent guidance on some key issues for model evaluation, it was too general to pick up on the specific nuances of each model.

Searching for parameter estimates
The searches that were developed helped to identify important data for inclusion in the model. However, the quality of life searches proved to be problematic: the published search filters did not focus on those measures specific to cost-effectiveness analysis and although the strategies developed as part of this project were more successful few data were found.

Effect of selection bias
Fourteen relevant references were identified, although three of these did not provide actual estimates of bias. Of the remaining 11 studies, five concluded that a non-randomised trial design is associated with bias and six studies found 'similar' estimates of treatment effects from observational studies or non-randomised clinical trials and randomised controlled trials (RCTs).

Implications for NICE appraisal process
Decision modelling is central to the NICE technology assessment process and it is essential to assess the quality of those models that are developed to inform the Appraisal Committee. One purpose of developing the synthesised guideline and checklist was to provide a framework for critical appraisal by the various parties involved in the health technology assessment process. First, the guideline and checklist can be used by groups that are reviewing other analysts' models and,

secondly, the guideline and checklist could be used by the various analysts as they develop their models (to use it as a check on how they are developing and reporting their analyses).

The Expert Advisory Group (EAG) that was convened to discuss the potential role of the guidance and checklist in the NICE TAR process felt that, in general, the guidance and checklist would be a useful tool in the NICE TAR process for the assessment team, technical leads and committee members. However, some caution must be applied when using the checklist, and it is particularly important to realise that the checklist is not meant to be used exclusively to determine a model's quality, and so should not be used as a substitute for critical appraisal.

Currently, no common checklist is used in the review process. It is hoped that further discussion between the assessment teams and NICE will lead to the use of the same checklists across the groups. This would include those used for economic evaluation in general, as well as decision models in particular.

Conclusions

The review of current guidelines showed that although authors may provide a consistent message regarding some aspects of modelling, in other areas conflicting attributes are presented in different guidelines.

A preliminary assessment showed that, in general, the checklist appears to perform well, in terms of identifying those aspects of the model that should be of particular concern to the reader. The checklist cannot, however, provide answers to the appropriateness of the model structure and structural assumptions, as these may be seen as a general problem with generic checklists and do not reflect any shortcoming with the synthesised guidance and checklist developed here. The assessment of the checklist, as well as feedback from the EAG, indicated the importance of its use in conjunction with a more general checklist or guidelines on economic evaluation.

The review of current guidance for good quality decision-analytic modelling for health technology assessment highlighted a number of methodological areas that have not received attention in the literature on good practice. There are a lot of these areas and, therefore, it was only possible to consider two specific methods areas in

decision modelling: the identification of parameter estimates from published literature, and the issue of adjusting treatment effect estimates taken from observational studies for potential bias. Literature reviews showed that both of these areas are under-researched and are areas in which further research is needed.

Recommendations for research

This project has highlighted many areas where further methods research may be of value. In particular:

- A review of the literature is needed pertaining to the quantification of selection bias in non-controlled studies and in controlled observational studies.
- Empirical research is needed to define further the level of bias in the different non-RCT study designs.
- Studies are needed which compare results from RCTs with those from other non-randomised studies.
- There is a need to assess the strengths and weaknesses of alternative ways to adjust for bias in a decision model.
- Studies are needed to determine how to prioritise searching for parameter estimates. The value of information methods is worth consideration.

Chapter 1

Introduction

Health technology assessment is undertaken to inform decision-making regarding the use of particular healthcare programmes and interventions. The NHS Health Technology Assessment programme is a national programme of evaluative research funded as part of the UK Department of Health's research and development programme (www.ncchta.org). Its purpose is to provide high-quality evidence on the costs, effectiveness and broader impact of health technologies for those who use, manage and provide care in the NHS. The HTA programme commissions both primary and secondary research. The programme is coordinated by the National Coordinating Centre for Health Technology Assessment (NCCHTA), which is based at the University of Southampton.

The process of estimating the effectiveness and cost-effectiveness of a technology inevitably involves the synthesis of a range of evidence. This suggests two key components to the health technology assessment process, which should be carried out in a mutually supportive and interactive manner. One involves gathering evidence from a range of primary studies relating to the disease and/or technology of concern. These data relate to factors including the epidemiology, natural history, costs and quality of life impact of a disease, and to the implications for these parameters for particular interventions. The second component involves the formal synthesis of all the available data within an analytical framework which is focused on the ultimate objective of the decision, for example, to select a particular intervention which is consistent with maximising health benefits from available resources. The process of decision-analytic modelling is now seen as central to the process of health technology assessment in general, and it plays a key role in the National Institute for Clinical Excellent (NICE) appraisal process in particular.[1]

Decision-analytic models represent an explicit way to synthesise the evidence currently available on the outcomes and costs of alternative (mutually exclusive) healthcare interventions. Usually their objective is to obtain a clear understanding of the relationship between incremental cost and effect in order to assess relative cost-effectiveness and to determine which interventions should be adopted, given existing information, including costs and relevant preferences. They provide a key role in translating the uncertainty associated with parameters (e.g. the cost of a particular adverse event, the quality of life impact of a condition or the relative risk reduction associated with a specific intervention) into the uncertainty associated with making a decision regarding the use of particular technologies.[2] They are also a valuable tool for quantifying the implications (in terms of resource costs and health gain forgone) of that decision uncertainty which, in the form of expected value of perfect information, can be used to prioritise future research.[1,3–6]

The use of decision-analytic modelling for health technology assessment has increased exponentially in recent years, and there is a need to consider how good practice in the field has been defined. Given the need for every model to focus on a specific decision problem, and to deal with the realities of data availability and the complexities of particular diseases and interventions, it is unlikely ever to be possible to establish guidelines that spell out every step of the modelling process, and against which particular models can be fully assessed.[7] However, there is a need to set out the general characteristics that all decision-analytic models should aim to achieve to help decision-makers to distinguish a good decision-analytic model from a bad one. This may at least help to identify those that fail some very basic tests of quality, and where more detailed quality assessment may not be required. It is also important to identify areas where the literature offers no description of good practice and where there is a potential need for methodological research to fill the gaps.

Study objectives

Given the speed at which economic evaluation methodology has progressed, it is timely to review, critically appraise, consolidate and further develop existing guidelines on the use of decision-analytic modelling in health technology assessment. In light of this, the study contains four broad elements:

- a systematic search of the literature to identify and describe published guidelines for assessing the quality of decision-analytic models in health technology assessment
- development of a synthesised guideline and accompanying checklist using available good practice guidelines; application of this checklist to three case-study decision-analytic models that have been developed as part of the NICE technology appraisal process
- the provision of guidance on key theoretical, methodological and practical issues not yet covered in published guidelines
- the consideration of the implications of this research for what might be expected of future decision-analytic models relating to the NICE technology appraisal process.

Chapter 2

A framework for good practice in decision-analytic modelling studies: review of current guidelines

A systematic search was undertaken to identify and summarise currently available guidelines for assessing the quality of decision-analytic models that have been undertaken for health technology assessment. It is recognised that a wealth of literature exists on good decision modelling outside health technology assessment. However, given the timescale for this project, it was felt that extending the scope beyond health technology assessment would produce an unmanageable number of papers to review. This chapter details the methods and results of this review.

Methods

Search strategy
The broadness of the search topic and the ambiguous nature of the terms included led to the development of a pragmatic focused search to avoid being overwhelmed with records to assess. The following databases were searched for guidelines:

- EconLit
- EMBASE
- Health Economic Evaluations Database (HEED)
- MEDLINE and PREMEDLINE
- NHS Economic Evaluation Database (NHS EED)
- Science Citation Index (SCI)
- Social Science Citation Index (SSCI).

Search strategies, adapted for each database, were developed to identify relevant papers from NHS EED, MEDLINE, EMBASE, EconLit, SSCI, SCI and HEED. A search for discussion papers published by academic centres and health technology assessment agencies, and selective searches of the Internet were also carried out. The full search strategies can be seen in Appendix 1.

On the basis of an assessment of abstracts, a shortlist of potential guidelines was established and copies of relevant papers were obtained.

Inclusion criteria
Papers were deemed relevant for inclusion in the

review if, on the basis of title and abstract (if available):

- the paper provided general guidance on the elements of a good decision-analytic model for health technology assessment, or
- the paper provided explicit criteria against which to assess the quality or validity of a decision-analytic model built for the purpose of health technology assessment.

These inclusion criteria included decision-analytic models for effectiveness analysis as well as cost-effectiveness analysis, but excluded papers that focused on just one aspect of decision-analytic modelling methodology or those relating to specific disease areas. 'Decision-analytic modelling' is defined here as a model using summary data to establish a preferred healthcare intervention on the grounds of effectiveness or cost-effectiveness, or both.

Data extraction
Those papers describing an overall approach to quality assessment in decision-analytic modelling for health technology assessment were summarised. Data were extracted from the papers using a deductive iterative thematic approach, which was based on the general themes identified in earlier work:[7] structure, data and consistency. Data extraction was undertaken by one researcher and checked by a second. Any discrepancies were discussed and resolved between the two researchers.

Results

On the basis of the title and abstract, 222 full papers were obtained, 26 of which were potentially relevant for the review of guidelines. Fifteen papers[7–21] gave general guidance for what might constitute a good decision-analytic model as well as offering specific guidelines covering model quality. *Table 1* describes the types of paper identified, including those that were not included in the review.

TABLE 1 *Summary of papers ordered for review*

Reason for order	No. of references
Guidelines for good decision-analytic modelling	26
General guidance on decision analysis	124
Validation of a specific model	36
Sensitivity analysis in decision-analytic models	18
Data sources for decision-analytic models	6
Papers that address the issue of bias in decision-analytic modelling	8
Presentation of decision-analytic models' results	2
Incorporating attitude to risk into decision-analytic models	2
Total	222

Data from the 15 guidelines were extracted qualitatively using a thematic approach. Each guideline was summarised and then formulated into themes and subthemes under the broad headings of Structure, Data and Consistency (see Appendix 2). A detailed review of those guidelines is given below. This part of the document describes what each guideline says and the consistency between them, but does not seek to assess the strengths and weaknesses of their arguments or to synthesise across guidelines; this is the purpose of Chapter 3.

Summary of existing guidelines of good practice

The guidelines differed in terms of their comprehensiveness and, in some, the area(s) of model quality they considered. *Table 2* gives a general summary of the specific areas of decision-analytic modelling considered by each of the guidelines.

Background
Both users and developers of decision-analytic models agree that a good decision-analytic model should be transparent[7–11,13,20,22] and the structure, data and process of building the model should be detailed enough to enable a competent analyst to reproduce it. It should be tailored to the purpose for which it is to be used, be useful for informing decisions at which it is aimed and be readily communicated.[8,20]

To increase the transparency of the modelling process, and to increase the confidence of the users of model outputs, there have been several

attempts to identify what constitutes good practice in decision-analytic modelling. Within these guidelines, however, it is important to recognise that it is not possible (or indeed desirable) to develop a rigid set of steps that a decision modeller should follow to develop what might be termed 'a good model'.[10] This might lead to the distortion of research priorities (questions will focus on simple questions where good evidence is already available) and the possible rejection of valuable models, and will almost inevitably discourage novel methodology.[7] Rather, development of guidelines provides a framework for what should constitute good practice in decision-analytic modelling and, for the relevant dimensions of quality assessment, can provide details of the appropriate methodology. There is a need, however, for a guideline that is sufficiently generic to apply across models of different types and diseases, without being so general that it is of little value to users and developers of models.

The guidelines that are available range from guidelines for general reporting to papers that have sought to define and suggest methods to assess validity in decision models. Within the published literature, the process of developing a framework for quality assessment has been iterative, with more recent guidelines extending, and further developing, earlier work as methodological advances are made.

The purpose of this review is to synthesise the current guidelines and to identify areas where they do not agree. In reviewing the guidelines available, they have been summarised according to three general components of a model: structure, data and consistency.

Structure
Determining the structure of the model is key to the model's development. Several dimensions relating to the structure of the model have been addressed in published guidelines, each of which is summarised below.

Statement of decision problem/objective
There should be a clear statement about the decision problem prompting the analysis from the outset of the modelling process,[7,10,11,15,18,19] and the resulting structure of the model should be consistent with the stated decision problem.[7] The statement should include information about the disease(s) or condition(s) being modelled, descriptions of the patient or study population, the possible diagnostic and therapeutic actions and interventions, and the possible outcomes after

TABLE 2 General summary of included guidelines

Element	Akehurst et al., 2000[8]	Buxton et al., 1997[9]	Eddy 1985[18]	Habbema et al., 1990[14]	Halpern et al., 1998[11]	Hay 1998[12]	ISPOR Task Force, 2001[13]	McCabe and Dixon, 2000[10]	Nuijten et al., 1998[15]	Ramsey and Sullivan, 1999[17]	Sculpher et al., 2000[7]	Sendi et al., 1999[16]	Sonnenberg et al., 1994[21]	Soto 2002[19]	Weinstein et al., 2001[20]
Structure															
Statement of decision problem/objective			✓	✓	✓			✓	✓	✓	✓			✓	
Justification of modelling approach					✓			✓	✓	✓	✓			✓	
Statement of scope/perspective	✓		✓		✓	✓		✓	✓	✓	✓		✓	✓	
Structural assumptions	✓		✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Consistency of structure	✓		✓		✓	✓	✓	✓			✓		✓	✓	
Strategies/comparators		✓		✓	✓	✓	✓		✓	✓	✓		✓	✓	✓
Model type	✓	✓			✓	✓	✓		✓		✓		✓	✓	
Time horizon	✓	✓			✓	✓	✓		✓		✓		✓	✓	
Health states					✓		✓		✓		✓			✓	
Cycle length							✓		✓		✓			✓	
Parsimony	✓	✓	✓		✓		✓				✓				
Data															
Data identification															
General	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Prospective data					✓	✓	✓		✓	✓	✓			✓	
Retrospective data					✓	✓									
Expert opinion					✓										
Resource use/costs					✓		✓				✓			✓	
Utilities							✓				✓			✓	
Iterative data identification															
Data modelling															
General					✓	✓	✓			✓	✓				
Data synthesis					✓	✓	✓				✓				
Discounting					✓		✓								
Analysis of trial data							✓		✓		✓				
Treatment effect							✓		✓						
Transition probabilities							✓								
Mortality															
Extrapolation														✓	

TABLE 2 General summary of included guidelines (cont'd)

Element	Akehurst et al., 2000[8]	Buxton et al., 1997[9]	Eddy 1985[18]	Habbema et al., 1990[14]	Halpern et al., 1998[11]	Hay 1998[12]	ISPOR Task Force, 2001[13]	McCabe and Dixon, 2000[10]	Nuijten et al., 1998[15]	Ramsey and Sullivan, 1999[1]	Sculpher et al., 2000[7]	Sendi et al., 1999[16]	Sonnenberg et al., 1994[21]	Soto 2002[19]	Weinstein et al., 2001[20]
Risk factors															
Utility					✓		✓			✓			✓	✓	
Charges and costs							✓			✓					
Adjustment over time/between countries							✓								
Half-cycle correction							✓				✓				
Data incorporation	✓	✓	✓		✓		✓		✓	✓	✓		✓	✓	
Assessment of uncertainty															
General statement regarding sensitivity analysis		✓	✓								✓				
Parameter	✓			✓	✓	✓	✓		✓	✓	✓		✓	✓	
Structural	✓				✓	✓	✓			✓					
Methodological	✓														
Heterogeneity versus variability					✓		✓		✓		✓		✓	✓	
Value of additional research			✓	✓	✓						✓				✓
Consistency															
Internal consistency	✓		✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
External consistency	✓	✓	✓		✓		✓	✓	✓		✓	✓	✓	✓	✓
Between-model consistency	✓	✓			✓		✓	✓			✓	✓	✓	✓	✓
Predictive validity					✓	✓	✓	✓			✓	✓	✓	✓	✓

treatment,[11,14] in addition to a general statement of the overall objective of the model.[19] A clear definition of the model's objective will enable readers of the model's outputs to assess the relevance to their own particular situation.[17]

The objective or purpose of the model needs to be defined from the outset, as it will have a fundamental impact on several aspects of the model, such as its perspective, the comparators included, the model's complexity, data sources and choice of outputs.[10] Soto[19] argues that the study question addressed by the model, along with reasons for its development, including the goal and objectives of the study, should be clear, transparent, relevant and achievable according to available data,[19] implying that the decision problem should be governed by available data.

The type of economic evaluation that is being undertaken (e.g. cost-effectiveness analysis, cost–benefit analysis) should be described and justified.[15]

Justification of the modelling approach
The rationale for the study should be described, and should include details of the information upon which the study question is based.[17] Models can only be developed when alternative interventions, outcomes and probabilities can be specified, or at least estimated.[11] The modelling approach and its methodology (model type, structure) should be justified;[10,15] this will require indication of the lack of any alternative information or an appraisal of the information that demonstrates its weakness.[10] It has been argued that it must be clear that, given the question to be answered, a model is the most (or only) appropriate or practical approach.[11,19]

Statement of scope/perspective
The perspective of the model should be stated and determined at the outset of the study,[11] and the model should be structured so that the inputs, outputs, costs and consequences are relevant to the decision-making perspective.[13] In accordance with most economic evaluation guidelines, many authors believe that the societal perspective should be used in decision-analytic modelling studies.[23,24] However, there is continued controversy over the relevance of the societal perspective for some decision-makers.[22] A narrower perspective, such as that of the relevant decision-maker, can also be used as long as it is presented alongside the broader societal perspective.[19,22] Nuijten and colleagues[15] recommend that the perspective of the modelling study should accord with country-

specific guidelines, in addition to rules for reimbursement and treatment patterns.[15] Halpern and colleagues[11] argue that the perspective should agree with the purpose of the model[11] and International Society for Pharmacoeconomics and Outcomes Research (ISPOR)[13] suggests that if a perspective narrower than societal is used, there should be discussion, at least qualitatively, of the implications of broadening the perspective.[13]

The scope of a decision model should be clearly specified.[8] It will depend on the specific question addressed by the decision model, the disease or condition and the interventions under study.[11] As with the perspective of the model, this should be determined from the outset of the study.[11] In general, the model's scope refers to its limits or boundaries and should include a statement of the perspective of analysis, the technologies involved, the population and the setting studied, and the time horizon to which it relates.[8,11,15,18] The population under study within the model will depend on the objective of the model. If the model is dependent on the extrapolation of particular clinical trial results to final outcomes, then details of the trial population should be given (i.e. inclusion and exclusion criteria). If the model is developed for such a specific population, factors that may limit the applicability of the results to a wider population should be discussed.[15] It is important that the scope of the model is described in sufficient detail for a reader of the model's results to be able to judge applicability within their own particular setting.[17] The main factors that could limit the applicability of the results should be detailed.[18] Data availability may limit both the scope and perspective of the analysis,[11,15,18,21] or at least refine it.[7]

The outcomes used in the model should reflect the perspective and scope of the model. It is generally recommended that models should include long-term or final outcomes, and the choice of outcome should be justified.[19] All relevant outcomes to the condition and disease, including adverse events, relapse and death, should be incorporated,[11,17,21] and clinical outcomes should show effectiveness rather than efficacy.[19] Nuijten and co-workers have argued that clinical events that do not differ statistically between intervention and control can be neglected by the model.[15]

Rationale for structure
The structure of the model should be consistent with a coherent theory of the health condition under evaluation[7,13,19] and its presentation should be clear.[17] Halpern and colleagues[11] suggest that

the steps, branches or health states in the model should follow the course of treatment or the progression of the disease, and that progressions should make clinical and mathematical sense.[11] Input from clinicians and decision-makers may be useful at this stage to ensure that the decision problem is relevant.[11,19] All sources of data and expertise used to develop the model structure should be described.[19]

Sonnenberg and colleagues[21] describe a conceptual framework to explain the relationship between the decision problem and the resulting model. The framework consists of four levels:

- Biological truth: this is not usually directly or completely knowable.
- Theoretical model: this describes our understanding of the biological truth, analogous to the scientific theory that best explains the facts known at a given point in time. Normally this amounts to an enumeration of the important choices and events and the relationships between them.
- Practical model: the most detailed model that can be constructed on the basis of the data available. This practical model may also be constrained by a variety of simplifying assumptions, reflecting not only limitations of the data but also limitations to the size and complexity of the model.
- Implementation model: this represents the programming of the practical model in the relevant computer software.

Sonnenburg and colleagues[21] explain how the move from the theoretical model to the practical model will involve assessment of available data and the requirement of simplifying assumptions to limit the size and complexity of the model. They argue that the detail of the implementation model should be no more than supported by available data.[21] Later guidelines explain how the structure of the model should reflect the essential features of the disease and its interventions, irrespective of data availability, and that ignoring data that are incomplete or difficult to interpret cannot be justified.[7,13] There is an acceptance in more recent guidelines, however, that data availability will inevitably play some role in structuring and refining the scope of the model. For example, the model may relate to an 'average' patient group as opposed to several specific subgroups, owing to data availability.[7]

The causal relationships included in the model should be explained and substantiated by best available evidence.[8] However, this does not mean that all of the causal links within the model need to be proven scientifically,[13] nor does it suggest that the clinical area must be represented with absolute detail.[10] Rather, there should not be any evidence that contraindicates the modelled relationships,[13] and simpler structures (as opposed to attempting to represent the clinical area with absolute accuracy), which do not materially affect the model's results, are preferred.[10]

Sonnenberg and colleagues[21] explain how the model should include key variables to represent important factors in the patient population. However, it is not clear what is meant by 'key variables'. On a similar note, Ramsey and Sullivian[17] explain how events following chance nodes within decision trees should be "all that are possible for the patient".[17] It is important that the possible pathways of the model are feasible and sensible for the particular patient group under investigation[10,11] and that the model structure is able to describe current practice.[10] Eddy[18] describes this as first order validation (i.e. that the structure of the model should make sense to those who have a good knowledge of the problem[18]) and others describe it as (clinical) face validity.[11,22]

Structural assumptions
Although limited availability of information should not be considered an insurmountable barrier to model development, the specifics of the model will be influenced by available information. As a result, this must be explored early in the modelling process.[11] It is important that any underlying theory used to develop the structure of the model and any assumptions or choices that are made are transparent and justified.[8,15,18–20,22] For example, this may include the assumed relationships between parameters, and all simplifying assumptions.[21] The choice of assumptions should be realistic and logical, and should reflect available data and routine medical practice in the chosen setting to which the model relates.[19,22] However, the absence of data is not in itself a justification for simplification.[10]

If evidence regarding the structural assumptions is incomplete, then the limitations of the evidence supporting the chosen model structure should be acknowledged and, if possible, sensitivity analysis using alternative model structures should be performed.[13] In theory, any simplifying assumptions that are made should not have material impact on the model's results.[10] However, at the very least, a discussion of how the modelling assumptions and inherent biases in the model may

affect the results, estimating both the direction of the bias and the magnitude of effect, should also be included.[11,18,19]

The validity of the model will rest on whether its assumptions are reasonable in the light of the needs and the purposes of the decision-maker and whether, after close examination, its implications make sense.[19,22] One approach to testing the appropriateness of structural assumptions is to undertake comparison with other models and to explain and justify any differences.[10] Another is to use experts to comment on the assumptions used in the model.[11,19] Finally, it should be noted that some of the guidelines imply that Markov-type models will be the rule in cost-effectiveness analysis. Although these models are widely used, part of the process of structuring a model involves assessing whether other types of model such as discrete even simulation models would be more appropriate.[25]

Strategies/comparators
A broad range of feasible strategies or comparators should be evaluated within the model[7,11,13,14,19,21] and each option should be mutually exclusive.[17] The choice of options may be based on reviews of the literature and/or expert opinion, and will include evidence from local treatment patterns, related clinical studies and clinical guidelines.[11,14,15,19] However, options should not be limited to the availability of direct evidence from clinical trials or what represents current practice.[7,13] Hay and Jackson[22] recommend that the comparator should be a standard of care,[22] although a single standard may not exist.

Interventions that are executed in the model, conditional on the results of a test, may also be included as first line interventions; for example, a see and treat policy.[14] It is also suggested that the model should include extreme strategies[7,11,14,21] even if they are clinically unrealistic. This is because they serve as useful anchor points against which all other strategies can be measured and provide a means of assessing the internal consistency of the model. No therapy, watchful waiting, prevention and screening options may also be relevant.[15,21] Details should be given regarding the options included in the model, including the characteristics of the studies from which they are based.[19]

There should be a balance between including the full range of feasible options (given the scope and perspective of the model) and the need to keep the model manageable (within the time and resource constraints)[7,13] interpretable and research based.[13] However, it is important that the decision problem is not framed inappropriately by excluding a relevant comparator.[9] Any assumptions and compromises that are necessary to satisfy the time and resource constraints should be transparent and justified.[7]

Model type
The choice of model type should be appropriate for the decision problem and the intended use of the model.[11] Most commonly found in the health technology assessment modelling literature are tree structures or Markov models. In decision trees, the events of interest may happen only once and during some discrete period, with the timing of events only reflected by the outcome measure(s). Markov processes are particularly useful for modelling chronic diseases, or where exposure to risks is continuous and its likelihood changes over time.[21,22] The chosen model type should involve the simplest structure that adequately reflects the time dependence of events being modelled and the complexity of the interaction of various consequences of each decision option within the specified scope of analysis.[7,11,19,21] The choice of model type should be justified: this can be achieved with a detailed description of the disease process and the prognoses of patients over time.[15]

Time horizon
The time horizon of the model should be specified within a statement of the scope of the model[8,11] and the analytical horizon should be justified.[15] The analyst should distinguish between the length of treatment effect and the time horizon of the model.[9] It should be sufficient to encapsulate all major clinical and economic outcomes[19] and indicate when cost and effect differences between options are stable.[7] However, it is recognised that this is an output of the model rather than an a priori specification.[7] The time horizon will depend on the clinical area in consideration[19] and should match that of the actual process being considered in the analysis.[21]

For pharmacoeconomic studies, the time horizon should encapsulate the duration of time for which a drug can be expected meaningfully to impact on the patients' health (morbidity and mortality).[22] In general, the model should extend far enough into the future to be able to reflect important and valued differences between the options included.[13] Lifetime horizons are appropriate for most models,[7,13] in particular Markov models.[15] For

decision tree models, the time horizon is usually from onset of treatment until recovery or death.[15] Shorter horizons can be justified for some disease processes or interventions[7] on the basis of no differences in survival or long-term chronic sequelae between options.[13] The lack of long-term follow-up data should not be used as a rationale for failing to extend the time horizon of the model to the period relevant for the decision.[13]

Health/disease states

The guidelines vary in their use of the terms 'health' and 'disease' states to describe the states that may be used in a state transition model, such as a Markov-type model.

The health states used in the model should be described and justified.[7,15,19] Both Nuijten and colleagues[15] and ISPOR[13] describe how health states can be defined according to the underlying disease or observable health states/local treatment patterns,[13,15] but ISPOR[13] describe how structural bias is avoided by modelling underlying disease status and calibrating outputs to data on observable clinical status. Sculpher and colleagues[7] state that disease states should be chosen to reflect the underlying biological process of the disease in question and the impact of interventions, rather than health service inputs, because the model should reflect accepted theory/clinical classifications of disease.[7]

There are no strict rules for the appropriate number of health states, but the chosen structure should be sensitive to changes in the underlying disease while remaining manageable. Reasons to include additional subdivisions of health states may be based on their clinical importance or their relation to mortality, quality of life or cost. Health states that may not be considered clinically important may still be important to include for these other reasons. Likewise, the inclusion of health states that do not affect the models' results may be included for face validity reasons.[13] It is recognised that the number and choice of health states will be a trade-off between descriptive realism and simplicity, but any choices that are made should be justified.[7] However, states should not be omitted because of lack of data.[7,13]

Cycle length

There is little guidance regarding the methodology for determining the appropriate choice of cycle length in a discrete time state transition model. It is reasonable to argue that multiple changes should not occur within a single cycle,[13] but ISPOR[13] does not indicate the

'changes' to which it refers. Nuijten and colleagues[15] describe how cycle length can be determined by the nature of disease or local treatment patterns, whereas Sculpher and colleagues[7] argue that cycle length should be based on the nature of disease. The cycle length should be the minimum interval over which the pathology or symptoms of the patient are expected to alter.[7] They argue that basing a cycle length on clinical review times is inappropriate because finding the optimal length of follow-up or review should only be one aspect of the options evaluated by the model. Data availability should not govern the chosen cycle length,[7] and it is important that the chosen cycle length is justified.[7,13,19]

Parsimony

Most experts agree that a model should be as simple as possible to address adequately the decision problem.[7,8,11,13] However, Buxton and colleagues[9] argue that the principal reason for parsimony is that it facilitates understanding by decision-makers, and that the level of simplicity will depend on the sensitivity of the policy implications to added complexity.[9]

The model should only introduce variables and structural components that are relevant to the scope of the model,[8,11] while ensuring that the essential features of the disease and interventions over time are included.[7,13] Theoretically possible but non-observable branches should not be developed.[11]

Simplifying assumptions are almost inevitable in models owing to constraints on time and resources. However, it is important that the simplifications that are made are described in detail and justified on the basis that they will not materially affect the results of the model.[13] ISPOR[13] suggests a structural sensitivity analysis that uses a less aggregated model to test this. Necessary model complexity may change over time, as more data become available.[11]

Data

Data issues are categorised into three broad sections: data identification, premodel data analysis and data incorporation. In addition, guidelines relating to the appropriate assessment of uncertainty have been separated out. The majority of previous guidelines make general statements about data issues,[8–10,15,17,18,20–22] and a few make recommendations that are specific to particular elements of data identification or premodel data analysis.[7,11,13,19] In summarising

the evidence, an attempt has been made to separate the general advice from the specific.

Data identification

Both Akehurst and colleagues[8] and Weinstein and colleagues[20] describe how the data used in the model should be consistent with available and appropriate information.[8,20] However, they offer no guidance regarding how to identify what is available or appropriate for the model. Halpern and colleagues[11] extend this argument and describe a top–down approach to data identification, starting from those of the highest quality such as epidemiological studies, randomised controlled trials (RCTs) or prospective naturalistic studies.[11] Others also explain how data should come (where possible) from well-designed[18,19] or high-quality[17] studies, but are not explicit in what they mean by these terms.

The quality and relevance of all data, including trial data, identified for the model, and the potential for, and likely direction of, bias must be assessed.[10,11,15,19] Soto[19] recommends that recognised rules for assessing the quality and reliability of meta-analyses and experimental and observational studies should be used in assessing data inputs[19] and makes reference to several guidance documents. Halpern and colleagues[11] offer detailed guidance for the assessment of prospective data. They offer evaluation criteria, and recommend that assessments be made with regard to availability of those data (i.e. whether they were published in a peer review journal), sample size, period (duration) and frequency of data collection, the degree of patient follow-up, patient population characteristics and methods of data collection.

Halpern and colleagues[11] state that, when trial-based data are used, all relevant trials should be used, and that the use of subsets requires justification. They also describe how clinical trial data should not always be regarded as the gold standard for modelling. With regard to retrospective data, they[11] provide specific guidance and recommend that the completeness of a retrospective data set should be assured before use. This should include assessment of the patient population comprised in the data set and the potential of bias, and assurances that specific data points are not systematically missing.

Where no data are identified for particular parameters, expert opinion is a legitimate source of information for the model. However, it is important that the identification and methods used to elicit expert opinion are documented

clearly.[7,15,17,19] This should include details of the inclusion criteria for experts, sample size, the types of questions asked, the method for data collection and number of iterations. The selection of experts should be based on individuals with substantial credibility in their fields who are likely to be opinion leaders, and they should come from a variety of practice settings and geographical locations.[11,19] Where individuals' opinions are sought, structured workbooks, which contain existing information and closed-ended questions, are recommended.[11] Group interviews should contain between five and eight experts.[11] Delphi, modified Delphi or nominal group techniques have been recommended.[11,13,15] Soto[19] suggests that a definition of consensus should be determined before expert opinion is sought, and explains how parameter estimates should be based on majority as opposed to individual opinion. However, it is important that a consensus opinion is not forced because there may well be genuine variability in opinion, based on experts' own experiences, which should be incorporated into the model.[7]

Soto[19] recommends that the values derived from expert opinion should be validated. However, the validity of experts' opinions cannot be ascertained without data against which to compare it.[11] Tests for face validity to ensure clarity and completeness are useful.[11]

Methods used to determine the units of healthcare utilisation should be described[19] and decisions regarding appropriate cost estimates should be based on a sound rationale.[22] Resource utilisation data derived from trials may not be generalisable and it is advised that real clinical practice data are used where possible.[19] Retrospective data are a common source of resource utilisation estimates and it is important to ensure that the patient group matches patients in general in order to improve generalisability.[19] Standard sources such as, in the UK, the Monthly Index of Medical Specialities, British National Formulary (BNF), the Chartered Institute of Public Finance and Accountancy (CIPFA) and NHS Reference Costs represent a useful starting point.[11]

Any trade-offs that are made in terms of, for example, appropriateness of study population versus study design should be documented[10] and any choices that are made should be justified.[13] Hay and Jackson[22] recommend that, where choices are made, conservative values of all parameters should be chosen for the base case. However, it may not be possible to evaluate empirically the implications of choosing one combination of data

over another.[10] The process of gathering all information should be documented explicitly, including any primary data collection methods and search criteria used.[11,15,17,19]

It is important to recognise that the identification of best available data for every parameter within the model may not represent a good use of resources.[7] ISPOR[13] recommends that a full systemic review should be carried out for key parameters of the model and that all known data should be incorporated for these key parameters, including those that fall short of the conventional thresholds of statistical significance. The group does not, however, define what they mean by 'key' parameters.

The notion of the relative importance of parameters is described in McCabe and Dixon's[10] work. They describe how the relative importance of one data source over another is not constant and that hierarchy of data should be secondary to the identification of a hierarchy of parameters within the model. Whether acquiring all existing data for all parameters is a good use of resources or not is an empirical question, and the model itself can be used to determine the optimal data to incorporate.[7] A case should be made for all reasonable opportunities to obtain new data prior to modelling to have been considered.[13] In this context, reasonable means that the cost and delay in obtaining further information are justified by the added value of the additional data.[13] Models should be repeatedly updated and/or abandoned and replaced as new evidence becomes available to inform their structure or input values. Models that have been shown to be consistent with subsequent evidence, but have not been revised to incorporate such evidence, should be abandoned until such a process is undertaken.[13] If there are several iterations of the model, each of which incorporates new evidence deemed cost-effective to collect and incorporated by an earlier version of the model, the optimality of data inputs is justified.[7] However, such formal methods are not expected to be used[13] and in current practice there is rarely a series of models.[7] It is sufficient to give a heuristic argument as to why the data identified for the model are optimal.[13] It should be clear that all sources of information that are available at relatively low cost in terms of researcher time have been searched for and that the most appropriate parameter values have been used.[7]

The accuracy of the results of a model will be limited by the accuracy of the data incorporated.[18] However, this is not a limitation of modelling studies in isolation, and no model should be criticised because the data fall short of scientific rigour.[7,13] It is, therefore, important that the results of models are reported conditional on their input data.[13]

Premodel data analysis
This relates to the mathematical and statistical processes that are undertaken to transform the data identified for the model into a form that can be incorporated into the model.[13] All premodel data analysis methodology should be disclosed and justified by providing evidence of its general acceptance or empirical validity.[13] In general, the process of data analysis should follow accepted methods of epidemiology and statistics such as meta-analysis.[7,13] It is important that sensitivity analysis is performed to assess the implications when alternative but equally defensible modelling approaches may lead to materially different results. However, the base-case analysis should relate to assumptions where there is most support.[13]

Several specific issues relating to premodel data analysis have been discussed in previous guidelines, each of which is detailed below.

Data synthesis
Data should be pooled using recognised techniques such as random effects meta-analysis.[7,11,13,17] The process should be carried out in a rigorous and systematic fashion and incorporate all relevant literature.[11]

Discounting
When the model relates to a longer time-frame, both costs and benefits should be discounted to present values.[11,22] Halpern and colleagues[11] and Hay and Jackson[22] additionally recommend that both costs and benefits should be discounted at the same rate, with alternative assumptions explored through sensitivity analysis.

Analysis of trial data
Where direct trial data are available for incorporation into a model, the data should be analysed at the patient level if at all possible.[22] All trial recruits should normally be included in the analysis [i.e. intention to treat (ITT) analysis as opposed to therapy results] and the generalisability of these types of data should be discussed.[11,22]

Deriving estimates of absolute treatment effect/adjusting absolute treatment effect for variation in baseline events
It is often appropriate to derive relative risks (or odds ratios) between treatment options in trials

and then to superimpose these onto baseline probabilities derived from other sources (which are usually population based).[13]

Transition probabilities
It is important that transition probabilities are derived appropriately when they are based on probabilities estimated over an interval other than cycle length. This involves manipulation first into a rate, and then into a probability corresponding to the cycle time used in the model.[7,13] The methods used should be documented.[15]

When transition rates or probabilities depend on events or states that may have been experienced in periods, this dependence should be modelled. This may be achieved either by incorporating a clinical/treatment history into the specification of the health states or by including history as a covariate in specifying the transition probabilities.[13]

Mortality
All-cause mortality should be derived from life tables unless an alternative source can be justified. It is not generally necessary to correct all-cause mortality for disease-specific mortality unless the disease represents a significant part of all-cause mortality.[13]

Extrapolation
The methods and assumptions that are used to extrapolate short-term results to final outcomes (e.g. trial-based intermediate outcomes to survival) should be documented and evidence should be provided that the methodology is valid.[15,19] This should include justification of the choice of survival function (e.g. exponential or Weibull forms).[13] All-cause mortality should generally be derived from life tables (see above).[13]

Risk factors
Evidence supporting the additive or multiplicative effects of risk factors on baseline probabilities or rates of disease incidence or mortality should be sought.[13]

Utility
The source of utility weights, methods used to obtain them and methodology used to transform health state estimates into quality of life scores should be documented.[11,13,17,19,21]

ISPOR recommends that combining domain-specific utility weights into multiattribute utility functions should be undertaken using validated health-related quality of life instruments, which have prespecified scoring systems based on standard gamble or time trade-off techniques.[13] Guidance from Halpern and colleagues[11] and Ramsey and Sullivan[17] relates to the primary collection of utility data. Halpern and colleagues[11] approach their guidance from a theoretical perspective, and suggest that, in general, standard gamble questions should be asked of non-healthcare personnel (patients or informed public), but that the most appropriate approach will depend on the condition and intervention under study.[11] Ramsey and Sullivan[17] provide guidance for clinical readers of modelling studies and describe how the credibility of utility estimates might be judged on:[17]

- the size of the sample from which the utility values are derived
- whether the utility weights match the readers' own clinical experience with similar patients
- whether the estimates are conservative (chosen to bias the outcome away from the new option).

Sonnenburg and colleagues[21] describe how outcomes should incorporate quality of life, and how the utility structure should contain all the relevant attributes,[21] but offer little in terms of appropriate methodology.

Charges and costs
It is important that the costs used in the model represent the value of resources rather than what is charged for them.[17] The methods used in transforming charges to cost should be explicit.

Adjustment over time/between countries
Adjustment for inflation should be based on the consumer price index, its healthcare components or one or more of its subcomponents, such as services or equipment. The choice between these will depend on whether the resource being costed is best represented by a general market basket or a healthcare market basket (ISPOR[13]). Purchaser power parities are the appropriate method for adjustments between countries.[13]

Half cycle correction
In state transition models with discrete time intervals, the model should include a half-cycle correction to adjust for the implicit bias of assuming that transitions are occurring at the end or beginning of the cycle.[7] This is most important when the difference in (quality-adjusted) survival between options evaluated in the model is less than the equivalent of one cycle.[13]

13

Data incorporation
In incorporating data into the model, measurement units, time intervals and population characteristics should be mutually consistent.[13] All data should be evaluated in terms of their appropriateness to the purpose of the model and their comparability with other data sources.[10] The use of the data should be justified, and should include an assessment of the importance of the parameter in addition to assessment of the importance of the value used.[10] It is important that justification is given for any choices made,[8] such as the reasons for choosing certain data over others.[10,17] Any differences between the data inputs of different models should be explained and justified.[10]

The analyst must make clear the source of all data and values used in the model and any assumptions that are made.[8,11,15,17–19,21] This will include event probabilities, healthcare resource utilisation (it is important to separate resource utilisation associated with different treatment pathways and justify how resource utilisation data are allocated to health states), utilities (including method of collecting utilities) and unit costs.[11,15,18] The inclusion of indirect (productivity) costs will depend on the condition being modelled, the perspective of the model and the intended audience. If indirect costs are included in the model, however, the source and values used must be specified clearly.[11]

The strengths and weaknesses of each data source (including inherent biases) should be described in sufficient detail for the reader to be aware of the type of the data source.[15,18] In particular, the reader should be able to distinguish between data that are collected from randomised studies and softer data such as expert opinion.[9] Where data come from trials, details of the results of the trial and of the study design, such as follow-up period and patient characteristics, should be given.[15]

It should be clear which data are being used for the base-case analysis. The range of values used in sensitivity analysis should also be present.[13,18] The methods used to determine the range of values should be justified.[15]

Data can be incorporated into the model as point estimates, or as a distribution. There is disagreement as to the appropriate method for incorporating data, and the method of data incorporation will be influenced by how the analyst evaluates parameter uncertainty within the model (see below). Halpern and colleagues[11]

describe the main advantage of the probabilistic approach as being that repeated sampling can assess the uncertainty around modelled outcomes.[11] Despite this, they recommend the use of deterministic modelling. They describe how a point estimate and a range of values for the likelihood of a given modelled event should be used. They justify this on a number of grounds. First, because of the difficulty in determining the distribution of event likelihood; second, because it is inappropriate to specify independent probability distributions where probabilities are correlated; and, finally, that in most circumstances, the additional time, expense and complexity involved in probabilistic modelling are not worth incurring. ISPOR[13] is also reluctant to recommend a probabilistic approach for input parameters based on sampling uncertainty, arguing that it is not always necessary or cost-effective.

Assessment of uncertainty
Uncertainty is ubiquitous in all health technology assessment decision models.[7] The dimensions of uncertainty have been described elsewhere[26] but, within modelling, researchers have tended to distinguish between parameter, and structural and methodological uncertainty, all of which require assessment.[8] Two exceptions in the guidelines are from Buxton and colleagues[9] and Eddy,[18] who both describe how uncertainty should be explored via sensitivity analysis but do not specify the types of uncertainty that may be relevant to be addressed in this way. Buxton and co-workers[9] explain how sensitivity analysis can assess the robustness of the model. Eddy[18] describes how it can indicate the importance of uncertainty and poor data.

Structural
The limitations of the evidence supporting the chosen model structure (the mathematical form in which parameter values are combined) should be acknowledged[17] and a sensitivity analysis using alternative model structures should be performed.[13] No proven method exists to evaluate these structural uncertainties except to compute cost-effectiveness estimates for each alternative structural assumption and to examine the appropriateness of the results.[22] A structural sensitivity analysis using a less aggregated form may provide reassurance that the structural assumptions do not materially affect the model's results.[13]

Methodological
Although this is an area of uncertainty that should be addressed,[8] just one guideline has explained

how methodological uncertainty, such as that associated with the appropriate discount rate, should be assessed by running the model with alternative discounting assumptions.[11]

Parameter
The principal focus of uncertainty in guidelines has been on parameter uncertainty. In addressing parameter uncertainty, it is important to distinguish between uncertainty resulting from the process of sampling from a population, and heterogeneity.[7] Where parameter estimates differ between patient subgroups, this should be described as heterogeneity, and analysis should be undertaken for the different subgroups individually.[7] The results of the decision model, therefore, should reflect differences between alternative patient subgroups in terms of treatment and management pathways and disease progression,[21] and it is important that modelled subgroups are disaggregated according to strata that have different event probabilities, quality of life or costs.[13,19] This is particularly important when recurrent event rates over time are correlated within subgroups.[13] The inclusion of alternative unit cost data can be regarded as a specific case of heterogeneity.[7] Halpern and colleagues[11] describe this as an adaptation of the model to a different setting.

The parameters for which sensitivity analyses are undertaken are a matter of judgement for the analyst.[13] Some suggest that parameters with the greatest level of uncertainty, such as those derived from expert opinion or those with greatest influence on the model outcomes, should be subjected to sensitivity analysis,[11,19] whereas others suggest that it should be the key clinical variables and the main cost drivers.[15,17] A clinician interpreting the results of the sensitivity analysis of a model may want to ensure that all clinically important variables have been subjected to sensitivity analysis and that the best/worst case scenarios tested meet or exceed expectations.[17] Nevertheless, the parameters and the values chosen for the sensitivity analysis should be presented and justified.[11,13,15,19]

There is disagreement about the most appropriate approach to exploring parameter uncertainty. Parameter uncertainty can be addressed by univariate, multivariate or probabilistic sensitivity analysis, analysis of extremes, joint confidence intervals, bootstrap techniques or Monte Carlo simulation.[14,19,22] ISPOR[13] recommends either deterministic or probabilistic sensitivity analysis and suggests that, where cohort simulation is used, probabilistic sensitivity analysis can be performed.[13] The group also explains that, if first order Monte Carlo simulation is used, evidence should be provided to show that random simulation errors are smaller than the effect sizes of interest.[13] Likewise, Monte Carlo simulations should be carried out with fixed random number seeds to minimise random simulation error.[13] Soto[19] recommends univariate analysis of key variables in the model to identify circumstances where the value of each variable has an important impact on results, but goes on to say that the appropriate approach to handling the uncertainty associated with multiple parameters is to undertake multivariate sensitivity analysis.

Halpern and colleagues[11] suggest that multivariate sensitivity analysis should be performed, thereby providing evaluation of a broader range of uncertainty than unidimensional analysis, but only on correlated parameters. They suggest that interpretation of independent modifications of uncorrelated parameters is difficult. Sculpher and colleagues[7] describe how parameter uncertainty can be reflected directly in the model by making the parameters random variables and incorporating a distribution as opposed to point estimates, but point out that the distributions used should be justified. Sonnenberg and colleagues[21] describe how parameter uncertainty is best represented using a probability distribution and Monte Carlo simulation rather than a point estimate, but point to difficulties in the interpretation of their results. Hay and Jackson[22] agree that there is controversy over the value of probabilistic sensitivity analysis.

If probabilistic analysis is undertaken, care should be taken to ensure that the interdependence between parameters is reflected in joint distributions,[13] and that it is second order uncertainty that is reflected.[7] The preferred method of obtaining input distributions is to use posterior distributions from formal meta-analysis and Bayesian analysis, but practical considerations may lead to the use of expert opinion.[13]

The most important reason for performing sensitivity analysis is to determine whether it would be worthwhile to seek better data for future decisions.[11,18] Recommendations for the conduct or future design of research to obtain further data can be based on informal interpretation of the implications of sensitivity analysis or formal value of information techniques.[13] Formal value of information analysis will determine optimal data to incorporate quantitatively[7] by balancing the

decision-making costs and consequences of obtaining and waiting for better data with the costs and consequences of synthesising currently available data to inform that decision.[20]

Consistency

Methods by which the validity or consistency of a health technology assessment decision-analytic model should be assessed have been a subject of debate among experts. As Eddy[18] commented in 1985, "There is no simple and universally applicable procedure for validating a model. Each case must be considered by itself." Because a gold-standard test for validity of outcomes rarely exists, assessment of the structure and inputs to the model, as well as the outputs, is required.[10] Methods for assessing the consistency of a model have been split into four intersecting groups: internal consistency, external consistency, between model consistency and predictive validity. Each of these groups is detailed below.

Internal consistency

The practical model should behave as the theoretical model predicts.[11,21] Sonnenberg and colleagues[21] describe some common errors that can form a set of rules by which the model can be critiqued from the perspective of structure and programming:

- invalid model syntax
- conditioning of an action on an unobservable event
- violations of symmetry
- failure to link variables
- failure to apply consistent bias
- incorrectly modelling the results of a diagnostic test
- incorrectly modelling a treatment.

Habbema and colleagues[14] and Halpern and colleagues[11] also discuss model symmetry, but do not suggest that violations from symmetry are always errors; rather, that they should be considered carefully because they may indicate errors. Habbema and colleagues[14] describe two types of symmetry, both of which require consideration:

- physiological symmetry: parts of the decision tree that relate to a patient's underlying physiology should be symmetrical because actions cannot affect the possible states of nature that may occur
- complications symmetry: morbidity, mortality or costs that can accrue from one action in one context can also accrue from the same action in another context.

The issue of consistent bias has also been raised in more recent guidelines. Halpern and colleagues[11] argue that the optimal outcome cannot be determined a priori, so the direction of bias cannot be set during the model development. They suggest that all inherent biases are logged and discussed in terms of both their direction and their likely impact on outcome.

Models should only be used after careful testing to ensure that the mathematical calculations are accurate and consistent with the specifications of the model, and that data have been incorporated appropriately.[7,10,13,20] The model should be mathematically well defined for all combinations of parameter values specified as feasible, and no such values should result in inconsistencies in the mathematical logic of the model.[8] Simple tests include ensuring that all probabilities lie between zero and one[19] and that outcomes following chance nodes sum to one.[11,17,19] The model should be able to match the data it used to estimate parameters, and failure to do this strongly suggests that the structure of the model is faulty.[18]

Sensitivity analysis using extreme or null values can be used to determine whether the model is behaving appropriately.[7,11,13,16,20,21] Likewise, the model can be rebuilt in a second software package and the results can be compared.[7,16] Any incongruities or departures from expected results must be investigated and explained or rectified. They may indicate errors in the model, or reveal a new insight.[21]

External consistency

The model should demonstrate face validity; its results should be amenable to intuitive explanation[10,13,16,19,20,22] and its structure should make intuitive sense.[11] The structure and incorporated data should be consistent with appropriate information and the outputs of the model should be assessed in relation to best available evidence.[8,13,20,22,] Such calibration data should be independent from data used to populate the model.[13,17,18,20] Buxton and colleagues[9] describe how the analyst should validate the model by assessing the level of agreement with the results of intervention studies, and Nuijten and colleagues[15] state that the results should be compared with other studies or expert opinion. However, it is unlikely that data will exist to compare the results of the model,[11] and there will always be an inevitable trade-off between using all available evidence for the model's development and having independent data against which to compare the model output.[16,18]

McCabe and Dixon[10] suggest that it may be appropriate to compare the final outputs of the model with prospectively collected data, but it is more likely that intermediate outputs of the model can be compared with independent data[10,11] such as the time in particular health states.[7] Any differences should be explained and justified.[10,15] It is important to note that other data will have been collected for different reasons than the development of a model, so lack of convergence of these data should not automatically lead to rejection of the model.[7] Models should also not be criticised if an opportunity for independent calibration data does not exist. They should be open to criticism if independent data do exist and the model has not been calibrated against them, or the two have been compared and the model is not consistent with the data.[13] It is likely that the process of calibration will be iterative and, in reporting methods of calibration, analysts should be explicit in their approach.[20] The model should be subject to peer review to ensure that it is what it claims to be.[20,21]

Between-model consistency

One approach to validating a model is to compare its results with those from other independent models that have addressed the same question.[7–11,13,15,16,19,21] The models should be developed independently from one another to allow such between-model corroboration.[13] The extent to which the alternative models reach different conclusions should be explained and justified.[7,10,13,15,16] However, as different models are developed at different points in time, lack of convergence of model results should not lead to automatic rejection,[7] but modellers should be able to explain why their model reaches different conclusions to other models of the same disease and question.[20]

Predictive validity

Some guidelines suggest that a model should demonstrate predictive validity[16,19,22] and that the best way to validate a model is to carry out a prospective study to ascertain that the results are similar.[19] Others have said that tests for predictive validity are valuable but not essential.[13,20]

Others have argued that models are intended to aid decision-makers and not necessarily to predict the future; therefore, it is inappropriate to expect predictive validity.[7,20] As Sonnenberg and colleagues[21] note, validity in the context of decision models refers to the ability of the decision model to recommend optimal decisions. Short of a clinical trial of a decision model, the validity of the

recommended decision cannot be assessed because, *ex ante*, there is no gold standard for the quality of the decision.

Comparing the outcomes predicted by the model for a new and previously unobserved programme with the actual outcomes of the programme once it is implemented may be meaningless, as the actual programme may be implemented under different conditions to those assumed in the model.[18] It is inappropriate, therefore, to expect a model to predict the future accurately because it can only incorporate the data available at the time the model is brought to bear on the decision.[7,20] The criteria for determining whether tests for predictive validity are required depend on the benefits in terms of improving the model for decision-making and the costs of delaying the flow of information while obtaining additional data.[13,20] A model should not be criticised for failing to predict the future, but should be amenable to respecification and recalibration as new data become available.[7,13] It is not necessary that every data point or structural assumption in the model is validated in prospective study, because results should be reported conditional on input data.[13]

Conclusions

It is important to note that guidelines relating to specific disease areas were not included in this review. Although the possibility that this exclusion will result in some useful material having been missed cannot be ruled out, it is unlikely that much of this would have been useful to a general health technology assessment audience.

This chapter has described the range of attributes for decision-analytic models in health technology assessment described in the literature. These have been grouped under the general themes of structure, data and consistency. In some areas these guidelines present a consistent message; for example, the need to be explicit in presenting an analysis. In other areas, the attributes presented in different guidelines conflict; for example, the extent to which the structure of, and data inputs within, a model should be determined by the data available. Chapter 3 attempts to synthesise these guidelines and to provide a checklist that can be used as a framework for critical appraisal.

Several issues are not currently addressed within the available guidance. These include literature searching for parameter estimation in decision

models and adjusting for bias in treatment effect estimates from observational studies used in decision models. These two specific aspects are considered further in Chapter 4. Other issues include methods for evidence synthesis,

appropriate extrapolation methodology and adjusting model parameters between jurisdictions. Key references relating to these areas can be found in Appendix 7.

Chapter 3
Development of a best practice guideline

The purpose of this chapter is four-fold. First, it provides a best practice guideline in decision modelling for cost-effectiveness analysis. This is done under the main themes that have emerged in the literature and incorporates much of the evidence available from the systematic review of best practice guidelines reported in Chapter 2. That review has shown that some attributes of good practice are contradictory to others. Given the need to come up with a consistent guideline, decisions have been taken on the most appropriate attributes. Where aspects of published guidelines have been rejected, this is explained and justified.

Second, to provide a practical framework for critical appraisal, a checklist has been constructed using the statements of best practice. Third, to assess how comprehensive and useful this checklist is, it has been applied to three recent decision-analytic models that have been developed for the NICE technology appraisal process (case-study models). Finally, to assess the potential value of the synthesised checklist in general, and for the specific purpose of guiding quality assessment for the NICE technology assessment process, an Expert Advisory Group (EAG) was convened. The group's role was to advise on all elements of the project. In particular, it had an important role in the final element, which related to the implications of this research for what might be expected of future decision-analytic models relating to the NICE technology assessment process. The EAG was drawn from users (in particular NICE and the TAR groups) and methodologists in the field. The EAG was not involved in the production of the report in any way.

Developing the synthesised guidance

Using the summary of guidelines developed in Chapter 2, a statement of best practice was developed. Each theme and subtheme was taken in turn, and its relevance was discussed in relation to the development of general guidelines for decision-analytic modelling in health technology assessment. Areas of guidance that were relevant to economic evaluation in general, rather than decision-analytic modelling specifically, were omitted. Where previous guidelines were contradictory, a consensus decision was taken by the research team regarding the most appropriate item for the synthesised guideline. A checklist was constructed from the synthesised guideline, using the suggested headings and statements. This checklist was then applied to three recent decision-analytic models undertaken as part of the NICE appraisal process and fully reported in detail in a publicly available Technology Assessment Review (TAR). The models have been chosen to reflect the variation in structure of decision-analytic models undertaken as part of the NICE process. This element of the study sought to ascertain whether a checklist of best practice is useful in establishing the quality of a model.

Statements of good practice

It should be emphasised that the statements below are specific to decision-analytic modelling in economic evaluation and do not cover more general attributes of good practice in economic evaluation such as those published in other guidance,[23] unless those attributes may be expected to influence the modelling process, for example, time horizon.

Structure
The structure of the model relates to the definition of pathways and/or events, their sequencing and the causal relationships between them.

Statement of decision problem/objective
- There should be a clear statement of the decision problem prompting the analysis. This should include details of the disease or condition under evaluation, the patient group and the diagnostic and/or treatment pathways.
- The objective of the evaluation (e.g. the maximand) and of the model should be defined. These should be consistent with the stated decision problem.
- The primary decision-maker should be stated clearly as this will have implications for the choice of relevant data.

- It is not necessary to state the type of economic evaluation that will be undertaken as the type of evaluation will be implicit from what is reported about the model (i.e. perspective, time-frame, outcomes, etc.).

Justification of the modelling approach

- It is unnecessary to provide justification for choosing a modelling framework per se.[10,15] This is because a model is simply an analytical framework with the purpose of synthesising the relevant evidence that is available at a particular point in time, and bringing this to bear on the particular decision problem under consideration. It can be argued that most, if not all, economic evaluations include some form of modelling.
- Decisions have to be made regardless of data availability.[1] For this reason, the statement that models can only be developed when alternative interventions, outcomes and probabilities can be specified or estimated (as suggested by Halpern and colleagues[11]) would be unhelpful in a synthesised guideline.
- Some previous guidelines have argued that there are other study designs (e.g. trials) that are superior to modelling studies.[11,19] However, trials are sources of data, rather than an analytical framework as such, and most trial-based economic evaluations require some form of additional modelling to address the decision problem.
- The statement about the decision problem should justify the modelling approach and the key features of the model. These include the time horizon, disease states/pathways and data inputs (see later statements).

Statement of scope/perspective

- The perspective of the model, in terms of which costs and consequences are considered relevant, should be stated clearly, and the model inputs should be consistent with that stated perspective and overall objective.
- The choice of appropriate perspective is a judgement that will often depend on the ultimate audience. A guideline relating to decision-analytic modelling for health technology assessment in general should not be prescriptive about this choice. Clearly, guidelines from particular decision-making agencies (e.g. NICE) are likely to specify the appropriate perspective for their decisions.[27]
- The scope of the decision model should be specified and justified. The model's scope refers to its limits or boundaries. Some of these factors will be defined within the statement of the decision problem; for example, the patient group.
- Although data availability should not be an overriding factor in the development of the structure of the model, it is possible that data availability may need to refine a model's scope.
- The outcomes estimated in the model should reflect the perspective and scope of the model and should be consistent with the objective of the evaluation.
- Arguments in the literature favouring final and/or lifetime outcomes over more short-term outcomes[19] are appropriately part of general guidelines for economic evaluation. However, these need not be defined in a modelling guideline as the appropriate outcomes will be defined by the definition of the decision problem, and by the objective and scope of the model.
- Clinical events for which a difference between strategies is not logically or theoretically expected can be excluded. However, clinical events cannot be omitted purely on the basis of no statistically significant difference between groups, as has been suggested.[15] This is because what constitutes 'statistical significance' is ultimately an arbitrary decision.
- The development of a model should be considered an iterative process, so clinical events should be included if they may be expected to differ in the future, when more data are available.

Rationale for structure

- The structure of the model should be consistent with a coherent theory of the health condition under evaluation.
- The treatment pathways (disease states or branches) should be chosen to reflect the underlying biological process of the disease in question and the impact of the intervention.
- The model structure should describe a series of causal relationships, and justifying those relationships will provide a rationale for the chosen model structure. In the case of assumptions being made, these should be transparent and justified (see below).
- All sources of data used to develop and inform the structure of the model (i.e. the theory of disease) should be described.
- The structure of the model should reflect the essential features of the disease/condition irrespective of data availability, although it is inevitable that absence of particular data may limit a model's scope (see above).
- The structure should not be dictated by current patterns of service provision, as a change to

those patterns is what is being evaluated. However, the model should include one or more strategies that are representative of current practice.

Structural assumptions
- All structural assumptions should be transparent and justified and should be reasonable in the light of the needs and purposes of the decision-maker.
- Uncertainty associated with structural assumptions should be explored via sensitivity analysis (see section 'Assessment of uncertainty', p. 23).

Strategies/comparators
- There should be a clear definition of the options under evaluation. This should be included in a statement of the decision problem.
- All feasible and practical options relating to the stated decision problem should be evaluated. However, options should not be constrained by the a priori views of the decision-maker, such as those relating to outcomes, data availability or the constraints of current practice.
- There should be justification for the exclusion of a feasible option.

Model type
- The appropriate model type will be dictated by the stated decision problem and the choices made regarding the causal relationships within the model.

Time horizon
- A model's time horizon should extend far enough into the future to reflect important differences between options.
- Lifetime horizons are appropriate for most models, but shorter horizons can be justified in some cases on the basis of there being no differences in mortality, long-term morbidity and cost between options.
- It is important to distinguish between the time horizon of the model and the duration of treatment effect. Both should be specified and justified.
- The lack of long-term follow-up data should not be used to justify a failure to extend the horizon of the model to the period relevant to the decision.

Disease states/pathways
- Here disease states relate to those used as part of the structure of a state transition model; these may be more appropriately defined as pathways in decision tree models.

- Disease states should reflect the underlying biological process of the disease in question and the impact of interventions.

Cycle length
- For discrete time models, the cycle length should be dictated by the natural history of disease. It should be the minimum interval over which the pathology or symptoms are expected to alter.
- Routine clinical practice with regard to follow-up should not determine the choice of cycle length.

Parsimony
- Parsimony relates to the complexity of the model, and statements relating to parsimony per se, such as those given in previous guidance,[7–9,11,13] do not provide helpful guidance regarding the quality of a model.
- Model complexity is neither good nor bad. If a high level of complexity is required to model the disease adequately and to address the decision problem, then this should not be regarded as a weakness of the model.
- The level of complexity is determined by the elements of structure as described previously. So long as these have been described in sufficient detail, and reasonable justification provided, the level of complexity is acceptable.
- A distinction should be made between transparency and accessibility. The former relates to whether choices about structure and data are explicit and clear in a model. The latter is concerned with the ease with which the model can be changed and alternative results generated. There should be no compromise in the development of a model simply to make it accessible to non-specialists.

Data
This section relates to the data that are identified, preanalysed and incorporated into the model.

Data identification
- Methods for identifying data should be transparent.
- It should be clear that the data identified are appropriate given the objectives of the model.
- There should be justification about any choices that have been made about which specific data inputs are included in a model.
- It should be clear that particular attention has been paid to identifying data for those parameters to which the results of the model are particularly sensitive. Identification of all available data for every parameter may not represent the best use of research resources.

- There should be a description of the quality of the data identified for use in the model.
- Where expert opinion has been used to estimate particular parameters, sources of opinion and methods of elicitation should be described. Given that decision models should appropriately assess uncertainty in parameters, enforcing consensus in expert opinion (e.g. using standard Delphi methods) is inappropriate.

Premodel data analysis

Here, premodel data analysis relates to the analysis of data before they are incorporated into a decision model.

- All methodology should be described and based on justifiable statistical and epidemiological methods.
- There are numerous data analysis issues relevant to decision models, and it is not possible to cover each of these in a general guideline. However, some are useful as a series of examples that relate to particular types of data. These are detailed below.

Baseline data
- Baseline probabilities may be based on natural history data derived from epidemiological/observational studies or relate to the control group of an experimental study.
- Rates or probabilities based on intervals, which differ from the cycle length, should be transformed into transition probabilities appropriately.
- If there is evidence that transition probabilities may vary, this should be reflected in decision models. The use of time-dependent or constant transitions should be justified.
- If a half-cycle correction has not been used on all transitions in a discrete time state transition model (costs and outcomes), this should be justified.

Treatment effects
- Relative treatment effects derived from trial data should be synthesised using recognised techniques such as random/fixed effects meta-analysis.
- The methods and assumptions that are used to extrapolate short-term results to final outcomes (e.g. trial-based intermediate outcomes to ultimate health effects) should be documented and justified. This should include justification of the choice of survival function (e.g. exponential or Weibull distributions).
- Assumptions regarding the continuing effect of treatment once it has ceased should be

documented and justified. If evidence regarding the long-term effect of treatment is lacking, alternative assumptions should be explored through sensitivity analysis.

Costs
- Costs incorporated into the model should be justified and the sources of these data should be described. Methods should accord with standard guidelines for costing within economic evaluation.
- Discounting methods should accord with general guidelines for economic evaluation. In many instances, specific discount rates are likely to be defined by the relevant decision-maker.

Quality of life weights (utilities)
- Utilities incorporated into the model should be appropriate for the specified decision problem.
- Sources of utility weights and methods used to derive them should be justified and referenced appropriately.

Data incorporation
- All data incorporated into the model should be described and sources detailed.
- The sources of all data should be given and reported in sufficient detail to allow the reader to be aware of the type of data that have been incorporated.
- Although ideal, it is not always possible for the data incorporated into a model to be directly comparable (e.g. based on the same patient sample). Where data are not directly comparable, the choices and assumptions that have been made should be explicit and justified.
- Data can be incorporated into a model as point estimates or as distributions. It should be clear how this process has been undertaken.
- In the choice between a deterministic or probabilistic modelling approach it must be recognised that:
  - when a model is non-linear, an unbiased estimate of mean costs and effects can only be obtained by incorporating the uncertainty in mean parameter values directly through probabilistic analysis
  - the combined effect of parameter uncertainty is only truly reflected when data are incorporated probabilistically.
- If data have been incorporated as distributions as part of probabilistic analysis, the choice of distribution, and how the distribution has been defined, should be described. The choice of distribution should reflect the characteristics of the parameter. Given the need to quantify uncertainty in mean cost-effectiveness results,

distributions should reflect parameter uncertainty (e.g. uncertainty in mean inputs). This has been called second order uncertainty.[3] Assumptions regarding the correlation between parameter values should be stated.

Assessment of uncertainty
- Modellers should distinguish between the four principal types of uncertainty detailed below. If a particular form of uncertainty has not been addressed, this should be justified.
- Modellers should use both probabilistic sensitivity analysis and scenario analysis to explore the various forms of uncertainty.

Methodological
- Methodological uncertainty relates to whether particular analytical steps taken in the analysis are the most appropriate (e.g. a fixed effect versus random effect assumption in data synthesis). This form of uncertainty should be handled via scenario analysis (e.g. running alternative versions of the model with different inputs).

Structural
- There should be evidence that structural uncertainties have been evaluated using sensitivity analysis.

Heterogeneity
- It is important to distinguish between uncertainty resulting from the process of sampling and from a population and variability due to heterogeneity (i.e. systematic differences between patient subgroups). The latter should be dealt with by running the model separately for different subgroups.

Parameter
- If data have been incorporated into the model as point estimates, univariate or multivariate sensitivity analysis should be used to express parameter uncertainty. The ranges used for sensitivity analysis should be stated and justified.
- Standard approaches to sensitivity analysis, such as varying one or more parameters at one time and examining the effect on the results, are limited as the choice of parameters to vary (and the range over which they are varied) is arbitrary.
- In addition to obtaining an unbiased estimate of the mean costs and outcome(s), probabilistic sensitivity analysis is the most appropriate method of handling parameter uncertainty because it facilitates the assessment of the joint effect of uncertainty over all parameters[26] (see section 'Data incorporation', p. 22).

Consistency
Consistency relates to the quality of the model overall; that is, to the combination of structure, assumptions and data inputs.

Internal consistency
- There should be evidence that the internal consistency of the model, in terms of its mathematical logic, has been evaluated. That is, a change in a particular parameter value should have a predictable effect on results.
- Sensitivity analysis, using extreme or null values, can be used to assess internal consistency.
- One check on internal consistency is to programme the model using a second software package to check consistency with the first results.

External consistency
- The results of a model should make intuitive sense and any counterintuitive results should be explained.
- All relevant available data should be incorporated into a model. Data should not be withheld for purposes of assessing external consistency.
- Calibration data are those that have not been directly incorporated into a model but that can be compared with intermediate results of the model. Data useful for the calibration of the model will not have been directly incorporated into the model because such data are usually aggregated or represent intermediate outputs. For example, in developing a cervical cancer screening model, an intermediate output of the model might be the age-specific incidence of cervical intraepithelial neoplasia (CIN) III. National statistics relating to the incidence of CIN III can be compared with this model output. Such data will not have been directly incorporated into the model, although they may have been used in combination with other data to inform a particular element of the model such as a transition probability. Any major differences between calibration data and those from the model may not necessarily uncover a weakness in the model, but any differences should be justified.
- The results of a model should be compared with those of previous models and any differences should be explained.

Predictive validity
- There is no empirical validity test of a decision-analytic model. Models are not intended to predict the future, rather they are developed as

aids to decision-making at a particular point in time.

- A model should be amenable to respecification as more data become available.

Applying the guidance: the use of case studies

Background and methods

A checklist was developed (see Appendix 3), on the basis of the synthesised guideline defined above in terms of statements of good practice, which forms the basis of the resource available to those undertaking critical appraisal of decision models. The checklist was used as part of three case-study reviews to assess its value. These case studies are models that have previously been used within the NICE technology appraisal programme. The case studies were chosen on the basis that they represented alternative model types, addressed both preventive and acute treatments, covered different time-frames and had varying degrees of complexity. This exercise aimed to identify whether the checklist was useful in reviewing HTA decision models, and whether application of the checklist was consistent between reviewers.

The three models were:

- Payne N, Chilcott J, McGoogan E. Liquid-based cytology in cervical screening. A report by the School of Health and Related Research, University of Sheffield, for the NCCHTA on behalf of NICE; May 2000[28]
- Kendrick M, Johnson M. Long term treatment of multiple sclerosis with interferon beta may be cost effective. *Pharmacoeconomics* 2000; 18:45–53.[29]
- Turner D, Wailoo A, Nicholson N, Cooper N, Sutton A, Abrams K. Systematic review and economic modelling for the prevention and treatment of influenza A and B. A report by the Departments of Epidemiology and Public Health and Microbiology and Immunology, University of Leicester and SCHARR, University of Sheffield, for the NHS R&D HTA Programme; April 2002.[30]

Two independent reviewers (ZP and LG) applied the checklist to each of the three models. A modelling expert (E Fenwick), who had not been involved in the process of developing the synthesised guideline, also provided a review of each of the models without the use of the checklist. She was asked to undertake her review as

if peer-reviewing an article submitted to a journal. The purpose of this was to identify whether there were particular issues with, and weaknesses of, a model that were not being picked up by systematic application of the checklist.

Case-study reviews

The full reviews of each model undertaken using the checklist are provided in Appendix 4(a) and (b). An overall summary, made by both of the reviewers, is provided in *Table 3*.

Consistency between the two reviewers using the checklist

The checklist cannot be used as a points system to determine the relative quality of a model. Rather, it provides general questions, which help the reviewer to focus on the details of the reporting of the key elements of a model.

This exercise has shown that the checklist is useful in so far as it provides a systematic approach to model review and focuses the reviewers on the model's (or the reporting of the model's) strengths and weaknesses.

The two reviewers reached similar conclusions regarding their overall assessment of the models, although there were some differences in answers to each specific question within the checklist. Upon further investigation, there were two reasons for this. First, there were differences in interpreting the term 'justification'; the extent to which a structural or data decision/assumption is 'justified' relies on a value judgement by the reviewer. Second, one of the reviewers had a detailed knowledge of one of the disease areas, which made the review of the model more straightforward because of an awareness of the data from which the model was developed.

The issue of knowledge of the disease area was important in each of the case studies. Without prior knowledge, the reviewers found it difficult to determine whether the chosen model structure and structural assumptions were appropriate, or whether all potentially feasible strategies were evaluated. This has implications for the peer review of decision models, either for journals or for decision-making organisations such as NICE.

Comparison of the reviews with and without the checklist

Here the comments on the models of the reviewers using the checklist are compared with

TABLE 3 *Results of checklist application to case studies*

Model	Element	Reviewer 1	Reviewer 2
Payne et al., 2000[28]	Overall assessment	The authors are transparent regarding the additions and assumptions that are made to the Sherlaw–Johnson model in order to evaluate LBC but there is little discussion about the appropriateness of the choice regarding which model to reproduce, or the data that it incorporates. However, some of the main caveats of the original model are discussed. Because of a detailed understanding of the area it is possible to conclude that the structure of the original model (i.e. Sherlaw–Johnson model) is appropriate, and that its reproduction here, with the addition of LBC, appears sound	This model was an update of a previously published screening model, looking at the cost-effectiveness of alternative screening strategies for cervical cancer. It appeared to be of a reasonable quality. The checklist allowed the identification of key elements of the model and focused the reviewer on its relative strengths and weaknesses
	Structure	The implicit assumption of the model is that the structure of, and data incorporated into, the Sherlaw–Johnson model are optimal. Without prior knowledge of the area it is impossible to determine the appropriateness of this assumption. Because this model is a reproduction of a previously published model, the authors give few details of the causal relationships in the model, or the data from which the structure was originally developed. Notwithstanding these issues, the authors are clear about the decision problem, the objectives of the model and the main structural assumptions	Very few details were available on the model structure; however, it would seem to be an appropriate representation of the disease process given the objectives and perspective of the model, which were clearly defined
	Data	All data used in the model are referenced appropriately, although it is not explicit why particular choices (i.e. choice of published model to reproduce and cost assumptions) were made. There is no detail regarding the pre-model data analysis methodology, presumably because the data are largely taken from a previously published model. Structural uncertainty is not addressed, but the other types of uncertainty are evaluated. The use of triangular distributions for parameters was based on a search for simplicity, which may mean that the results do not adequately reflect the uncertainty associated with parameter inputs	Very little detail was given about the original model (from which a lot of the model data derived); therefore it was difficult to assess the methods used to identify and extract data as well as the quality of the model data. All data sources were, however, stated and referenced. Parameter and methodological uncertainty was addressed in the sensitivity analysis
	Consistency	Both internal and external consistency is addressed in detail	Some model validation, in terms of a comparison with the earlier model and published data was undertaken, which suggested that the model was a fair representation of the disease process and produced results that were not counterintuitive
Kendrick and Johnson, 2000[29]	Overall	The difficulty with this model is that the write-up lacks transparency. Furthermore, a number of assumptions have been made, which appear to have little justification. The appropriateness of the statistical techniques employed is questionable	This was neither a state transition nor decision model; therefore, many of the questions on the checklist did not apply. A regression model was used to extrapolate existing trial data on the effectiveness of β-interferon for MS, to a longer-time frame (up to 20 years). The write-up was particularly poor and as such a lot of the checklist questions were unanswered

continued

TABLE 3 Results of checklist application to case studies (cont'd)

Model	Element	Reviewer 1	Reviewer 2
	Structure	There is no clear statement of the decision problem, objectives or scope of the model. There is no transparency in details regarding the structure of the model. A graphical representation of the model itself is not presented. It is clear that important aspects of MS (e.g. its relapsing–remitting nature) are not taken into account. The progression of EDSS is a proxy for progression of MS, but EDSS is an ordinal scale and the assumption of constant progression over time is questionable. There is no evidence to provide a basis for the extrapolation assumption. The assumed relationship between EDSS and utility is also questionable on similar grounds. Death does not appear to be taken into account, and the chosen model time-frame appears arbitrary	Very few details were reported on the model structure; therefore, it was not possible to assess the appropriateness of the techniques used. Statements of objective and scope were not defined
	Data	The authors are not sufficiently clear about the clinical data identified and incorporated into the model. The cost and utility data are referenced appropriately, but the premodel data analysis methods required to incorporate QALYs into the model are questionable. There is no assessment of any form of uncertainty	Very few details of the data incorporated into the model and techniques used. No assessment of uncertainty
	Consistency	There is no evidence that the internal consistency of the model was checked before use. The results are compared with those from other models and differences in results are explained in terms of the assumption regarding treatment effects beyond 2 years	Some comparison with shorter term models. The logic of the model results was not tested
Turner *et al.*, 2002[30]	Overall	This model adheres to the majority of the principles of good practice in decision modelling for health technology assessment. Because of the transparency in the reporting of this model, one was able to understand the process undertaken by the researchers in addressing the particular research question. The main difficulty is in identifying whether the chosen model structure is appropriate	This model was developed to assess the cost-effectiveness of alternative prophylactic vaccines for preventing influenza. A decision tree model was used to assess treatment and prophylaxis for four patients groups. The report included a particularly good level of detail on the methods and data used, which made the checklist easy to apply
	Structure	The structure of the model is specified clearly, although there is no justification for the chosen pathways or the causal relationships between events. However, all assumptions are defined clearly, along with the likely impact on results. Without detailed knowledge of the area it is difficult to determine whether all potentially feasible options have been assessed. It is difficult to identify the time-frame of the model, and impossible to determine whether it is sufficient to reflect the important differences between options	The report contains a clear statement of objective, perspective and scope of the model. The structural assumptions made in the model were described in detail, but not always justified. The structure of the model is described in detail; however, it is difficult to assess its appropriateness without prior knowledge of the disease area

TABLE 3 Results of checklist application to case studies (cont'd)

Model	Element	Reviewer 1	Reviewer 2
	Data	The data identification and incorporation methods are transparent. All premodel data analysis methodology is based on clearly reported epidemiological/statistical methods and reported in detail in appendices to the main report. The only shortcoming is that expert opinion was used for one of the main parameters in the model, but there are no details of the methods used to illicit values. Uncertainty was addressed with probabilistic sensitivity analysis and a series of one-way sensitivity analyses	All data are referenced and described in detail. Expert opinion used to derive EQ-5D scores for adverse events; however, details of methods used were not included
	Consistency	There is no evidence that the internal consistency of the model had been checked before the model was evaluated, but the results of the model are compared with those of other models and differences explained appropriately	Model results compared with three other models. No internal validation of the model

EDSS, Expanded Disability Status Scale; EQ-5D, measure of health status for use in evaluating health and healthcare; LBC, liquid-based cytology; MS, multiple sclerosis.

those of the expert reviewer, who did not use the checklist.

Liquid-based cytology in cervical screening[28]
The expert reviewer concluded that the report suggested a well-constructed model, populated and analysed using probabilistic methods.

All major concerns regarding the modelling methods used, which were identified by the expert reviewer, were also identified by the reviewers using the checklist. The expert reviewer questioned further some of the simplistic assumptions that were made in the model, such as a constant risk of cervical cancer between the ages of 18 and 64 years and a constant mortality risk.

In addition, the expert reviewer noted that the authors had not excluded weakly or extendedly dominated options before calculating incremental ratios.

β-Interferon for multiple sclerosis[29]
Both the reviewers and the expert agreed that this report provided insufficient details about the model. The model seems to fall short of the guideline in several places. Specifically, there were problems with data incorporation, methods of analysis and the way in which the model was structured.

Neuramindase inhibitors for prevention of influenza A and B[30]
Both the reviewers and the expert agreed that the report provided a good description of what appeared to be a good quality model. All major concerns regarding the modelling methods used, identified by the expert reviewer, were identified by the reviewers using the checklist. The expert reviewer questioned further some of the simplistic assumptions that were made in the model, such as the assumption that vaccination for children did not impact on hospitalisation rates and the exclusion of the side-effects of Zanamivir.

In addition, the expert reviewer noted that the authors had not calculated incremental ratios correctly, comparing treatments with no treatment, rather than the next best option.

The potential value of the synthesised checklist
A meeting involving all members of the EAG and the project team was convened towards the end of the project, to inform the various elements of the research and to discuss the potential value of the synthesised guidance and checklist, both for the NICE appraisal process and for health technology assessment more generally. Names and affiliations of members of the EAG are given in Appendix 5.

Format of meeting
Copies of the draft report were circulated to all members of the EAG and project team before the meeting. Attendees were asked to focus on the content of the synthesised guidance and checklist, and to consider how these might be of use in future decision-analytic models relating to the NICE technology appraisal process and health technology assessment in general. Comments made during the meeting were documented and are summarised below, along with the project team's response.

Feedback on the synthesised guidance
In general, the feedback from the EAG was positive. A number of suggestions for changes to the checklist and guidance were made, and comments about how the material contained in the checklist corresponded to the statements made in the accompanying guidance (Appendix 6a,b). These have been incorporated into an amended version of the checklist (Appendix 6c). At the EAG meeting, there was also a discussion of the potential role of the checklist in the NICE technology assessment and appraisal process. This discussion is picked up in Chapter 5.

Conclusions

The synthesised guideline presented in this chapter essentially provides a list of statements of good practice in decision modelling for health technology assessment. This list covers the key attributes that previous authors have identified as being important in critically assessing models. Their attributes are not without controversy and, in assembling the guideline, it has been necessary to reach judgements about appropriate methodology where the literature indicates some inconsistency. Although explanations have been provided for how these conflicts have been resolved in the synthesised guideline, it is recognised that different perspectives may be adopted.

The synthesised guideline has been used to develop a checklist, which aims to provide a framework for reviewers to use when critically appraising a model. Given that it is not possible to codify every aspect of a decision model, the checklist should not be relied upon as a stand-

alone method of assessing quality. Rather, it provides the reviewer with a way of structuring their review and complements the reviewer's understanding of the specific details of the decision problem and the technologies under evaluation.

The checklist has been used as a basis for reviewing three decision models developed as part of the NICE appraisal process. These reviews have been compared with the assessment of an experienced modeller who considered the three models without the aid of the checklist. It is recognised that this does not represent a validation of the checklist. Given additional time and resources, more could be learned about the checklist's value through its use by a larger number of reviewers to assess a range of different models. It is hoped that this may be possible in due course (see Chapter 5). It should be noted, however, that designing an experiment to assess the quality of a review with and without the checklist is not straightforward. This is because, if this comparison is undertaken using different reviewers (i.e. one reviewer uses the checklist, the other does not), there is always likely to be some difference between reviewers' experience, both of decision modelling generally, and of a specific technology or disease area. If the experiment is undertaken using the same reviewer (i.e. an

individual reviews the model with and without the checklist), identifying the true incremental value of the checklist is not easy.

Despite the limitations of the assessment of the checklist here, the exercise provides a starting point for its practical use and continuous development. In general, the checklist appears to have performed well, in terms of identifying those aspects of the model that should be of particular concern to the reader.

There is a common problem when reviewing models. This relates to the assessment of the appropriateness of the model structure and structural assumptions. A general checklist cannot ask specific questions pertinent to particular models, yet experts cannot assess appropriateness without detailed knowledge of the disease area.

As a final point, this exercise has illustrated the importance of using the checklist of model quality in conjunction with a more general checklist of guidelines on economic evaluation methodology, as the checklist does not cover issues such as the calculation of incremental cost-effectiveness ratios and the reporting of results. Application of the checklist should not be used as a substitute for critical appraisal.

Chapter 4
Key methodological and practical issues not covered in existing published guidelines

Given that decision models are being used more widely as part of health technology assessment, it is important that the rigour of such studies is constantly enhanced. The main focus of the report has been on general guidelines for decision modelling, with the objective of providing a synthesised guideline to embody good practice. There is, however, a large number of specific issues for methods in decision modelling. For example, the preanalysis of data, before their incorporation into a model, is an area that, in principle, relates to the whole of biostatistics and clinical epidemiology. In other words, providing a general guideline for decision modelling is no substitute for the continuous development of specific methods relating to every stage in modelling.

The purpose of this chapter is to consider two specific methods areas in decision modelling that have received relatively little consideration in the literature. Given the time constraints of this project, these topics were selected on the basis of the team's experience, rather than any systematic review of the methods literature. The first methods topic is the identification of parameter estimates from the literature. Previous guidance suggests that data identified for use in decision-analytic models should represent optimal estimates, given resource and time constraints. However, there is little guidance on the way in which data, other than the main treatment effects, can be identified in an efficient manner. The question of how to search and identify data for all parameters within a decision-analytic model is, therefore, the focus of the first topic presented here.

The second topic relates to bias in parameter estimates. Despite an emphasis in the methods literature on the need to derive treatment effect estimates from randomised controlled studies, which, in principle, exhibit absence of selection bias, this may not be possible for all decision problems. In the absence of trial data, there is little guidance on how to adjust estimates of treatment effect from observational studies where there are risks of selection bias. There is an issue, therefore, of how one should take account of such bias.

A scoping review was undertaken to identify some of the key issues for both of these methods topics. To identify literature on searching for parameter estimates for models, the results of the searches in Chapter 2 were sifted, and a search was conducted of the Cochrane Methodology Register (CMR) using the terms: data source*, search*, database* with economic, model* or quality next life (* is the truncation symbol in the Cochrane Library). Searches were also carried out for the literature on bias although, as explained on p. 48, the results of these searches were not used.

In addition, a bibliography (Appendix 7) provides a list of key references relating to a range of other specific methods topics relating to the design and construction of decision-analytic models. This reference list was assembled by a group of experienced decision modellers well acquainted with current literature. The reference list is not designed to be exhaustive, but it is intended to provide a starting point for developers of models to ensure a structured and methodologically sound approach to building decision-analytic models for health technology assessment.

Appropriate methods for the identification and quality assessment of secondary parameter estimates such as utilities, costs, incidence and prevalence

Background
For the NICE TAR process (if not for decision-analytic models more generally), the identification and synthesis of treatment effect data for incorporation into models is currently carried out in a well-defined and systematic way.[31,32] However, the methods used to identify other data needed for the decision-analytic model are not as well established. In thinking about information sources for estimating all the parameters within a decision model, the following issues can be considered.

Decision models will usually include a large range of different types of parameters, such as treatment effects, baseline event rates (which may relate to the natural history of the condition), quality of life effects, health state values (or utilities), resource use and unit costs. It will very rarely be the case that one source of information will provide data for all these parameter estimates. Indeed, the most appropriate source of information is likely to differ markedly for the alternative parameters. For example, RCTs and systematic reviews of RCTs will almost certainly be the focus of any search for treatment effect parameter estimates, as these study types should provide estimates that are least susceptible to selection bias. However, many parameters within a decision model are not easily populated by RCTs, as trials often have a limited duration of follow-up, are often carried out in a clinical setting and do not necessarily involve real-life data or real-life patients (particularly vulnerable groups of patients). Therefore, this focus on trials may be inappropriate for such parameters and other study designs may need to be considered.

The process of thorough searching for information to populate each parameter in a decision model is extremely resource intensive. Decision models often have to be developed rapidly to inform particular decisions at a certain point in time. For example, models for the NICE TAR process are developed over a period of less than 6 months. Given such time constraints, it is rarely possible to undertake thorough searches to inform all the parameter estimates. There are also budget constraints with this approach as, in addition to staff time, there is the cost of document acquisition for potentially relevant papers identified from the results of these searches.

As recognised in the guidelines review in Chapters 2 and 3, there are two important features that should be in place in the process of identifying evidence for decision models, even if thorough searches are not possible. The first is that searching should be systematic. The search process and how the parameter estimates were ascertained should be clearly described and, in principle, the results should be reproducible by another researcher. The second feature is that the searching should be efficient. That is, given the resource constraints, searching should focus on those parameters that are expected to have the largest influence on the results of the model. In other words, priorities need to be set in searching for information to populate a decision model, and focusing on those parameters to which the results and conclusions of a model are expected to be most sensitive is likely to increase efficiency.

So far, it has been assumed that the structure of the model is developed and then the literature is searched to identify information to inform the parameter estimates required in the model. But how is model structure determined? Although the structure is often based on expert advice, the modellers' prior knowledge and the structure of earlier models, literature searching may also be able to inform the model structure. Indeed, as part of the process of literature searching for parameter estimation, it is possible that information may be identified that suggests a change in the model structure. This case study will address some of these issues discussed above.

Objectives

The data required to populate some model parameters may be quite specific, thus requiring a different approach to that of searches for effects data. One way to create models based on the best evidence and to limit bias in populating models could be to carry out a range of focused searches. In this part of the project a preliminary investigation was undertaken into the feasibility of carrying out focused search strategies to populate a case-study model, and approaches to conducting such parameter-specific searches were explored.

Methods

A dialogue was initiated between a decision modeller and an information specialist to produce searches that aimed to provide relevant records from a relatively small sample of records. This enabled the decision modeller to specify in detail the data required to populate the model. In this case the questions related to a model developed to assess the cost-effectiveness of prophylactic antibiotics for preventing urinary tract infections (UTIs) in children.[33] A list of questions, relating to specific types of model data (and their distributions), was generated to help to focus the searches. It was also hoped that the data identified would inform the structure of the model.

Searches for the answers to these questions followed an iterative approach, and the information specialist and the modeller discussed the results of the searches at each stage. Relevant text words and indexing terms were found by assessing a sample set of papers from the first set of searches and the search strategies were subsequently modified. If, in the view of the modeller, an excessive number of records was identified, the search strategy was adapted to be more focused and to retrieve fewer records. To focus the searches a combination was used of indexing terms restricted to major subject

headings, subheadings and text words restricted to the record title. This is necessarily a very different approach to the one used when searching as part of a systematic review. Search strategies for systematic reviews tend to be highly sensitive (retrieving many relevant papers) and often low in precision (also retrieving many additional but irrelevant papers). In focusing the search strategies for the parameters in the decision model there is the risk of missing relevant papers. However, this pragmatic approach was adopted because the number of questions and the time available did not allow extensive searches for each question.

Case-study background
Decision problem
A decision model was developed to link the number of recurrent UTIs that a child may experience over a period of 3 years to the number of pyelonephritic attacks, which in turn will determine the likelihood of developing progressive renal scarring.

Model structure
Progressive renal scarring is linked to the development of end-stage renal disease. Prophylactic antibiotics work through the impact of reducing the frequency of UTI. The model was analysed separately for children with and without vesicoureteral reflux (VUR), for boys and girls, and for children aged 1 or 3 years (i.e. eight patient subgroups in total).

The model was probabilistic. That is, parameter estimates were incorporated as mean values plus a measure of the precision of the mean. Costs were assessed from an NHS perspective, and benefits were expressed as QALYs. Simulation methods were used to determine the probability that alternative therapies were cost-effective at a range of threshold values that the NHS may attach to an additional QALY. Value of information analysis was used to quantify the cost of uncertainty associated with the decision about which therapy to adopt, and this indicates the maximum value of future research.

Sources of data
As noted above, the most appropriate sources of evidence are likely to vary according to the nature of the parameters of interest. There is a number of broad categories of parameters for which estimates are likely to be required in decision models. A suggested categorisation of the parameters with suggested sources of relevant research is presented below.

Baseline event rates
This category concerns data on health-related event rates associated with existing forms of treatment. In some instances, this may relate to event rates without active therapy (i.e. natural history data). These data may be needed for different subgroups of patients. Furthermore, for decision models informing NHS decision-makers, there may be a need to identify UK sources of evidence for these parameters. Potentially useful information sources were identified as:

- MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed/)
- EMBASE (http://www.embase.com)
- Incidence and Prevalence Database (IPD) (http://library.dialog.com)
- WHOSIS Evidence for Health Policy (http://www3.who.int/whosis/menu.cfm).

Health-related quality of life and its valuation
This includes data on the health states that patients experience following particular events. It also includes the values associated with those health states (utilities), which are based on the preferences of a relevant group (e.g. patients or the public). Potentially useful information sources were identified as:

- MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed/)
- EMBASE (http://www.embase.com)
- Harvard Center for Risk Analysis Cost-utility Analysis (CUA) Database/Cost-effectiveness Analysis (CEA) Registry (http://www.hsph.harvard.edu/cearegistry/)
- Quality of Life Instruments Database (QOLID at MAPI Research Institute) (http://www.qolid.org/)
- Health and Psychosocial Instruments (http://www.ovid.com)
- Health Related Quality of Life (http://www.cdc.gov/hrqol/)
- Australian Centre on Quality of Life (http://acqol.deakin.edu.au/)
- SCI (http://wos.mimas.ac.uk/).

Resource use and unit costs
This category includes data on the implications of treatments and health-related events for physical units of resource use (e.g. days in hospital, drug use), together with the monetary value of those resources (unit costs). Potentially useful information sources were identified as:

- BNF (http://www.bnf.org)
- Drug Tariff (Department of Health and National Assembly for Wales) (http://www.drugtariff.com)

- Monthly Index to Medical Specialities (MIMS) (Haymarket Publishing Services) or eMIMS (http://www.emims.net/login/main/login.htm)
- Department of Health Hospital Episode Statistics (HES) England. London: Department of Health (http://www.doh.gov.uk/hes/)
- McCormick A, Fleming D, Charlton J. Morbidity statistics from general practice: England and Wales. London: HMSO
- Organization for Economic Co-operation and Development Compendium of Health Data (www.oecd.org/health/healthdata)
- NHS Reference Costs (http://www.doh.gov.uk/nhsexec/refcosts.htm)
- World Health Organization (WHO)-CHOICE (http://www3.who.int/whosis/menu.cfm?path= evidence,cea&language=english)
- Unit Costs of Health and Social Care/Unit Costs of Community Care (Personal Social Services Resource Unit) (http://www.pssru.ac.uk/UC2003.htm)
- Health Service Financial Database (CIPFA)
- NHS Economic Evaluation Database (NHS EED) (http://www.york.ac.uk/inst/crd/nhsdhp.htm)
- Health Economic Evaluations Database (HEED) (http://www.ohe-heed.com).

Relative treatment effects
This category includes data on the clinical effectiveness of the intervention. Often decision models are structured in such a way that baseline event rates (i.e. events with usual treatment or natural history) are separated from the relative treatment effect of the new therapy compared with its comparator (e.g. an odds ratio or a relative risk). Possible data sources for baseline event rates are discussed in 'Baseline event rates', p. 33; here, the focus is on the relative treatment effect. Methods of retrieval of effects data are well developed and the key sources of information have been researched. As discussed above, estimates of relative treatment effects would ideally come from a systematic review or one or more RCTs where the risk of selection bias is minimal. Therefore, a hierarchy of information sources for evidence to populate models with relative effective data may be suggested as:

1. Cochrane Database of Systematic Reviews (CDSR) (http://www.update-software.com/cochrane/) (contains full-text systematic reviews)
2. Database of Abstracts of Reviews of Effectiveness (DARE) (http://nhscrd.york.ac.uk/darehp.htm) (contains abstracts of quality assessed systematic reviews)
3. HTA database (http://www.york.ac.uk/inst/crd/htahp.htm)

(contains unevaluated systematic reviews and other research)
4. CENTRAL (http://www.update-software.com/cochrane/) [contains the largest single collection of controlled clinical trials and RCTs]
5. MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed) or EMBASE (http://www.embase.com) (both contain bibliographic details of all study designs).

Other parameters
In a given decision model, a number of other parameters may need to be estimated. In the case study discussed here, resistance to drugs and side-effects, which may affect the cost, relative treatment effects and frequency of the treatment, may be important. Potentially useful information sources for these data are:

- MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed/)
- EMBASE (http://www.embase.com)
- TOXLINE (http://toxnet.nlm.nih.gov/).

Not all of the sources listed above were investigated in this case study owing to limited time and resources. The databases chosen for this study were EMBASE and MEDLINE, because of their wide availability to researchers, their size and the frequency with which their use is cited in TARs and other research. The IPD was also used for baseline event rates. This was chosen as it is a full-text database, which may prove easier to search for baseline events data as the title, abstract or indexing terms of relevant records often do not indicate that this information is in the full text. Therefore, the value of this database was investigated in terms of providing quick and efficient access to relevant data. For resource use and unit costs data, HEED and NHS EED were used, as these are convenient collections of records of economic evaluations and cost studies conducted in the past 10 years. For relative treatment effects, CDSR was searched, followed by DARE, the HTA database and the CENTRAL database. If necessary, MEDLINE and EMBASE would have been searched too. This order of database selection reflects the hierarchy of evidence and once enough evidence was found from the top of the hierarchy searching was stopped.

Searchable questions
After discussion between the information specialist and decision modeller, questions that were deemed searchable were categorised as follows.

Baseline event rates
In this case study, there were 32 questions categorised as relating to baseline events rates. These questions were about the incidence and prevalence of UTIs, rates of occurrence of UTIs and the relationship between different events associated with UTIs.

Incidence and prevalence
Of the 32 baseline events data questions, half were related to the incidence and prevalence of UTIs and formed questions 1–16. These questions were as follows.

In infant girls aged less than 1 year:

Question 1: What is the frequency of acute UTIs with recurrent UTI but no VUR?
Question 2: What is the frequency of acute UTIs with mild VUR?
Question 3: What is the frequency of acute UTIs with severe VUR?
Question 4: What is the incidence and prevalence of: recurrent UTI (no VUR), mild VUR, and severe VUR?

Questions 1–4 above were repeated for infant boys aged less than 1 year (questions 5–8), girls aged one to 1–10 years inclusive (questions 9–12) and boys aged 1–10 years inclusive (questions 13–16).

Rates of occurrence Twelve of the baseline event rates questions (questions 17–28) were related to the rates of occurrence of UTIs.

In infant girls aged less than 1 year:

Question 17: What proportion of acute UTIs are pyelonephritic attacks in those with recurrent UTI (no VUR)?
Question 18: What proportion of acute UTIs are pyelonephritic attacks in those with mild VUR?
Question 19: What proportion of acute UTIs are pyelonephritic attacks in those with severe VUR?

Questions 17–19 above were repeated for infant boys aged less than 1 year (questions 20–22), girls aged one–ten years inclusive (questions 23–25) and boys aged 1–10 years inclusive (questions 26–28).

Relationships between events Four of the baseline event rates questions (questions 29–32) were related to the relationship between events associated with UTIs.

In boys and girls of all ages:

Question 29: What is the relationship between the number of pyelonephritic attacks and the risk of progressive renal scarring (most importantly by VUR no VUR, as well as age and maybe gender)?
Question 30: What is the relationship between developing progressive renal scarring and developing end stage renal disease?
Question 31: Are there any other significant consequences of progressive renal scarring (severe hypertension)?
Question 32: What are the consequences (and their likelihood) of end-stage renal disease?

Health-related quality of life and its valuation
The next set of questions (questions 33–35) were asked in relation to health-related quality of life of infants and children with UTI and its potential consequences.

Question 33: What is the impact on quality of life of acute UTIs and pyelonephritic attacks in infants and children (duration of symptoms would be good too)?
Question 34: What is the reduction in quality adjusted life expectancy (or, failing that, in life expectancy) as a consequence of end-stage renal disease?
Question 35: What is the reduction in quality-adjusted life expectancy (or, failing that, in life expectancy) of any other consequences of progressive renal scarring?

Resource use and unit costs
Three questions (questions 36–38) related to the resource use and unit costs of UTIs. Two of these questions (questions 37 and 38) were dependent on the results of previous searches (questions 31 and 32) as the consequences of end-stage renal disease and progressive renal scarring needed to be ascertained.

Question 36: What are the costs of treating acute UTIs and pyelonephritic attacks in infants and children (duration, drugs/dose, hospitalisations, primary care visits)?
Question 37: What are the costs of the consequences of end-stage renal disease?
Question 38: What are the costs of any other consequences of progressive renal scarring?

Relative treatment effects
Two questions (questions 39 and 40) related to the effectiveness of antibiotics and surgery.

Question 39: What trials have been carried out in the use of long-term low-dose antibiotics for boys and girls of all ages for UTIs, and mild or severe VUR?

Question 40: How effective is surgery in treating mild or severe VUR for boys and girls of all ages?

Other parameters: antibiotics resistance
The last two questions (questions 41 and 42) were regarding antibiotic resistance.

Question 41: What studies have looked at resistance to these antibiotics in boys and girls?

Question 42: What studies have looked at resistance to these antibiotics in general?

Baseline event data questions were further grouped to avoid duplicate searching. For example, searching for the incidence and prevalence of recurrent UTI (no VUR), mild VUR and severe VUR was difficult to separate out from the search for frequency of acute UTIs with recurrent UTI but no VUR. Limiting searches to boys and girls was not seen as practical because, in this case, most of the searches did not produce an unmanageable number of records, and limiting to boys and then girls did not limit the number of records retrieved significantly, but did increase the risk of excluding relevant papers. Many papers referred to UTIs in 'boys and girls' and in others the terms for 'boys' and 'girls' did not appear, but 'children' or other similar words did. The full search strategies are presented in Appendix 8(a).

Results

The number of records and the databases from which they were retrieved for each question, or group of questions, is presented below. The relevance of search results is discussed in terms of which records gave information that could assist in the estimation of parameters relevant to baseline event rates, health-related quality of life, resource use and unit costs, relative treatment effects and other information such as antibiotic resistance. The first set of searches for quality of life information, which focused on general quality of

life measures, did not produce relevant results relating to quality of life, and a second set of search strategies was devised which concentrated on identifying utility values rather than quality of life measures per se.

Number of records
Baseline event rates
The 32 questions relating to baseline events data were searched on MEDLINE, EMBASE and IPD, using 11 separate search strategies adapted for each database. To restrict the results to a manageable number of records the following approach was adopted to the search strategy in both MEDLINE and EMBASE: (1) the use of major indexing terms was implemented, (2) using subheadings where available, and (3) text word searching in the title and/or abstract. There was considerable overlap between the results of these searches across the databases and across the search results for the different questions. The results, therefore, were deduplicated between databases and across questions before being passed to the modeller. Only a selection of the results from the IPD database was downloaded, owing to the cost of downloading from this database. The overlap between the results from EMBASE and MEDLINE was quantified for each search. The number of duplicate records from the searches on IPD could not be ascertained, as not all of the records were downloaded from this database. However, all of the relevant records identified from IPD were unique and not found by either the MEDLINE or EMBASE searches.

Incidence and prevalence Questions 1–16 on incidence and prevalence were searched using four search strategies. Questions 1 and 5 and the first part of questions 4 and 8 were searched together, with similar search strategies in MEDLINE, EMBASE and IPD. These questions were asking for the frequency or incidence and prevalence of recurrent UTIs in infant boys or girls. The majority of the records were found on MEDLINE (108) (Table 4); 23 records were found using both EMBASE and MEDLINE, and after deduplication between EMBASE and MEDLINE 199 records were retrieved.

TABLE 4 *Results of searches for incidence and prevalence (Questions 1, 4, 5 and 8)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	108	85
EMBASE	1980 to week 6 2003	35	12
IPD	1994 to 1 February, 2003	79	Unknown

TABLE 5 *Results of searches for incidence and prevalence (questions 2–4, 6–8)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	28	17
EMBASE	1980 to week 6 2003	18	7
IPD	1994 to 1 February 2003	4	Unknown

TABLE 6 *Results of searches for incidence and prevalence (questions 9, 12, 13, and 16)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	201	128
EMBASE	1980 to week 6 2003	112	39
IPD	1994 to 1 February 2003	2	Unknown

TABLE 7 *Results of searches for incidence and prevalence (questions 10–12 and 14–16)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	30	6
EMBASE	1980 to week 6 2003	38	14
IPD	1994 to 1 February 2003	5	Unknown

TABLE 8 *Results of searches for rates of occurrence (questions 17 and 20)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	25	19
EMBASE	1980 to week 6 2003	8	2
IPD	1994 to 1 February 2003	59	unknown

Questions 2, 3, 6 and 7 and the second part of questions 4 and 8 were searched together, with similar search strategies in MEDLINE, EMBASE and IPD. These questions were asking for the frequency or incidence and prevalence of mild or severe VUR in infant boys or girls. The majority of the records were again retrieved from MEDLINE (28) and 11 records were found in both EMBASE and MEDLINE (*Table 5*). After deduplication between EMBASE and MEDLINE 39 records were retrieved.

Questions 9 and 13 and the first part of questions 12 and 16 were searched together, with similar search strategies in MEDLINE, EMBASE and IPD. These questions were asking for the frequency or incidence and prevalence of recurrent UTIs in boys or girls aged 1–10 years inclusive. The majority of the records were again retrieved from MEDLINE (201) and 73 records were found with both EMBASE and MEDLINE (*Table 6*). After deduplication between EMBASE and MEDLINE 242 records were retrieved.

Questions 10, 11, 14 and 15 and the second part of questions 12 and 16 were searched together, with similar search strategies in MEDLINE, EMBASE and IPD. These questions were asking for the frequency or incidence and prevalence of mild or severe VUR in boys or girls aged 1–10 years inclusive. For this set of questions EMBASE retrieved the largest number of records (38) and 24 records were found with both EMBASE and MEDLINE (*Table 7*). After deduplication 49 records were retrieved.

Rates of occurrence Questions 17–28 relating to rates of occurrence were grouped so that they could be captured using three search strategies. Questions 17 and 20 were searched together, with similar search strategies in MEDLINE, EMBASE and IPD. These questions were asking for the proportion of UTIs which are pyelonephritic attacks in those with recurrent UTI in infant boys and girls. The majority of the records were retrieved from the IPD (*Table 8*). After deduplication between EMBASE and MEDLINE

TABLE 9 *Results of searches for rates of occurrence (questions 18, 19, 21, 22, 24, 25, 27 and 28)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	13	4
EMBASE	1980 to week 6 2003	13	4
IPD	1994 to 1 February 2003	4	Unknown

TABLE 10 *Results of searches for rates of occurrence (questions 23 and 26)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	40	27
EMBASE	1980 to week 6 2003	19	6
IPD	1994 to 1 February 2003	59	Unknown

TABLE 11 *Results of searches for relationships between events (question 29)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	167	68
EMBASE	1980 to week 6 2003	146	47
IPD	1994 to 1 February 2003	10	Unknown

86 records were retrieved. Six records were found using both EMBASE and MEDLINE.

Questions 18, 19, 21, 22, 24, 25, 27 and 28 were searched together, with similar search strategies in MEDLINE, EMBASE and IPD. These questions were asking for the proportion of UTIs that are pyelonephritic attacks in those with mild or severe VUR in infant and child boys and girls. The searches did not limit by age or gender as a general search retrieved a very manageable number of results. The searches using MEDLINE and EMBASE retrieved an equal number of records (13) with a large degree of overlap (nine records were identified from both databases (*Table 9*). After deduplication between EMBASE and MEDLINE 21 records were retrieved.

Questions 23 and 26 were searched together, with similar search strategies in MEDLINE, EMBASE and IPD. These questions were asking for the proportion of UTIs that are pyelonephritic attacks in those with recurrent UTI in child boys and girls. The searches did not limit by age or gender as a general search retrieved a very manageable number of results. IPD produced the largest number of records (59) and 13 records were found using both EMBASE and MEDLINE

(*Table 10*). After deduplication between EMBASE and MEDLINE 105 records were retrieved in total.

Relationships between events The four questions relating to relationships between events were searched individually using four search strategies. Question 29, which asked about the relationship between the number of pyelonephritic attacks and the risk of progressive renal scarring, was searched using similar search strategies in MEDLINE, EMBASE and IPD. Using both EMBASE and MEDLINE 99 records were identified (*Table 11*) and after deduplication between EMBASE and MEDLINE 224 records were retrieved in total.

Question 30, which asked about the relationship between developing progressive renal scarring and developing end-stage renal disease, was searched with similar search strategies in MEDLINE, EMBASE and IPD. After deduplication between EMBASE and MEDLINE 111 records were retrieved. Using both EMBASE and MEDLINE 44 records were identified (*Table 12*).

Question 31 looked at whether there were any other significant consequences of progressive renal

TABLE 12 *Results of searches for relationships between events (question 30)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	64	20
EMBASE	1980 to week 6 2003	74	30
IPD	1994 to 1 February 2003	17	Unknown

TABLE 13 *Results of searches for relationships between events (question 31)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	10	8
EMBASE	1980 to week 6 2003	6	4
IPD	1994 to 1 February 2003	16	Unknown

TABLE 14 *Results of searches for relationships between events (question 32)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to January week 2 2003	41	24
EMBASE	1980 to week 6 2003	26	9
IPD	1994 to 1 February 2003	692	Unknown

scarring and was searched with similar search strategies in MEDLINE, EMBASE and IPD. Although IPD may not be the most appropriate database for this question it was included as the search strategy could easily be focused to provide a manageable set of results.

After deduplication between EMBASE and MEDLINE 30 records were retrieved and only two records were found using both EMBASE and MEDLINE (*Table 13*).

Question 32 asked for the consequences (and their likelihood) of end stage renal disease and was searched using similar search strategies in MEDLINE, EMBASE and IPD. However, the search could not be easily focused in IPD to produce a manageable set of records. It was thought that this was because the question is not strongly associated with incidence and prevalence so IPD might not be the most appropriate database for this question.

After deduplication 50 records were retrieved in MEDLINE and EMBASE and an extra 692 in IPD. Seventeen records were found in both EMBASE and MEDLINE (*Table 14*).

The 11 sets of search results for the baseline events rates questions retrieved a total of 2169 records (947 from IPD, 727 from MEDLINE and 495 from EMBASE). However, there was considerable overlap between the results of the searches in MEDLINE and EMBASE, and after deduplication within the results for each question 901 unique records were identified. A further deduplication of the records was carried out as there was also overlap in the results across each question. This further reduced the number of records from MEDLINE and EMBASE to 583 unique records. The information officer sifted the results from IPD because searching the full text retrieved large quantities of clearly irrelevant records and there is a cost for downloading each record from this database. Of the 25 records selected by the information officer from IPD, only one paper had already been found by the baseline events searches on MEDLINE and EMBASE.

Health-related quality of life and its valuation
Three questions were asked in relation to quality of life issues. The first of these questions (question 33) proved the most difficult to search for. Three MEDLINE search filters available in the literature were tested for their usefulness in relation to this question. In total these MEDLINE searches produced 143 records (*Table 15*). Following an assessment of the search results by the decision modeller it was decided that expanding these searches to other databases such as EMBASE would not be worthwhile.

TABLE 15 *Results of the first series of searches for quality of life issues (question 33)*

Database	Date range	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to February week 2 2003	Comprehensive outcome filter[34] plus duration terms	143
MEDLINE	1966 to January week 2 2003	Brief outcome filter[34]	28
MEDLINE	1966 to January week 2 2003	ScHARR outcome filter[35]	0
ScHARR, School of Health and Related Research, Sheffield, UK.			

TABLE 16 *Results of the second set of searches for quality of life issues (question 33)*

Database	Date searched	Filter used	No. of unique records retrieved	No. of unique records retrieved
MEDLINE	1966 to April week 3 2003	Utilities filter devised at CRD	95	51
EMBASE	1980 to week 17 2003	Utilities filter devised at CRD	122	78
CRD, Centre for Reviews and Dissemination, University of York.				

TABLE 17 *Results of quality of life searches (questions 34)*

Database	Date searched	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to April week 3 2003	9	7
EMBASE	1980 to week 17 2003	5	3

TABLE 18 *Results of quality of life searches (question 35)*

Database	Date searched	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to April week 3 2003	4	3
EMBASE	1980 to week 17 2003	3	2

A second set of strategies was therefore developed to try to produce more useful results. These were developed using terms relating to the utilities associated with the health states in the model. The second set of quality of life searches produced 173 records after deduplication. 44 of the records were found in both EMBASE and MEDLINE (*Table 16*).

Similar search strategies were used for question 34 regarding the quality-adjusted life expectancy or life expectancy as a consequence of end-stage renal disease, and question 35, which looked at the reduction in quality-adjusted life expectancy or life expectancy of any other consequences of progressive renal scarring. The searches for question 34 retrieved 12 records, of which two were found in both EMBASE and MEDLINE (*Table 17*).

The searches for question 35 retrieved six records, of which one was found in both EMBASE and MEDLINE (*Table 18*).

The search strategies for questions 34 and 35 would have been improved once the consequences of end-stage renal disease and renal scarring were known. However, these consequences were dependent on the results of previous questions (question 31 and 32) and time constraints did not allow feedback from the searches to answer these questions.

Resource use and unit costs
There were three questions involving resource use and unit costs. The first of these questions, in relation to the costs of treating acute UTIs and pyelonephritic attacks in infants and children

TABLE 19 Results for the searches for resource use (question 36)

Database	Date searched	No. of records retrieved	No. of unique records retrieved
NHS EED admin database	17 February 2003	57	35
HEED	26 February 2003	64	42

TABLE 20 Results of searches for relative treatment effects (question 39)

Database	Date searched	No. of records retrieved	No. of unique records retrieved
CDSR	Issue 1 2003	14	14
DARE	31 January 2003	30	16
HTA	31 January 2003	4	4

TABLE 21 Results for effectiveness search (question 40)

Database	Date searched	No. of records retrieved	No. of unique records retrieved
CDSR	Issue 1 2003	6	6
DARE	31 January 2003	0	0
HTA	31 January 2003	0	0
CENTRAL	Issue 1 2003	72	72

(duration, drugs/doses, hospitalisations, primary care visits) (question 36), was searched using paper-based sources such as BNF, Morbidity Statistics from General Practice and HES, as well as two economic evaluation databases, NHS EED and HEED. The paper sources were searched via their indexes and browsing. Similar search strategies were used as far as possible in HEED and NHS EED.

Photocopies of the relevant pages were made from each paper source and passed to the modeller, and 121 references were retrieved from the electronic databases, which after within-question deduplication reduced to 99 references (*Table 19*).

Question 37, regarding the costs of the consequences of end-stage renal disease, and question 38, on the costs of any other consequences of progressive renal scarring, were dependent on the results of previous questions (questions 31 and 32) and time constraints did not allow feedback from the searches to answer these questions. However, it is anticipated that similar resources and search strategies to question 36 would be used to identify relevant cost data once the consequences are known.

Relative treatment effects
There were two questions about relative treatment effects. Information to answer the first question

looking for trials carried out on the use of long-term low-dose antibiotics for boys and girls of all ages for UTIs, and mild or severe VUR (question 39) was sought in CDSR, DARE and the HTA database (*Table 20*). It was decided that, in accordance with the hierarchy of evidence, this search need not be expanded to find RCTs, as the searches for systematic reviews had already produced a number of relevant records.

Answers to the second effectiveness question on how effective surgery is in treating mild or severe VUR for boys and girls of all ages (question 40) were sought in CDSR, DARE, the HTA database and CENTRAL. The searches for systematic reviews produced six records (*Table 21*), but few of these appeared directly relevant. The search was therefore expanded to look for RCTs on CENTRAL. This produced 72 records (*Table 21*) and searching was stopped as the decision modeller thought this to be a manageable set of records with some relevant records. If the modeller had indicated that this set contained no relevant records the search would have been expanded to other databases such as MEDLINE and EMBASE.

The searches for information on relative treatment effects resulted in a total of 126 records and there was no duplication with the other searches carried out in this study.

TABLE 22 *Results of searches for antibiotic resistance (questions 41 and 42)*

Database	Date searched	No. of records retrieved	No. of unique records retrieved
MEDLINE	1966 to February week 2 2003	78 (5 specifically mentioning children)	36
EMBASE	1980 to week 6 2003	206 (9 specifically mentioning children)	164

Other parameters: antibiotic resistance
Lastly, searches were carried out for information on resistance to specified antibiotics to answer the following questions: what studies have looked at resistance to these antibiotics in boys and girls (question 41)? and what studies have looked at resistance to these antibiotics in general (question 42)?

The total number of records retrieved relating to the antibiotic resistance searches was 284; after within-question deduplication this reduced to 242 (*Table 22*).

Relevance of search results
Baseline event rates
The searches of MEDLINE and EMBASE produced 727 and 495 references, respectively. After deduplication (between databases and across questions), this reduced to 583 records. Abstracts and titles were screened for relevance and 56 papers were ordered. Of these records, 44 were identified from MEDLINE and 31 from EMBASE (with 19 records identified from both databases). Of the 56 papers ordered, five papers were used in developing and populating the model, although other papers did provide useful background material.

Of the five papers that were directly relevant to the model, all were found in EMBASE and none in MEDLINE. There were several reasons why these papers were not found on MEDLINE. First, two of the five records were from journals that are not indexed on MEDLINE and one record was not on MEDLINE at the time of searching (MEDLINE is not as current as EMBASE). The other two were not retrievable with the search strategy as one contained no abstract on MEDLINE (although an abstract was present in EMBASE) and the other paper did not contain any of the MEDLINE subject headings used in the search.

Although the MEDLINE search retrieved more papers than the search in EMBASE, and more papers were ordered from the MEDLINE results, the EMBASE search proved much more successful in finding papers of confirmed direct relevance to the model.

In relation to the baseline events data, the search on IPD produced 947 records. Of these, six (0.63%) were directly relevant. However, with hindsight the questions regarding the relationships between events were not very appropriate for this database. In particular, the search for the consequences of end-stage renal disease produced too many irrelevant records despite a relatively focused search strategy. Unsurprisingly, this database proved much more useful for incidence and prevalence and rates of occurrence data. The number of records retrieved from the IPD in relation to the incidence/prevalence/rates of occurrence data was 212. Of these, six (2.83%) were directly relevant. In particular, data were obtained for the model on the percentage split of UTIs by age group, the number of cases of end-stage renal disease following VUR and the number of VUR cases that resolve each year.

In relation to all of the baseline events data, the search on IPD provided the largest number of relevant records (6/947) and the search on EMBASE gave the best results in terms of having fewer irrelevant records that needed to be read to identify useful data (5/495). As the search on MEDLINE produced no directly relevant results this search performed least well of the three searches (*Table 23*). If, however, the results of the incidence and prevalence and rates of occurrence searches are viewed separately from the relationships of events data then IPD produced not only the largest set of relevant records (6), but also the highest precision (2.83%) for the incidence/prevalence and rates of occurrence parameters.

The search strategies on IPD were not repeated owing to the high online charges of this database. However, there were a number of problems with the data obtained from the records retrieved from IPD, which may have been preventable by manipulating the search strategy:

- Most of the prevalence/incidence data related to the USA, with international prevalence from Sweden (this is only in men aged over 45 years).

TABLE 23 *Performance of the searches for baseline events data (questions 1–32)*

	No. of records retrieved	No. of relevant records	Precision (no. of relevant records/ no. of records retrieved)
IPD	947	6	0.63%
EMBASE	495	5	1.01%
MEDLINE	727	0	0%

TABLE 24 *Performance of the searches for health-related quality of life (question 33)*

	No. of records retrieved	No. of relevant records	Precision (no. of relevant records/ no. of records retrieved)
EMBASE	122	1	0.82%
MEDLINE	95	2	2.11%

- The data refer to the numbers of cases of UTI, not recurrent UTIs (three or more UTIs in past 5 years) that were required for the model.
- A lot of the data were about cystitis, which was not relevant for this model.

There were also some other areas where the search strategies in EMBASE and MEDLINE could have been improved given more time. For example, excluding non-English language papers and studies relating solely to adults would have improved the precision of the search results (i.e. reduced the number of irrelevant records) and the exclusion of animal studies from the EMBASE searches, in particular, would also have improved the precision of the results.

Health-related quality of life and its valuation
The first set of searches using published search filters produced 143 records, of which 13 were deemed potentially relevant and their papers were ordered. However, not all of these were related to the quality of life questions specified by the decision modeller, but were considered potentially useful for other questions in the model. Of the 13 papers that were ordered, four were directly relevant to the model, but none of these was directly useful in terms of contributing quality of life information. The records were useful, however, for the baseline event rates, relative treatment effects, and resource use and unit cost parameters.

In summary, the first quality of life searches based on the published search filters did not identify any studies that were useful to inform quality of life model parameters or distributions. Most of the utility values that were attached to

UTI/pyelonephritic attacks were actually identified from the searches that focused on resource use.

Given the lack of relevant studies produced in the first round of quality of life searches, an alternative search strategy was developed, with a focus on identifying quality of life data relevant for economic models; that is, utilities associated with the health states defined in the model (UTI, pyelonephritic attack, dialysis and renal transplant). Additional terms relating to life expectancy and the duration of symptoms were also used. The revised search strategy can be seen in Appendix 8(a). Searching MEDLINE with the revised search strategy generated 99 records (51 records relating to quality of life, 13 to life expectancy and 35 to duration of symptoms). In EMBASE, 124 records were retrieved (85 records relating to quality of life, 14 to life expectancy and 25 to duration of symptoms). From the 129 records (after deduplication), three were potentially relevant to the model and were ordered. In reviewing the full-text of these papers, two were found to be directly relevant. Both of these papers were found in MEDLINE and one of them was found in EMBASE (*Table 24*).

The searches for information relating to life expectancy and quality-adjusted life expectancy proved more successful. Five of the 18 records retrieved were directly relevant to the model. Both of the searches on EMBASE and MEDLINE retrieved relevant records, although the precision of the searches was slightly higher on EMBASE. Three relevant papers were found on MEDLINE and three relevant papers were found on EMBASE. Only one relevant paper was found on both databases (*Table 25*).

TABLE 25 *Performance of the searches for life expectancy information (questions 34 and 35)*

	No. of records retrieved	No. of relevant records	Precision (no. of relevant records/ no. of records retrieved)
EMBASE	8	3	37.5%
MEDLINE	13	3	23.08%

TABLE 26 *Performance of the searches for resource use and unit costs (question 36)*

	No. of records retrieved	No. of relevant records	Precision (no. of relevant records/ no. of records retrieved)
HEED	64	2	3.12%
NHS EED	57	2	3.51%

TABLE 27 *Performance of the searches for relative treatment effects (questions 39 and 40)*

	No. of records retrieved	No. of relevant records	Precision (no. of relevant records/ no. of records retrieved)
CDSR	20	2	10%
DARE	30	7	23.3%
HTA	4	0	0%
CENTRAL	72	10	13.9%

Resource use and unit costs

The paper sources did not contribute to the resource use and costs questions. However, the database searching proved more successful. The resource use and costs searches produced 99 records, of which 20 were found to be potentially relevant to the model and the full papers were ordered. Sixteen of the 20 papers were found on HEED and six on NHS EED (four were found on both databases).

Once the full papers had been obtained, three of the 20 papers ordered were found to be directly relevant to the model (*Table 26*). Two of these papers were identified from NHS EED (one unique) and two on HEED (one unique). Although the searches of HEED produced more records and more potentially relevant records, both database searches contributed equally to this model.

To increase the precision of the searches, it is important to specify the UK setting in the search, as many studies identified were conducted in the USA and hence the resource use and costs may not be directly transferable. A staged search could have been adopted in which if no studies were identified for the UK (or even selected European countries) then US studies could be identified. However, as a relatively small set of records was retrieved in both of these databases this was not considered worthwhile in this instance.

Relative treatment effects

In total, 126 records were produced from the relative treatment effects searches, of which 17 were considered relevant. However, owing to the time constraints on the modelling exercise, these references were not acquired in full-text form. The relative treatment effects in the model were actually taken from a Cochrane Review.[36] Three of the four databases searched (CDSR, DARE and CENTRAL) provided potentially relevant results (*Table 27*). There were no specific problems with this search or the results that it produced. This was anticipated as the search approaches for effectiveness data are well established, and there are key databases, which contain effects information, that are relatively easy to search in a focused way.

In addition, several systematic reviews of effectiveness were identified in this search, which the modeller indicated would not have been ordered for the model but may have been obtained to provide background material.

Other parameters: antibiotic resistance

The searches for papers on antibiotic resistance

TABLE 28 *Performance of the searches for other parameters: antibiotic resistance (questions 41 and 42)*

	No. of records retrieved	No. of relevant records	Precision (no. of relevant records/ no. of records retrieved)
MEDLINE	78	2	2.56%
EMBASE	206	2	0.97%

identified 242 records. More than double the number of references were retrieved from EMBASE than from MEDLINE. This is because the search strategy in EMBASE could not be as easily focused as the one undertaken in MEDLINE because of the practice of extensive indexing of drugs in EMBASE. Of the 242 records, 21 were potentially relevant to the model and thus ordered. Of these 21 papers, ten were identified in MEDLINE and 16 in EMBASE, with five identified in both databases.

Four of the 21 papers ordered contained data directly relevant to the model. Of these, two were found in MEDLINE (and are indexed on EMBASE but were not found by the search strategy) and two were found by the EMBASE search strategy (*Table 28*) (and only one of these papers was indexed on MEDLINE but was not found by the MEDLINE search). This emphasises how focused the search strategies were, despite obtaining such a large set of results. The precision of the searches for this question was low, at 1.65% (4/242 × 100).

Discussion
This case study has shown that to provide efficient searches to inform the model it is important to establish a process of continual dialogue throughout the project between the modeller and the information specialist. However, this iterative process takes time and does not happen frequently in current modelling practice. In this case study, feedback from the modeller, although time consuming, proved invaluable in producing more focused searches with relevant records. It is likely that even closer collaboration would prove yet more effective in producing focused searches that provide relevant information to inform the model parameters.

Decisions are required throughout a modelling project about data requirements and the ability of the search questions to identify these data. In this example, focusing or broadening the search strategies was discussed for each parameter after an assessment of the first round of searches. The dialogue between modeller and information specialist was particularly useful for the searches of

the databases with no online charges, because searches on these databases can be repeated at no extra charge. For searches of databases with online charges, such as IPD, an iterative process of continually improving the searches after comments from the modeller would have proved too costly. Thus, the search strategies for MEDLINE and EMBASE were adapted for IPD and this database was only searched once. Another less costly solution when searching on databases with online charges may be for the information specialist to conduct the searches alongside the modeller. This would enable the modeller to provide initial feedback on the results and for the information specialist then to adapt the search strategies while online. In addition, the modeller would be able to select which records are of potential interest and this would reduce the number of records downloaded and thus the cost of downloading. In this case study the information specialist selected the records for downloading and was overcautious to avoid missing any relevant records.

The searches helped to identify important data for inclusion in the model. The initial population of the model was undertaken using searches conducted by the modeller. The modeller's methods included the identification of studies by searching the Cochrane Library and by contacting clinical experts in the topic area. This approach did not produce as many relevant studies as the searches for each parameter conducted by the information specialist in this case study. Further case studies could be undertaken in which the procedures typically used by modellers are compared with searching for information for each parameter in the model by an information specialist in collaboration with modellers. These case studies could explore whether the records obtained from the two approaches produce different information in terms of both quantity and quality, and how these differences affect the end model, on which ultimately healthcare decisions are made.

Focused searches that aimed to retrieve a manageable number of records, in the view of the modeller, were carried out on both MEDLINE and

EMBASE for questions relating to the baseline event rates, health-related quality of life and antibiotic resistance. Search techniques included using major indexing terms and text words in the title. Overall, these focused searches provided relevant data from both MEDLINE and EMBASE for the model used in this case study. The numbers of relevant records in this case study were too small, however, to enable any general recommendations to be made as to which database is more useful. However, there were certain areas in which the results suggested that one database may prove to be more useful than the other. For example, the searches for baseline events data proved more successful in EMBASE (which provided five directly relevant records) than in MEDLINE (which did not provide any directly relevant records). In contrast, in relation to antibiotic resistance, the drug indexing in MEDLINE allowed more focused searching and so the searches in MEDLINE provided marginally better precision (2.56%) than EMBASE (0.97%).

Some of the relevant papers found on EMBASE were not indexed on MEDLINE and the reverse was also true. This suggests that, although the searches missed relevant records because of their focused nature, they also would have missed relevant papers had only MEDLINE or EMBASE been searched. For instance, of the five relevant papers found for the baseline event rates parameters, three were not indexed on MEDLINE. Of the seven relevant papers found for the health-related quality of life and its valuation, two were unique to MEDLINE, one was unique to EMBASE and four appeared in both databases. This indicates that searching both MEDLINE and EMBASE is required as both produced useful unique relevant papers. Further case studies may provide stronger evidence about the best combinations of databases to search.

The titles and abstracts of many records do not indicate whether the data required are contained within the full text of the document. Thus, general medical databases such as MEDLINE and EMBASE may not necessarily be the best single sources of data for all questions. The searches of the IPD (which contains full-text records) proved useful and provided the greatest number of relevant records for the incidence and prevalence parameters, with few irrelevant records for those questions. Although IPD has some overlap with MEDLINE and EMBASE all the relevant records retrieved from IPD in this case study were unique. However, in light of the cost of this database, searches in IPD may need to be undertaken as a

last resort. For example, IPD could be used to search for very focused questions for which no data were found in the databases that are free at the point of use. Another alternative would be to budget for a search of IPD in the research proposal, and then this database could be used to generate relatively quick incidence and prevalence data for the model.

Unsurprisingly, the searches of both NHS EED and HEED provided good results for the resource use and cost questions. In this case study, HEED produced slightly more records (64, compared with 57 on NHS EED), but both databases contributed an equal number of relevant references for the model. Interesting relevant papers were retrieved from each database that were not indexed in the other database. This suggests that both databases should be searched. However, more research is needed to compare the usefulness of focused searches of both databases compared with broad searching on just one of these databases. To make a useful comparison of the contribution that NHS EED and HEED can make to the decision model, more research needs to be undertaken, not just in terms of the number of relevant papers found on each database, but also in terms of how much the papers contribute to populating decision models. There is also a considerable time lag with both these databases and an exploration into searching more current resources may prove useful.

The quality of life searches proved to be the most problematic of the areas considered. There are three possible reasons for this. First, published search strategies for quality of life data (which were tested out) do not focus on the measures that are required for cost-effectiveness models, in this instance, utility values. Second, very few studies have estimated utility values specifically for UTIs and related events. Third, the optimal search terms necessary to capture records that describe these data have not yet been researched and tested. This is another area that requires more exploration.

The searches for almost all questions produced a lot of duplicate records not only between the databases but also with other questions. For example, both the first and second iterations of the quality of life searches demonstrated considerable overlap with the resource use and antibiotic resistance searches.

Excluding the IPD results, the other searches found a total of 2162 records, which after between-database (within each question)

deduplication produced 1684 records, and after across-question deduplication produced 1213 records. This indicates the level of duplication of records both within each question (database duplication) and across each question (question duplication). There were 478 records that were duplicated within questions. This is to be expected, as there is considerable overlap between such databases as MEDLINE and EMBASE, and HEED and NHS EED. However, despite carefully grouping the questions by topic and devising specific search strategies for each group of questions, there was also considerable overlap across questions (471 records were found for more than one question). This case study, therefore, indicates that a topic for future studies could be whether more of the search strategies could be amalgamated, leading to less work for both the information specialist and the modeller.

This case study generated small numbers of papers from each database, so it is difficult to generalise the findings to other decision models in other topics. However, a preliminary investigation was carried out of the feasibility of using these search strategies to populate some parameters in a model on Crohn's disease. Information on two aspects of the condition were requested: health-related quality of life and baseline event rates. The search strategies for quality of life/baseline event rates conducted in this case study were tested out on the topic of Crohn's disease. However, the strategies retrieved very high numbers of records when applied to the large body of literature on Crohn's disease. The searches, therefore, had to be narrowed down considerably to retrieve a more manageable number of references for the modeller. More research needs to be undertaken with other decision models to ascertain how far the approaches and strategies devised for the case study can be applied easily and effectively to search for information to populate common topic areas of models.

From this case study it can be seen that focused searching in single databases can miss relevant records. One way to reduce this possibility is to expand the number of resources searched for each question. An acceptable level of focus in search strategies and the number of resources searched will ultimately be determined by the resources available to the project. It is important to record the decisions made about the levels of focus in the search strategies and the resources used, to enable readers to appreciate the inevitable compromises that may have been made in identifying evidence to inform the decision model.

Effects of selection bias on treatment outcomes

Background
The second area where, at the outset of the study, it was felt that more methods guidance for decision modelling was required, was how to handle bias in parameter estimates incorporated into decision models.

Under the general heading of uncertainty, it is important to distinguish between bias and precision. In recent years, literature has emerged on methods: (1) to quantify the imprecision with which input parameters in models are estimated; (2) to assess the implications of that imprecision for the uncertainty associated with decisions about the use of health technologies; and (3) to feed this into decisions about future research priorities.[1,2,37] However, uncertainty is also present in models because of the quality of the evidence used to derive input parameters. This applies, in particular, to the potential for selection bias to affect estimates of the relative treatment effect taken from non-randomised studies.

The need to develop methods with which selection bias can be reflected in the uncertainty built into decision-analytic models was identified as a gap in the guidelines of best modelling practice.

Objective
The original protocol set out to undertake a scoping exercise of the literature describing the effect of selection bias in terms of the uncertainty associated with an estimate of treatment effect. The authors were only interested in papers where this uncertainty had been quantified; for example, by inflating the standard error around the treatment effect estimate. The scoping review was also to have included papers on suggested methods of reflecting selection bias in decision models.

Methods
To inform this, a literature scoping exercise was undertaken to assess the availability of literature which has: (1) quantified the effect of selection bias in terms of the uncertainty associated with a measure of treatment effect (e.g. by increasing the standard error around the measure); and (2) suggested methods of reflecting selection bias in decision models. It is expected that this scoping exercise will provide signposts for future methods research, including full systematic reviews.

Searching the literature
Searching electronic databases for methodological

papers that explore quantification of bias is not straightforward and has been little researched. The indexing of research methods literature in databases such as MEDLINE is poor because the focus of the database tends to be on indexing clinical content. Retrieval is hampered by the fact that words that suggest that a paper may cover methods issues frequently occur in the title or abstract of clinical research papers even when methodological issues are not discussed. Consequently, methodological terms such as outcome, randomised, generalised and search strategy produce lots of irrelevant records. Many of the keywords relevant to the topic of bias are frequently cited in the abstract or indexing without bias being the main subject of the paper, and many papers mention the need to quantify or take account of bias without taking this any further. In addition, the lexicon of terms describing issues of relevance to a quantification of selection bias is not well established and it is, therefore, very difficult to derive an adequately sensitive search strategy that does not also produce large numbers of irrelevant records.

The size of the literature relating to bias is very large. Searching on MEDLINE alone using the term bias produced over 32,000 records, most of which will not be papers that quantify an exaggerated or underestimated treatment effect. A scoping search (see Appendix 8a) limited to specified research designs and using major MeSH, limiting 'bias' to the title and using proximity operators to link text terms closely, still retrieved 9447 papers in MEDLINE. This was considered an impractical number for sifting in the time available, and yet it was recognised that this search would still fail to retrieve many relevant papers that were not indexed correctly or that did not use the specified terminology.

Search strategy
A focused, and inevitably more pragmatic, search strategy was devised using major MeSH, MeSH subheadings and terms limited to the title and linked by proximity operators (see Appendix 8b). The search terms were chosen to achieve a balance between sensitivity (high retrieval rates of relevant records) and precision (high rates of relevance among retrieved records). The literature searches were restricted to records published in 1990 onwards owing to constraints of time and resources. Although financial constraints did not allow for translations, no language restrictions were applied.

Although this strategy did not aim to uncover all papers in this field, it was anticipated that it would retrieve some significant relevant papers. The following databases were searched (12–13 May 2003):

- DARE (CRD administration database version using the CAIRS interface)
- EMBASE (1980 to week 19 2003) (Ovid)
- HTA Database (http://nhscrd.york.ac.uk/welcome.htm)
- MEDLINE (1966 to April week 4 2003) (Ovid)
- NHS EED (CRD administration database CAIRS using software)
- Cochrane Database of Methodology Reviews (Issue 2, 2003) (web version http://www.update-software.com/clibng/cliblogon.htm)
- CMR (Issue 2, 2003) (web version http://www.update-software.com/clibng/cliblogon.htm).

Results
Records identified
These searches retrieved 768 references. Two reviewers independently assessed the records retrieved for relevance. After assessing half of the references, it was found that agreement was reached in 88% of cases. There was a high level of rejection and disagreement around relevance. For references that were judged to be relevant by at least one reviewer (27% of all references), agreement was reached in only 19% of cases. Most importantly, neither reviewer was confident that rejected titles did not contain relevant information. It was concluded that, ideally, it would be necessary to order and assess the full papers of all 768 references. This was not considered practicable within the available time and funding.

It was concluded that this question could not be addressed using the range of search strategies of decreasing sensitivity that had been devised. Therefore, the approach was adopted of contacting experts in the field and checking citations. Fourteen relevant references were identified (see Appendix 8c for the data extraction tables). Few of these references had been included amongst the 768 identified from the electronic databases, confirming the assessment that focused strategies would miss relevant records.

Summary of study findings
Three studies did not provide actual estimates of bias.[38–40] Of the remaining 11 studies, the following five concluded that a non-randomised trial design is associated with bias.

Ioannidis and colleagues[41] compared results of randomised and non-randomised studies that

evaluated medical interventions and examined characteristics that may explain discrepancies between randomised and non-randomised studies. They identified 45 diverse topics for which both randomised trials ($n = 240$) and non-randomised studies ($n = 168$) had been performed. Very good correlation was observed between the summary odds ratios of randomised and non-randomised studies; however, non-randomised studies tended to show larger treatment effects. The summary results of the two types of design differed beyond chance in seven cases (16%). Discrepancies beyond chance were less common when only prospective studies were considered (8%). In 28 cases (62%), the natural logarithm of the odds ratio differed by at least 50%, and in 15 cases (33%), the odds ratio varied at least two-fold between non-randomised studies and randomised trials. This led the authors to conclude that despite good correlation between randomised trials and non-randomised studies – in particular, prospective studies – discrepancies beyond chance do occur and differences in estimated magnitude of treatment effect are very common.

Kunz and Oxman[42] compared the results of randomised and non-randomised controlled trials of the same interventions. They found that failure to use random allocation and concealment of allocation was associated with relative increases in estimates of effects of 150% or more, relative decreases of up to 90%, inversion of the estimated effect and, in some cases, no difference. On average, failure to use randomisation or adequate concealment of allocation resulted in larger estimates of effect owing to a poorer prognosis in non-randomly selected control groups compared with randomly selected control groups.

Sacks and colleagues[43] compared the use of RCTs and historically controlled trials (HCTs) for clinical trials. They found that in 44 out of 56 HCTs (79%) the therapy was better than the control regimen, but only ten out of 50 RCTs (20%) agreed. For each therapy, the treated patients in RCTs and HCTs had similar outcomes. The difference between RCTs and HCTs of the same therapy was largely due to differences in outcome for the control group, with the HCT control patients generally faring worse than the RCT control groups. Adjustment of the outcomes of HCTs for prognostic factors, when possible, did not appreciably change the results. Sacks and colleagues concluded that biases in patient selection could irretrievably weight the outcome of HCTs in favour of new therapies, while RCTs may miss clinically important benefits.

Schulz and colleagues[44] examined whether inadequate approaches to RCT design and execution are associated with evidence of bias in estimating treatment effects. They found that trials in which concealment was either inadequate or unclear yielded larger estimates of treatment effects. Odds ratios were exaggerated by 41% for inadequately concealed trials and by 30% for vaguely concealed trials. This led the authors to conclude that inadequate methodological approaches in controlled trials, particularly those representing poor allocation concealment, are associated with bias.

Shadish and Ragsdale[45] sought to question the practice in psychotherapy meta-analyses of combining results from controlled experiments that use random and non-random assignment without examining whether the two methods give the same answer. With the use of outcome studies of marital and family therapy, 64 experiments using random assignment yielded consistently higher mean post-test effects and less variable post-test effects than 36 studies using non-random assignment. This difference was reduced by about half by taking into account various covariates, especially pre-test effect size levels and various characteristics of control groups. Shadish and Ragsdale concluded that studies using non-random assignment may produce acceptable approximations to results from randomised experiments under some circumstances, but that reliance on results from randomised experiments as the gold standard is still well founded.

Contrary to the five studies above, the following six studies found similar estimates of treatment effects from observational studies or non-randomised clinical trials and RCTs.

Balk and colleagues[46] examined the association between quality measures and treatment effect size in RCTs. Three of the 24 quality measures included are related to randomisation: 'Randomisation methods described', 'Central randomisation site' and 'Allocation concealment'. They found that none of these, nor any of the other 21 quality measures, was statistically significantly associated with treatment effect. Therefore, the authors concluded that individual quality measures are not reliably associated with the strength of treatment effect across studies and medical areas.

Benson and Hartz[47] compared the results of observational studies with those of RCTs. In most cases, the estimates of the treatment effects from

observational studies and RCTs were similar. In two of the 19 analyses of treatment effects, the combined magnitude of the effect in observational studies was outside the 95% confidence interval (CI) for the combined magnitude in the RCTs. Benson and Hartz concluded that little evidence was found that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than, or qualitatively different from, those obtained in RCTs.

Britton and colleagues[48] explored those issues related to the process of randomisation that may affect the validity of conclusions drawn from the results of RCTs and non-randomised studies. No obvious patterns emerged: neither the RCTs nor the non-randomised studies consistently gave larger or smaller estimates of the treatment effect. The type of intervention did not appear to be influential, though the authors state that more comparisons need to be conducted before definite conclusions can be drawn. Adjustment for differences in baseline prognostic factors in non-randomised studies often changed the treatment effect size, but not significantly; importantly, the direction of change was inconsistent. Britton and colleagues concluded that results of RCTs and non-randomised studies do not inevitably differ, and that the available evidence suffers from many limitations. Furthermore, they concluded that it may be possible to minimise any differences by ensuring that subjects included in each type of study are comparable. The effect of adjustment for baseline differences between groups in non-randomised studies is inconsistent but, where it is done, it should involve rigorously developed formulae. Existing studies have generally been too small to assess the impact of such adjustment.

Concato and colleagues[49] compared the results of RCTs and observational studies that examined the same clinical topic according to the type of research design. They found that, for the five clinical topics and 99 original articles evaluated, the average results of the observational studies were remarkably similar to those of the RCTs. From this they concluded that the results of well-designed observational studies (with either a cohort or case–control design) do not systematically overestimate the magnitude of the effects of treatment compared with those in RCTs on the same topic.

MacLehose and colleagues[50,51] investigated the association between methodological quality and the magnitude of estimates of effectiveness derived from RCTs and quasi-experimental and

observational (QEO) studies. In a comparison of the RCT and QEO study estimates of effectiveness of any intervention, where both estimates were reported in a single paper, the authors showed that QEO study estimates of effectiveness could be valid if important confounding factors are controlled for. In a comparison of the RCT and QEO study estimates of effectiveness for specified interventions, where the estimates were reported in different papers, the authors found no association between study quality and effect size for either intervention, after taking account of study design. The primary aim of quantifying any association between methodological quality and effect size was thwarted by several obstacles. For the first comparison, the authors were unable to draw strong conclusions because of the paucity of evidence and the potentially unrepresentative nature of the evidence they reviewed. For the second comparison, the authors were unable adequately to distinguish and measure the variations in different aspects of quality between studies.

Moyer and Finney[52] compared the participants, methodological features and post-treatment functioning in randomised and non-randomised studies of alcohol treatment. They found that abstinence and improvement rates following active treatment were similar for the two types of design, even when differences in study features were controlled. Moyer and Finney concluded that the contrasting strengths and weaknesses of randomised and non-randomised studies suggest that they should be considered as complementary forms of treatment evaluation in the alcohol treatment field and, perhaps, more generally.

Three studies suggest that it may be possible to minimise any differences by ensuring that subjects included in each type of study are comparable.[45,48,50]

Conclusions

The studies identified by this scoping exercise illustrate the type of research that has been conducted into the quantification of selection bias in clinical research. This research has focused on testing the importance of conducting good quality randomised trials when testing hypotheses relating to relative treatment effects in clinical research. Which study designs have been studied is not always transparent; in some cases poorly conducted RCTs (quasi-RCTs) have been pooled together with good quality observational studies.

Overall, the findings indicate that RCTs and observational studies do not necessarily produce different results. Furthermore, where their results do differ significantly from one another, the direction of that difference is neither constant nor predictable. Comparisons of RCTs with good quality cohort studies suggest that the findings from these two designs of study can be complementary. Furthermore, where the populations included in the available RCTs are highly selected, the results from the good quality-controlled cohort study may more accurately reflect those of the 'true' patient population.

Importantly, this scoping exercise failed to uncover any information on the quantification of bias in uncontrolled studies. Such research designs are likely to be less reliable and more susceptible to bias than any controlled trial or controlled observational study, and yet are frequently used as part of the evidence base for the relative treatment effects of healthcare technologies.

Just before submission of this report, an HTA report was published on 'Evaluating non-randomised intervention studies'.[53] One of the three reviews included in the report focused on empirical evidence of bias associated with non-randomised studies. The authors found eight studies, seven of which are among the studies described above, and reached similar conclusions: "The only deducible conclusions were (a) results of randomised and non-randomised studies sometimes, but not always, differ and (b) both similarities and differences may often be explicable by other confounding factors."[53–58]

In addition, Deeks and colleagues used non-randomised studies – generated from two large, multicentre RCTs by selectively resampling trial participants according to allocated treatment, centre and period – to examine the ability of case-mix adjustment methods to correct for selection bias introduced by non-random allocation. This led them to conclude that: "Standard methods of casemix adjustment do not guarantee removal of bias. Residual confounding may be high even when good prognostic data are available, and in some situations adjusted results may appear more biased than unadjusted results."[53–58]

In decision-analytic modelling for purposes of cost-effectiveness analysis, there are frequently situations where estimates of relative treatment effects from randomised controlled trials are not available. The need to make a decision – in some cases preliminary, subject to further research – means that data from observational studies have to be used to estimate these parameters. To reflect the limitations of such data sources in decision models, the ultimate objective has to be to reflect the possible bias in the parameterisation of the model. In practice, several strategies could be used to achieve this.[53–58] The first is to increase the uncertainty in the estimate of treatment effect; that is, to make the distribution that characterises the uncertainty in the parameter more diffuse.

The second approach is to include an additional 'bias' parameter in the model, which combines with the estimate of treatment effect from the observational study to produce a 'bias-adjusted' measure of treatment effect. A key issue with this approach, however, is whether, on average, the bias works to overestimate or underestimate the treatment effect for the new treatment. The literature reviewed here suggests that the direction of bias is not systematic in its direction. One solution is to assume that, on average, the biases will neither overestimate nor underestimate the treatment effect, but that there is considerable dispersion in the distribution of bias. In effect, this would amount to the same thing as adding additional uncertainty to the estimate of treatment effect.

More research is clearly required on how to handle potential bias in decision models. Indeed, the complexity of the problems is more extensive when randomised trial evidence is not available and observational evidence is the only means of estimating a treatment effect. For example, sometimes trial data are available, but from small studies, and larger observational data sets are known to exist. In this instance, the issue of whether to pool observational and trial evidence needs to be considered, and also whether the observational evidence needs to be down-weighted to reflect its potential biases.[53–58]

For the purposes of modelling to support reimbursement decision-making, there is a strong argument for the use of scenario analysis; that is, to present decision-makers with alternative estimates of cost-effectiveness based on alternative ways of characterising and reflecting the potential bias in estimates of treatment effects in decision models.

Chapter 5

Discussion

Review of guidelines for quality assessment in decision models

The review of current guidelines showed that although authors may provide a consistent message regarding some aspects of modelling, in other areas the attributes presented in different guidelines conflict. In developing the synthesised guideline and checklist it has been necessary to take a line on these areas of disagreement. However, further discussion and research may well be required on some of these. For example, the extent to which the availability of data should determine the structure of a model is an area of disagreement in the literature. Here, it has been argued that, in principle, structure should not be influenced by the extent or quality of the data available to populate a model, but the authors accept that, in practice, this will not always be possible and more detailed guidance would be of value for analysts.

The focus of the research here was on the decision models used to inform the NICE technology appraisal process. Therefore, the position taken in the synthesised guideline and checklist on some aspects of good quality in decision modelling reflects the needs and constraints of that decision-maker. It is recognised that, for other types of decision-maker, some features of the checklist may need to be altered.

A preliminary assessment of the checklist showed that, in general, the checklist appears to perform well, in terms of identifying those aspects of the model that should be of particular concern to the reader. The checklist cannot, however, provide answers to the appropriateness of the model structure and structural assumptions, as these may be seen as a general problem with generic checklists and do not reflect any shortcoming with the synthesised guidance and checklist developed here. The assessment of the checklist, as well as feedback from the EAG, indicated the importance of its use in conjunction with a more general checklist or guidelines on economic evaluation.[59]

Methods in decision modelling

The review of current guidance of good quality

decision-analytic modelling for health technology assessment highlighted a number of methodological areas that have not received attention in the literature on good practice. There are many of these areas and, therefore, it was only possible to consider two specific methods areas in decision modelling: the identification of parameter estimates from published literature, and the issue of adjusting treatment effect estimates taken from observational studies for potential bias.

Chapter 4 reported a case-study decision model used to explore the use of specific searches to inform parameter estimates. It demonstrated the need to use an iterative process of communication between the information specialist and the modeller when trying to identify appropriate model data. The searches helped to identify important data to populate the model. However, the quality of life (utility) searches proved to be problematic. It is difficult to recommend specific databases for parameter estimates as the case study generated a small number of papers from each database and it appears that a more focused search strategy may miss relevant records. However, the limited time available to assess papers means that less focused searching may not be practical. Further research is required to assess what is missed in the development of generic search strategies for particular types of data (quality of life) by less focused searching, and in assessing the additional value of these types of search in terms of identifying relevant data.

Again, there is much to be gained from sharing the experience of the teams undertaking technology assessment for NICE, and of those undertaking literature searches for decision models in health technology assessment more generally. Over time, efficient searches will evolve for the main categories of parameter estimate used in decision models.

To ensure efficiency in searching to populate decision models, methods need to be available to identify those parameters to which the results of a model are most sensitive. Although standard sensitivity analysis offers one approach, formal methods of value of information analysis offer the

potential of a major contribution in this area. Although these methods are increasingly being used to inform priority setting in primary research,[1,3–6] the role of informing priority setting in literature searching is underdeveloped and may be considered a future research priority.

When populating a model, an analyst may encounter estimates of relative treatment effect from non-randomised studies and, given the need to make a decision, it is often necessary to use these effect estimates. The non-random nature of these data does, however, mean that estimates of treatment effect can be subject to bias. Ideally, it would be desirable to reflect this in the parameterisation of the decision model. The starting point for doing this is the quantification of the extent and direction of bias; indeed, if reliable quantification of bias were available, a range of options would be available for how this could be reflected in a decision model.

Empirical studies have, in general, concluded that RCTs and observational studies do not necessarily produce different results and, when cohort studies are of particularly good quality, the two can generate similar results. Moreover, where the populations included in the available RCTs are highly selected, the results from the good quality controlled cohort study may more accurately mirror those of the 'true' patient population, making the results more generalisable. The review here also indicated that the direction of bias in observational data is not easily predictable. This makes adjustment within a decision model difficult to undertake.

Implications for the NICE technology assessment and appraisal process

Decision modelling is central to the NICE technology assessment process, given the complexity of the decision problems, the unavoidable nature of the decisions and the imperfections of the evidence base for all appraisals. However, it is essential to assess the quality of those models that are developed to inform the Appraisal Committee. Currently, the models submitted by consultees are subject to critical appraisal by the assessment teams. Those models developed by the assessment teams are peer-reviewed by external referees and commented upon by interested parties, including consultees. One purpose of developing the

synthesised guideline and checklist was to provide a framework for this critical appraisal by various parties. There are at least two ways in which this might be achieved. The first is the use of the guideline and checklist by groups that are reviewing other analysts' models. For example, the assessment teams may use it as a framework for their assessment of manufacturers' models; and independent referees may find it of value in their critical review of the assessment teams' models. The second way in which the guideline/checklist could be used is by the various analysts as they develop their models; in other words, to use it as a check on how they are developing and reporting their analyses.

An EAG (see Chapter 3) was convened and discussed the potential role of the guidance and checklist in the NICE TAR process. In general, feedback from the EAG was positive and the group was confident that the guidance and checklist would be a useful tool in the NICE TAR process for the assessment team, technical leads and committee members. A number of points emerged in discussion with the EAG and subsequently.

- It is important to consider how the checklist fits in with more general guidance on economic evaluation. It is important to realise that the checklist is not meant to be used exclusively to determine a model's quality. Indeed, it is expected that users will refer to the synthesised guidance at each part of the critical appraisal process.
- The checklist identifies good and bad models, but what about the middle ground? Often it is relatively easy for the Appraisal Committee to identify when a model is of particularly good or poor quality. However, the benefit of the guidance and checklist will be in the identification of those models that are not at the extremes; that is, those that have both good and bad qualities. The checklist must be able to identify these aspects and provide guidance on which are most likely to compromise the model's results.
- The checklist provides an explicit way of reviewing a model. Although the checklist is not a substitute for critical review, it can allow the interested party to gain an understanding of the model's key components and assumptions.
- There is a need to define who the checklist is aimed at. It is likely that there are several potential users of the checklist: the TAR teams, as a marker of good model practice; technical leads; and the Appraisal Committee.

- Currently, no common checklist is used in the review process. It is hoped that further discussion between the assessment teams and NICE will lead to the use of the same checklists across the groups. This would include those used for economic evaluation in general, as well as decision models in particular.
- The checklist described here may be considered the starting point in terms of a checklist for decision models, but it is recognised that it is far from the finished article. This report describes how the checklist was used to assess the quality of three case-study decision models developed as part of the NICE technology appraisal process. Through extensive use by the assessment teams, the value of the checklist can be further assessed. Over time, it can be continually developed, with the purpose of providing a more sensitive framework for critical appraisal.

Recommendations for research

The process of developing the synthesised guideline/checklist uncovered numerous areas where further methods research may be of value. However, the focus here is on those particular methods considered in this report. On that basis, the following areas for further research have been identified.

- A systematic scoping and review of the literature pertaining to the quantification of selection bias in non-controlled studies and in controlled observational studies (cohort studies and case–control studies) is needed. More empirical research is needed to define further the level of bias in the different non-RCT study designs. Studies are needed that compare results from RCTs with those from each of the other trial designs (quasi-RCTs, non-randomised clinical trials) and with observational studies (controlled cohort studies and case–control studies).
- On the basis of improved information on the extent and direction of bias in observational studies, there is a need to assess the strengths and weaknesses of alternative ways to adjust for bias in a decision model. This process could usefully be informed by case studies, perhaps in the context of recent NICE appraisals.
- There is a need to develop and test search strategies to identify particular types of data of use within decision modelling. There is also a need to assess whether a range of resources needs to be searched for each parameter or whether a checklist of one (or several) resource per parameter type can be derived.
- There is a need for research into how to prioritise searching for parameter estimates. The use of value of information methods, in particular, is worthy of consideration.

Acknowledgements

The authors would like to acknowledge the contribution made to this project by the Expert Advisory Group in terms of the specifics of the synthesised guideline and checklists, and the development of the authors' ideas in general relating to quality in decision modelling. However, the views expressed, and any errors, are solely those of the authors.

Contributions of the authors

Zoe Philips (Research Fellow) was responsible for the review of guidelines. She was also involved in the selection of studies, data extraction and report writing. Laura Ginnelly (Research Fellow) was involved in the review of guidelines and review of studies for Chapter 4. She was also involved in the selection of studies, data extraction and report writing. Mark Sculpher (Professor of Health and also Economics) provided input at all stages. He contributed to the review of guidelines and report writing. Karl Claxton (Senior Lecturer) provided input at all stages. He also contributed to the review of guidelines and report writing. Su Golder (Information Officer) devised the search strategies and carried out the literature searches. She was also involved in the writing of Chapter 4 of the report. Rob Riemsma (Reviews Manager) provided input at all stages. He was also involved in the writing of Chapter 4. Nerys Woolacott (Reviewer) was involved in the writing of Chapter 4. Julie Glanville (Associate Director) was involved in the literature searches for the review of guidelines and Chapter 4. She also commented on various drafts of the report.

References

1. Claxton K, Sculpher M, Drummond M. A rational framework for decision making by the National Institute for Clinical Excellence. *Lancet* 2002; **360**:711–15.

2. Briggs A, Goeree R, Blackhouse G, O'Brien B. Probabilistic analysis of cost-effectiveness models: choosing between treatment strategies for gastroesophageal reflux disease. *Med Decis Making* 2002;**22**:290–308.

3. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ* 1999; **18**:341–64.

4. Claxton K, Posnett J. An economic approach to clinical trial design and research priority-setting. *Health Econ* 1996;**5**:513–24.

5. Claxton K, Neumann PJ, Araki S, Weinstein MC. Bayesian value-of-information analysis – an application to a policy model of Alzheimer's disease. *Int J Technology Assess Health Care* 2001; **17**:38–55.

6. Claxton K, Neuman PJ, Araki SS, Weinstein MC. The value of information: an application to a policy model of Alzheimer's disease. *Int J Technol Assess Health Care* 2001;**17**:38–55.

7. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models: a suggested framework and example of application. *Pharmacoeconomics* 2000;**17**:461–77.

8. Akehurst R, Anderson P, Brazier J, Brennan A, Briggs A, Buxton M, *et al*. Decision analytic modelling in the economic evaluation of health technologies – a consensus statement. *Pharmacoeconomics* 2000;**17**:443–4.

9. Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, *et al*. Modelling in economic evaluation: an unavoidable fact of life. *Health Econ* 1997;**6**:217–27.

10. McCabe C, Dixon S. Testing the validity of cost-effectiveness models. *Pharmacoeconomics* 2000; **17**:501–13.

11. Halpern MT, McKenna M, Hutton J. Modeling in economic evaluation: an unavoidable fact of life. *Health Econ* 1998;**7**:741–2.

12. Hay JW. Economic modeling and sensitivity analysis. *Value Health* 1998;**1**:187–93.

13. ISPOR Task Force. Principles of good practice for decision analytic modeling in health-care evaluation. *Value Health* 2003;**6**:9–17.

14. Habbema JD, Bossuyt PM, Dippel DW, Marshall S, Hilden J. Analysing clinical decision analyses. *Stat Med* 1990;**9**:1229–42.

15. Nuijten MJC, Pronk MH, Brorens MJA, Hekster YA, Lockefeer JHM, de Smet P, *et al*. Reporting format for economic evaluation: Part II: Focus on modelling studies. *Pharmacoeconomics* 1998; **14**:259–68.

16. Sendi PP, Craig BA, Pfluger D, Gafni A, Bucher HC. Systematic validation of disease models for pharmacoeconomic evaluations. Swiss HIV Cohort Study. *J Eval Clin Pract* 1999;**5**:283–95.

17. Ramsey SD, Sullivan SD. Weighing the economic evidence: guidelines for critical assessment of cost-effectiveness analyses [Review]. *J Am Board Fam Pract* 1999;**12**:477–85.

18. Eddy D. Technology assessment: the role of mathematical modelling. In Mosteller F, editor. *Assessing medical technologies*. Washington, DC: National Academy Press; 1985. pp. 144–60.

19. Soto J. Health economic evaluations using decision analytic modeling. Principles and practices – utilization of a checklist to their development and appraisal. *Int J Technol Assess Health Care* 2002; **18**:94–111.

20. Weinstein MC, Toy EL, Sandberg EA, Neumann PJ, Evans JS, Kuntz KM, *et al*. Modeling for health care and other policy decisions: uses, roles, and validity. *Value Health* 2001;**4**:348–61.

21. Sonnenberg FA, Roberts MS, Tsevat J, Wong JB, Barry M, Kent DL. Toward a peer review process for medical decision analysis models. *Med Care* 1994;**32**:JS52-64.

22. Hay J, Jackson J. Panel 2: methodological issues in conducting pharmacoeconomic evaluations – modeling studies. *Value Health* 1999;**2**:78–81.

23. Drummond M, Jefferson T. Guidelines for authors and peer reviewers of economic submissions to the *BMJ*. *BMJ* 1996;**313**:275–83.

24. Canadian Coordinating Office for Health Technology Assessment (CCOHTA). Guidelines for economic evaluation of pharmaceuticals: Canada. *Int J Technol Assess Health Care* 1995;**11**:796–7.

25. Barton P, Bryan S, Robinson S. Modelling in the economic evaluation of health care: selecting the appropriate approach. *J Health Services Research and Policy* 2004;**9**:110–18.

26. Briggs A. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics* 2000;**17**:479–500.

27. Hjelmgren J, Berggren F, Andersson F. Health economic guidelines – similarities, differences and some implications. *Value Health* 2001;**4**:225–50.

28. Payne N, Chilcott J, McGoogan E. Liquid-based cytology in cervical screening. A report by the School of Health and Related Research, University of Sheffield, for the NCCHTA on behalf of NICE; May 2000.

29. Kendrick M, Johnson K. Long term treatment of multiple sclerosis with interferon beta may be cost effective. *Pharmacoeconomics* 2000;**18**:45–53.

30. Turner D, Wailoo A, Nicholson K, Cooper N, Sutton A, Abrams K. Systematic review and economic modelling for the prevention and treatment of influenza A and B. A report by the Departments of Epidemiology and Public Health and Microbiology and Immunology, University of Leicester and ScHARR, University of Sheffield, for the NHS R&D HTA Programme; April 2002.

31. Paisley S. Rapid reviews: less than, greater than or equal to systematic reviews. Presented at Information Specialists' Meeting: Methodologies for Technology Assessment Reports; London; 2002.

32. Royle P, Waugh N. Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system. *Health Technol Assess* 2003;**7**(34).

33. Claxton K, Ginnelly L, Sculpher M, Philips Z, Palmer S. A pilot study on the use of decision theory and value of information analysis as part of the NHS health technology assessment programme. *Health Technol Assess* 2004;**8**(31).

34. Brettle AJ, Long AF, Grant MJ, Greenhalgh J. Searching for information on outcomes: do you need to be comprehensive? *Quality in Health Care* 1998;**7**:163–7.

35. Paisley S, Booth A. Chapter 12: Health-related quality of life studies. In Etext on health technology assessment (HTA) information resources; 2002. URL:http://www.nlm.nih.gov/nichsr/ehta/chapter12.html. Accessed 19 November 2003.

36. Williams GJ, Lee A, Craig JC. Long-term antibiotics for preventing recurrent urinary tract infection in children (Cochrane Review). In *The Cochrane Library* (Issue 4). Chichester: John Wiley; 2003.

37. Claxton K. Bayesian approaches to the value of information: implications for the regulation of new pharmaceuticals. *Health Econ* 1999;**8**:269–74.

38. Ioannidis JP, Karassa FB. Bias in uncontrolled therapeutic trials in rheumatology due to selection of populations with extreme characteristics. *J Rheumatol* 2001;**28**:1881–7.

39. Marcus SM. Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. *J Clin Epidemiol* 1997;**50**:823–8.

40. Sackett DL. Bias in analytic research. *Journal of Chronic Disease* 1979;**32**:51–63.

41. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;**286**: 821–30.

42. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998; **317**:1185–90.

43. Sacks H, Chalmers TC, Smith HJ. Randomized versus historical controls for clinical trials. *Am J Med* 1982;**72**:233–40.

44. 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.

45. Shadish WR, Ragsdale K. Random versus nonrandom assignment in controlled experiments: do you get the same answer? *J Consult Clin Psychol* 1996;**64**:1290–305.

46. Balk EM, Bonis PAL, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, *et al*. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;**287**:2973–82.

47. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342**:1878–86.

48. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;**2**(13).

49. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**:1887–92.

50. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AMS. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;**4**(34).

51. Reeves BC, MacLehose RR, Harvey IM, Sheldon TA, Russel IT, Black AMS. Comparisons of effect size estimates derived from randomised and non-randomised studies. In Black N, Brazier, J, Fitzpatrick R, Reeves B, editors. *Health services research methods: a guide to best practice*. London: BMJ Publishing Group; 1998. pp. 73–85.

52. Moyer A, Finney JW. Randomized versus nonrandomized studies of alcohol treatment: participants, methodological features and posttreatment functioning. *J Stud Alcohol* 2002; **63**:542–50.

53. Deeks J, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, *et al.* Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;**7**(27).

54. Ibrahim J, Chen M-H. Power transform prior distributions. *Statistical Science* 2000;**15**:46–60.

55. Eddy D, Hasselblad V, Shachter R. A Bayesian method for synthesizing evidence. The Confidence Profile Method. *Int J Technol Assess Health Care* 1990;**6**:31–55.

56. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer. *Stat Med* 2000;**19**:3359–76.

57. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. London: John Wiley; 2003.

58. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* 2001;**10**:277–303.

59. National Institute for Clinical Excellence. *Guide to the methods of technology appraisal*. Draft for consultation. London: NICE; August 2003.

60. Halpern MT, Luce BR, Brown RE, Geneste B. Health and economic outcomes modeling practices: a suggested framework. *Value Health* 1998; **1**:131–47.

61. Ramsey SD. Evaluating evidence from a decision analysis. *J Am Board Fam Pract* 1999;**12**:395–402.

62. Murphy Jr GM, Jia XC, Song Y. Macrophage inflammatory protein 1-alpha mRNA expression in an immortalized microglial cell line and cortical astrocyte cultures. *J Neurosci Res* 1995;**40**:755–63.

61

Appendix I

Search strategies for guidelines of good practice in decision modelling

Search strategies and bibliographic databases used

The following sources were searched for studies relating to guidelines for good practice in decision-analytic modelling. A range of free text terms and subject headings was used, as appropriate. Details of the search strategies are below.

CRD internal administration databases
NHS EED (searched 19 December 2002)

Databases on the Internet
EMBASE (1980 to week 50 2002)
(searched 17 December 2002 on OvidWeb Gateway at http://gateway.ovid.com/athens/)
MEDLINE (1966 to October week 5 2002))
(searched 17 December 2002 on OvidWeb Gateway at http://gateway.ovid.com/athens/)
PREMEDLINE (up to 13 December 2002)
(searched 17 December 2002 on OvidWeb Gateway at http://gateway.ovid.com/athens/)
SCI (1981 onwards)
(searched 19 December 02 on Web of Science at http://wos.mimas.ac.uk/)
SSCI (1981 onwards)
(searched 19 December 2002 on Web of Science at http://wos.mimas.ac.uk/)

CD-ROM resources
EconLit (1969 to November 2002)
(searched 18 December 2002 on ARC Silverplatter)
HEED (Issue: December 2002)
(searched 19 December 2002 on stand-alone CD-ROM)

A date restriction of 1990 was placed on all of the literature searches as the methods of economic evaluation are rapidly evolving and it was thought that papers before this date would not be relevant to current practices in health technology assessment. Although the project had no capacity to undertake translations, no language restrictions were applied to the searches.

CRD internal administration databases

NHS EED
Searched 19 December 2002 on CAIRS
NHS EED was searched via the NHS CRD's internal administration system. This provides a more up-to-date version of the database than the version on the Cochrane Library or the Internet and includes additional records to those in the public databases. The search strategy used was as follows:

S (economic(w)model$ or econometric(w)model$ or markov(w)model$ or mathematical(w)model$ or cost$(w)model$ or pharmacoeconomic$(w)model$ or decision(w)model$ or decision(w)tree$ or decision(w)triage or decision(w)data or decision(w)analytic$ or decision(w)analysis or crystal(w)ball)

S (checklist$ or check(w)list$ or standards or standardi?ation or peer(w)review or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline$ or validation or checkpoint$ or properly or critically appraise or problems or limitations or rating(w)scale$ or framework$ or protocol$ or audit or principles or methodolog$ or validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons)

S (economic(w)evaluation$ or economic(w)analysis or economic(w)stud$ or economic(w)submission$)/til

S (checklist$ or check(w)list$ or standards or standardi?ation or peer(w)review or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline$ or validation or checkpoint$ or properly or critically appraise or problems or limitations or rating(w)scale$ or framework$ or protocol$

63

or audit or principles or methodolog$ or validate or validation or evaluating or properties or guidance or integrity or pros or cons)/til

S s1 and s2

S s3 and s4

S s5 or s6

This identified 1371 records.

Databases on the Internet

EMBASE (1980 to week 50 2002)
Searched 17 December 2002 on OvidWeb Gateway (http://gateway.ovid.com/athens/)

1. (checklist? or check list? or standards or standardi?ation or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?).ti
2. (properly or critically appraise or problems or limitations or rating scale? or framework$ or protocol? or audit or principles or methodolog$).ti. or *"Methodology"/
3. (validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons).ti.
4. (decision adj (tree or triage or data or analytic or analysis)).ti.
5. (*"STOCHASTIC MODEL"/ or *"MATHEMATICAL MODEL"/ or *"STATISTICAL MODEL"/ or *"THEORETICAL MODEL"/ or *"MODEL"/ or *"INDIVIDUAL BASED POPULATION MODEL"/ or *"PROCESS MODEL"/ or *"decision support system"/ or *"medical decision making"/ or *"decision theory"/) and cost$.hw.
6. (*"STOCHASTIC MODEL"/ or *"MATHEMATICAL MODEL"/ or *"STATISTICAL MODEL"/ or *"THEORETICAL MODEL"/ or *"MODEL"/ or *"INDIVIDUAL BASED POPULATION MODEL"/ or *"PROCESS MODEL"/ or *"decision support system"/ or *"medical decision making"/ or *"decision theory"/) and economic$.hw.
7. ((economic? or pharmacoeconomic? or decision? or cost? or costing?) and model$).ti.
8. (markov or crystal ball).ti.
9. (*"health economics"/ or *"economic evaluation"/ or *"cost benefit analysis"/ or *"cost control"/ or *"cost effectiveness analysis"/ or *"cost minimization analysis"/ or *"cost of illness"/ or *"cost utility analysis"/) and model$.hw.
10. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision model$) adj2 (checklist? or check list? or standards or standardi?ation or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?)).ab.
11. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision model$) adj2 (properly or critically appraise or problems or limitations or rating scale$ or good practice$ or framework$ or protocol$ or audit or principles or methodolog$)).ab.
12. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision model$) adj2 (validate or validation or evaluating or properties or guidance or integrity or avoiding bias or evaluation or pros or cons)).ab.
13. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (checklist? or check list? or standard$ or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?)).ab.
14. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (properly or critically appraise or problems or limitations or rating scale$ or framework$ or protocol$ or audit or principles or methodolog$)).ab.
15. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons)).ab.
16. ((economic evaluation? or economic analysis or economic stud$ or economic submission?) and guideline$).ti
17. or/10-15
18. or/4-9
19. or/1-3
20. 19 and 18
21. 16 or 17 or 20 limit 20 to yr=1990-2004

This resulted in 631 records.

MEDLINE (1966 to October week 5 2002)
Searched 17 December 2002 on OvidWeb Gateway
(http://gateway.ovid.com/athens/)

1. (checklist? or check list? or standards or standardi?ation or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?).ti.
2. (properly or critically appraise or problems or limitations or rating scale? or framework$ or protocol? or audit or principles or methodolog$).ti.
3. (validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons).ti.
4. (decision adj (tree or triage or data or analytic or analysis)).ti.
5. *"models, economic"/ or *"models, econometric"/
6. (*"decision support techniques"/ or *"data interpretation, statistical"/ or *"decision theory"/ or *"models, statistical"/ or *"likelihood functions"/ or *"linear models"/ or *"logistic models"/ or *"proportional hazards models"/) and *"costs and cost analysis"/
7. ((economic? or pharmacoeconomic? or decision? or cost? or costing?) and model$).ti.
8. (markov or crystal ball).ti.
9. *"markov chain"/
10. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision model$) adj2 (checklist? or check list? or standards or standardi?ation or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?)).ab.
11. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision model$) adj2 (properly or critically appraise or problems or limitations or rating scale$ or good practice$ or framework$ or protocol$ or audit or principles or methodolog$)).ab.
12. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision model$) adj2 (validate or validation or evaluating or properties or guidance or integrity or avoiding bias or evaluation or pros or cons)).ab.

13. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (checklist? or check list? or standard$ or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?)).ab.
14. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (properly or critically appraise or problems or limitations or rating scale$ or framework$ or protocol$ or audit or principles or methodolog$)).ab.
15. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons)).ab.
16. ((economic evaluation? or economic analysis or economic stud$ or economic submission?) and guideline$).ti
17. or/10-15
18. or/4-9
19. or/1-3
20. 19 and 18
21. 16 or 17 or 20 limit yr=1990-2002

This search identified 670 records.

PREMEDLINE (up to 13 December 2002)
Searched 17 December 2002 on OvidWeb Gateway
(http://gateway.ovid.com/athens/)

1. (checklist? or check list? or standards or standardi?ation or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?).ti.
2. (properly or critically appraise or problems or limitations or rating scale? or framework$ or protocol? or audit or principles or methodolog$).ti.
3. (validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons).ti.
4. (decision adj (tree or triage or data or analytic or analysis)).ti.
5. ((economic? or pharmacoeconomic? or decision? or cost? or costing?) and model$).ti.
6. (markov or crystal ball).ti.
7. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision

model$) adj2 (checklist? or check list? or standards or standardi?ation or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?)).ab.

8. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision model$) adj2 (properly or critically appraise or problems or limitations or rating scale$ or good practice$ or framework$ or protocol$ or audit or principles or methodolog$)).ab.

9. ((markov model$ or economic model$ or mathematical model$ or cost$ model$ or pharmacoeconomic model$ or decision model$) adj2 (validate or validation or evaluating or properties or guidance or integrity or avoiding bias or evaluation or pros or cons)).ab.

10. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (checklist? or check list? or standard$ or peer review$ or rules or critiquing or criteria or good or bad or correct$ or bias or fundamentals or recommend$ or best or strength$ or weakness$ or quality or qualities or validity or guideline? or validation or checkpoint?)).ab.

11. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (properly or critically appraise or problems or limitations or rating scale$ or framework$ or protocol$ or audit or principles or methodolog$)).ab.

12. ((decision tree or decision triage or decision data or decision analytic$ or decision analysis or crystal ball) adj2 (validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons)).ab.

13. ((economic evaluation? or economic analysis or economic stud$ or economic submission?) and guideline$).ti

14. or/7-12

15. or/4-6

16. or/1-3

17. 15 and 16

18. 13 or 14 or 17

This search identified 19 records.

SCI (1981 onwards) and SSCI (1981 onwards)
Searched 19 December 2002 on Web of Science (http://wos.mimas.ac.uk/)

The Web of Science search interface only allows a limited number of characters to be searched in any one field at any one time. The search was therefore divided into three phases and any duplicate records were identified after loading into the Endnote library.

Search 1
The following terms were restricted to the title field only and identified 204 records;

(economic evaluation* or economic analysis or economic stud* or economic submission*) and (checklist* or check list* or standard* or peer review or rules or critiquing or criteria or good or bad or correct* or bias or fundamentals or recommend* or best or strength* or weakness* or quality or qualities or validity or guideline* or validation or checkpoint* or properly or critically appraise or problems or limitations or rating scale* or framework* or protocol* or audit or principles or methodolog* or validate or validation or evaluating or properties or guidance or integrity or pros or cons)

Search 2
The terms in searches 2 and 3 were restricted to article title, keywords and abstract.

(economic model* or econometric model* or markov model* or mathematical model* or cost* model* or pharmacoeconomic model* or decision model* or decision tree* or decision triage or decision data or decision analytic* or decision analysis or crystal ball) and (checklist* or check list* or standards or standardi?ation or peer review or rules or critiquing or criteria or good or bad or correct* or bias or fundamentals recommend* or best or strength* or weakness* or quality or qualities or validity)

This search identified 175 records.

Search 3
economic model* or econometric model* or markov model* or mathematical model* or cost* model* or pharmacoeconomic model* or decision model* or decision tree* or decision triage or decision data or decision analytic* or decision analysis or crystal ball) and (guideline* or validation or checkpoint* or properly or critically appraise or problems of limitations or rating scale* or framework* or protocol* or audit or principles or methodolog* or validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons)

This search identified 432 records.

CD-ROM resources

EconLit (1969 to November 2002)
Searched 18 December 2002 on ARC
SilverPlatter

1. (checklist* or check list* or standards or standardi*ation or peer review* or rules or critiquing or criteria or good or bad or correct* or bias or fundamentals or recommend* or best or strength* or weakness* or quality or qualities) in ti
2. (validity or guideline* or validation or checkpoint*) in ti.
3. (properly or critically appraise or problems or limitations or rating scale* or framework* or protocol* or audit or principles or methodolog*) in ti.
4. (validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons) in ti.
5. Economic methodology in de
6. (decision near (tree or triage or data or analytic or analysis)) in ti.
7. ((economic* or pharmacoeconomic* or decision* or cost* or costing*) and model*) in ti.
8. (markov or crystal ball or economic evaluation* or economic analysis or economic stud* or economic submission*) in ti
9. markov in de
10. criteria-for-decision-making-under-risk-and-uncertainty in de
11. ((markov model* or economic model* or mathematical model* or cost* model* or pharmacoeconomic model* or decision model*) near (checklist* or check list* or standards or standardi*ation or peer review* or rules or critiquing or criteria))
12. ((markov model* or economic model* or mathematical model* or cost* model* or pharmacoeconomic model* or decision model*) near (good or bad or correct* or bias or fundamentals or recommend* or best or strength* or weakness* or quality))
13. ((markov model* or economic model* or mathematical model* or cost* model* or pharmacoeconomic model* or decision model*) near (qualities or validity or guideline* or validation or checkpoint*))
14. ((markov model* or economic model* or mathematical model* or cost* model* or pharmacoeconomic model* or decision model*) near (properly or critically appraise or problems or limitations or rating scale* or good practice* or framework* or protocol* or audit))

15. ((markov model* or economic model* or mathematical model* or cost* model* or pharmacoeconomic model* or decision model*) near (principles or methodolog*))
16. ((markov model* or economic model* or mathematical model* or cost* model* or pharmacoeconomic model* or decision model*) near (validate or validation or evaluating or properties or guidance or integrity or avoiding bias or evaluation or pros or cons))
17. ((decision tree or decision triage or decision data or decision analytic* or decision analysis or crystal ball) near (checklist* or check list* or standard* or peer review* or rules or critiquing or criteria or good or bad or correct* or bias or fundamentals or recommend* or best or strength* or weakness* or quality or qualities or validity or guideline* or validation or checkpoint*))
18. ((decision tree or decision triage or decision data or decision analytic* or decision analysis or crystal ball) near (properly or critically appraise or problems or limitations or rating scale* or framework* or protocol* or audit or principles
19. ((decision tree or decision triage or decision data or decision analytic* or decision analysis or crystal ball) near (validate or validation or evaluating or properties or guidance or integrity or evaluation or pros or cons or methodolog*))
20. #1 or #2 or #3 or #4 or #5
21. #6 or #7 or #8 or #9 or #10
22. #16 and #17
23. #11 or #12 or #13 or #14 or #15
24. #18 or #19
25. health or medical or care or (economic evaluation*) or (cost effectiveness) or (cost utility*) or (cost benefit*) or (technology assessment) or (technology appraisal*) or pharmacoeconomic*
26. 20 and 21
27. 22 and PY> 1990

This resulted in a set of 109 records.

HEED (Issue: December 2002)
Searched 19 December 2002 on CD-ROM

The following search terms were restricted to the title and abstract and identified 2085 records:

(check* or standard* or peer or rules or critiquing or criteria or good or bad or correct* or error* or fundamentals or recommend* or best or strenght* or weakness* or quality or

qualities or valid* or guideline* or properly or critically or problems or limitations or rating or framework* or protocol* or audit or principles or

methodolog* or evaluating or properties or guidance or integrity or bias or evaluation or pros or cons) and (model* or decision* or economic)

Appendix 2
Summary of guidelines available in structured format

Akehurst and colleagues[8]

Structure
Statement of scope/perspective
The scope of a decision model should be clearly specified. This would include the technologies involved, the population studied and the time-frame to which it relates.

Structural assumptions
Any underlying theory and assumptions specified relating to the structure of the model should be made clear. Justification should be given for any choices made.

Consistency of structure
The structure and data in the model should be consistent with the most appropriate information (external consistency). The causal relationships included in the model should be explained and substantiated by best available evidence (inferential soundness).

Time horizon
Within the statement of scope, the time horizon of the model should be specified.

Parsimony
A model should avoid unnecessary complexity and introduce only important variables and structural components that are relevant to the scope of the model.

Data
Data identification
The data used to populate the model should be consistent with the most appropriate information.

Data incorporation
The analyst must make clear the source of all data used in the model and any assumptions that are made. Justification should be given for any choices made.

Uncertainty
The implications of all forms of uncertainty, including methodological, structural and parameter uncertainty, must be explored appropriately.

Consistency
A good decision model should be reproducible; that is, replicated by a competent analyst with differences in results explained entirely by random variation.

Internal consistency
A good decision model should be mathematically well defined for all combinations of parameter values specified as feasible. No such values should result in inconsistency in mathematical logic of the model.

External consistency
The structure and data in the model should be consistent with the most appropriate information. The outputs of the model should be assessed in relation to best available empirical evidence.

Between-model consistency
The outputs of the model should be assessed in relation to evidence from other models.

Buxton and colleagues[9]

Structure
Strategies/comparators
Care should be taken to avoid framing the problem in an inappropriate way (e.g. by excluding a relevant alternative).

Parsimony
The model should be as simple as possible to aid understanding by decision-makers. How simple will depend on the sensitivity of the policy implications to added complexity.

Time-frame
The analyst should be careful when choosing the time-frame of the model. The analyst should distinguish between the length of the treatment effect and the time-frame of the model.

Data
Data incorporation
The analyst should respect the quality of the data in the model. Results should be presented in a way

that enables the reader to distinguish between hard data (e.g. those collected from controlled studies) and soft data (e.g. expert opinion).

Uncertainty
The analyst should explore uncertainty, and should test the robustness of the results by undertaking a thorough sensitivity analysis.

Consistency
Between-model consistency
The analyst should attempt to validate the model by assessing the level of agreement with other models.

External consistency
The analyst should attempt to validate the model by assessing the level of agreement with the results of intervention studies.

Eddy[18]

Structure
Statement of decision problem/objective
A health technology assessment modelling report should contain a statement of the problem. There should be a description of the settings to which the results of the analysis can be applied and a list of the main factors that could limit the applicability of the results.

Structural assumptions
A health technology assessment modelling report should contain:
- a description of the relevant factors and outcomes
- a description of the model
- a list of assumptions pertaining to the structure of the model (e.g. factors included, relationships and distributions)
- a discussion of how the modelling assumptions may affect the results, and estimating both the direction of the bias and the magnitude of effect should also be included.

Consistency of structure
The structure of the model should make sense to people who have a good knowledge of the problem (first order validation).

Data
Data identification
The extent to which current knowledge is incorrect will be reflected within the model. Because of this, to the greatest extent possible, models should be based on observations from well-

designed studies. The accuracy of the results of a model is limited by the accuracy of the data it uses. This is a limitation that is not limited to modelling, but a model does make data needs explicit.

Data incorporation
A health technology assessment modelling report should contain:

- a description of the data sources (including subjective estimates), with a description of the strengths and weaknesses of each source
- a list of assumptions pertaining to the data
- a list of the parameter values that will be used for the base-case analysis, and a list of the ranges in those values that will be used in the sensitivity analysis
- presentation of the results of the base-case model and sensitivity analysis.

If the analysis recommends a policy, the report should also contain:

- a list of the outcomes that required value judgements
- a description of the values assessed for those outcomes and the source of those values
- the policy recommendation and a description of the sensitivity of the recommendation to variations in the values.

Parameter uncertainty
Models can indicate the importance of uncertainty and poor data via sensitivity analysis. The results and implications of sensitivity analysis should be presented. Models can be used to estimate the value of additional research to obtain better data.

Consistency
There is no simple and universally applicable procedure for validating a model. Each case must be considered by itself. The decision to use a model should be based on a comparison with the validity of other techniques that may be used to assess the technology. A health technology assessment modelling report should contain a description of the validation methods and results.

Internal consistency
Any model should be able to match the data used to estimate parameters. Failure to pass this test strongly suggests that the structure of the model is faulty (second order validation).

External consistency
Comparing the predictions of the model with observations that were never used to construct the

model can make third order validation. However, there is a trade-off between using all available evidence for increased accuracy and not being able to test for third order validation.

Predictive validity
A fourth order validation can be defined as comparing the outcomes predicted by the model for a new and previously unobserved programme with the actual outcomes of the programme once it is implemented. This may be meaningless as the actual programme may be implemented under different conditions to those assumed in the model.

Habbema and colleagues[14]

Structure
Statement of decision problem/objective
The first stage of a clinical decision analysis is to define the clinic problem and structure it as a decision tree. This includes descriptions of the patient, of possible diagnostic and therapeutic actions, and of the possible outcomes after treatment.

Strategies/comparators
Checking should take place to see whether the decision strategies in the tree are all clinically realistic and also whether all realistic strategies are included. It is reasonable to consider whether strategies that are executed conditional on the results of a test should also be included unconditionally in the tree. Clinically unrealistic alternatives may be included in the decision analysis to verify that they are inadequate and to check or interpret the results of realistic ones.

Data
Parameter uncertainty
Presenting calculations using best guesses as input data is not sufficient for gaining an insight into the problem and the stability of the results. Uncertainty and misspecification can have a strong influence on the outcomes of strategies modelled. This uncertainty should be explored. Types of sensitivity analysis to assess the uncertainty of a model are:

- influence analysis
- threshold analysis
- full Bayesian analysis
- bayesian influence analysis
- attribute analysis
- generalisation analysis
- scenario analysis.

Consistency
Internal consistency
Symmetry: departures from symmetry should be considered carefully because they may indicate modelling errors. There are two types of symmetry to consider:

- physiological symmetry: the part of a decision tree that relates to a patient's underlying physiology should be symmetrical because actions cannot affect the possible states of nature that may occur
- complications symmetry: morbidity, mortality or financial costs that can accrue from one action in one context can also accrue from the same action in another context.

Halpern and colleagues[60]

Structure
Statement of decision problem/objective
The question that the model is to answer must be explicitly stated to ensure that the reason for developing the model is made clear. The disease(s) or condition(s) being modelled, interventions in question, specific study populations and study perspective need to be stated. Good practices cannot result unless there is a clear starting basis. Good practice requires the decisions made in the planning and development of a model to be appropriate; if these decisions are flawed, the value of the model will be greatly limited regardless of subsequent development efforts.

Statement of scope/perspective
Within the statement of the study question, the perspective of the study must be stated. This needs to be determined at the outset of the study. This refers to the limits or boundaries of a model: the time-frame, populations, comparators, setting, country or region, payment system, discount rate used (if applicable), index patients or disease characteristics. The scope will depend on the disease/condition and intervention under study. Much of the scope may be determined by the available data; alternatively, models may be developed to generate estimates of outcomes or other parameters (such as event probabilities) where none exists.

Model perspective also needs to be determined and clearly stated. Many believe that the society perspective should be used. However, the perspective should agree with the purpose of the model. The outcomes of the model will depend on

71

the study question but, in general, it is recommended that models include long-term or final outcomes. Models should include all relevant outcomes, including adverse events, relapse, death and loss to follow-up.

Justification of modelling approach
The modelling approach should be justified. It must be clear that, given the question to be answered, a model is the most (or only) appropriate approach. Models can be developed only when alternative interventions, outcomes and probabilities can be specified, or at least estimated.

Structural assumptions
Any potential biases and their likely direction should be recorded in a log. All suspected biases present in models, including those present by design, should be discussed and the impact of biases on modelling outcomes explored.

Consistency of structure
The scenario(s) or strategies to be modelled should be clinically relevant. The model must have clinical face validity, and the selected patient group must be relevant for the specific medical condition. This will usually involve clinical/specialist input. Policy makers may also make a valuable contribution, as they can provide assessment of the relevance of the study question within the social, political and budgetary environment. The steps, branches or health states in the model should follow the course of treatment or the progression of the disease, and progressions should make clinical and mathematical sense.

Strategies/comparators
All relevant treatment or intervention strategies must be assessed. The set of treatment alternatives should be based on a review of the literature and expert opinion. It is useful to include extreme strategies, even if they are not practical, as at least tests of the model's internal validity. The exploration of irrelevant or nonsensical treatment options has more potential to occur in modelling than in other study methodologies.

Model type
The appropriate model type will be determined by the intended use of the model (descriptive versus prescriptive, clinical decision model versus policy model, disease progression versus treatment comparisons), how probabilities are incorporated (deterministic, stochastic), the use of time (decision tree versus Markov) and the control of variability (heterogeneity). Models should involve the simplest time structure feasible for

appropriately assessing the modelled condition(s) and intervention(s).

Time-frame
The time-frame of the model should be stated within the scope. Longer term models will require discounting of costs and benefits. Discount rates should be stated within the model scope.

Parsimony
Models should only be as complex as required. Theoretically possible but non-observable branches should not be developed. Variables in the model should represent important clinical or patient factors. Variables not relevant to the condition should be excluded. Developing the model as a series of nodules, which can be evaluated individually, may enhance a model's simplicity. Necessary model complexity is likely to change over time, as more data become available. It is recommended that, when starting a model, one should err on the side of complexity and then simplify the model when appropriate or necessary.

Data
Data identification
General
In approaching the development of a model, the breadth and sources of information pertaining to the condition and/or intervention under study must be assessed. Limited availability of information should not be considered an insurmountable barrier to model development; models are most needed when little information is available from other sources. However, as the specifics of the model development will be influenced by available information, this must be explored early in the modelling process. It is important to begin with the most rigorous data available. Models should be based on the highest quality of data available. Whenever possible, data should come from epidemiological studies, RCTs or prospective naturalistic studies. Both the quality and relevance of the data must be evaluated.

Prospective data
If trial-based data are used, all relevant trials, rather than a subset of trials, should be used. If only a subset of trials is used in the model, justification for using a subset is required. Clinical trial data should not always be regarded as the gold standard for modelling solely on the basis of coming from a randomised controlled setting. RCTs are not always well designed, and problems in trial design can lead to spurious results. The use of small RCTs can lead to inaccurate model parameters and the generalisability of trial-based

results can be questioned. Therefore, the quality and relevance of trial-based data, as well as other studies, must be evaluated before use. The evaluation of prospective data should include:

- availability of data (are they published in a peer-reviewed journal?)
- sample size
- period (duration) and frequency of data collection
- degree of patient follow-up
- patient population characteristics
- methods of data collection.

Retrospective data
When using retrospective data, the completeness of a retrospective data set should be assured to ensure that the data were not derived from a potentially biased subset of patients or that specific data points are not systematically missing. The specific patient population comprising the retrospective data should be considered.

Expert opinion
The selection of experts to participate in model building is important. It should include individuals with substantial credibility in their fields that are likely to be opinion leaders. They should come from a variety of practice settings and geographical locations. The credibility and reputation of experts should be evaluated before inclusion. If individual data collection is being performed, a structured workbook, which includes existing information, and closed ended questions should be used. If group interviews are taking place, five to eight experts should be involved. Authors do not recommend extensive piloting of instruments to elicit expert opinion (validity of experts' opinions cannot be ascertained), but do recommend face validity tests to ensure clarity and completeness. To reduce biases in expert opinion, modified Delphi panels should be used. The nominal group technique may also be appropriate for certain information generated by clinical opinion.

Cost data
The quality of the values used for costs in terms of source and method of collection needs to be assessed. Standard national sources (e.g. MIMS) represent agreed-upon standards and are a useful basis for analysis.

Data modelling
Data synthesis
Appropriate methods should be used to combine data from disparate sources such as random effects meta-analysis. The technique should be carried

out in a rigorous and systematic fashion incorporating all relevant literature.

Discounting of costs and health effects to present values
If the model relates to a long period, discounting of costs and benefits to present values is required. The base-case model should have the same discount rate attached to cost and outcome. The effect of different discount rates for costs and benefits should be assessed in sensitivity analysis.

Analysis of trial data
When using information from RCTs, ITT results should always be used, as this will reflect the real world to a greater extent than on-therapy results. Judgement on whether ITT data reflect the real world needs to be made.

Utility estimates
The source of any utility weights, and the methods used to collect them, must be described. Utility weights should be collected from non-healthcare personnel, either patients or the informed public, based on the model and study constraints. The most appropriate methodology for collecting utilities will depend on the condition or intervention under study. In general, the standard gamble technique should be used.

Data incorporation
The use of deterministic models is recommended in most circumstances as the additional time, expense and complexity involved in stochastic modelling are not worth the gain in precision. In many cases it is reasonable to use only a point estimate and a range of values for the likelihood of a given modelled event. It may be difficult to determine the distribution of an event's likelihood, in which case a deterministic approach is appropriate. Moreover, probability distributions within a model may be correlated, in which case the specification of independent probability distributions would be inappropriate. The main advantage of a stochastic approach is that repeated sampling can be used to evaluate the uncertainty around modelled outcomes.

To evaluate the appropriateness of model parameters, the sources of all values should be listed. This will include event probabilities, healthcare resource utilisation (it is important to separate resource utilisation associated with different treatment pathways), utilities (including method of collecting utilities) and unit costs. The inclusion of indirect costs will depend on the condition being modelled, the perspective of the model and the intended audience. If indirect costs

73

are included in the model, the source and values used must be specified clearly.

Uncertainty
Parameter
Parameters with the greatest level of uncertainty (e.g. values from expert opinion), the greatest degree of variation, or greatest influence on the model outcomes should be included in sensitivity analysis. The sensitivity values for each parameter should be justified. Justification should be given for the choice of values used in sensitivity analysis. Multidimensional sensitivity analysis should be performed, but by only varying correlated parameters. This permits evaluation of a broader range of variability than unidimensional analysis, but provides easier interpretation than with independent modifications of uncorrected parameters.

Methodological
The effect of discounting should be assessed.

Heterogeneity versus variability
The inclusion of different costs is not a sensitivity analysis, but rather an adaptation of the model to a new setting.

Consistency
Internal consistency
Probabilities should sum to one. The model should exhibit symmetry (the modelled prognosis must be the same for the same condition/treatment combinations in different sections of the model). The distinction between decision and chance nodes is important. The model should make intuitive sense. Sensitivity analysis with extreme values can be used to assess whether the model is behaving appropriately. Any incongruities should be evaluated.

Between-model consistency
Comparing a model's results with those from similar models can validate a model.

External consistency
It is unlikely that data will exist to be able to compare the results of the model with actual data. It may be possible to compare intermediate outcomes from the model with actual data.

Hay and Jackson[22]

Structure
Statement of scope/perspective
Societal perspective is preferable, although there

is continued controversy over the relevance of the society perspective for some decision-makers. The use of a narrower perspective can be presented alongside the broader societal perspective.

Structural assumptions
In every study, several choices are required to fit the model to the research question. The choice of assumptions should be realistic and should reflect available data. The validity of the model will rest on whether its assumptions are reasonable in the light of the needs and the purposes of the decision-maker and whether, after close examination, its implications make sense.

Consistency of structure
A model should be shown to demonstrate face validity.

Strategies/comparators
A standard of care should be used as the appropriate comparator.

Model type
Rather than modelling repeated events as a separate branch in a complicated decision tree, state transition models may be more appropriate. Other alternative model structures include difference equations, deterministic models, stochastic models and discrete event simulations.

Time horizon
The time horizon of the model should be the duration for which a drug can be expected to impact meaningfully on a patient's health.

Data
Data identification
General
Wherever choices are made, conservative values of all parameters should be chosen, and the base case should represent the most plausible assumptions.

Costs
Decisions regarding cost estimates should be transparent and based on sound rationale.

Premodel data analysis
Methods for discounting costs and health effects to present values
In making discounting decisions, both costs and benefits should be discounted at the same rate.

Methods for analysing trial data
As far as possible, data should be analysed at the individual level for both costs and outcomes. Analysis of data from all study subjects is necessary

to interpret clinical trial data for pharmacoeconomic modelling. However, although ITT analysis is important, it is not the only way to analyse RCT data.

Uncertainty
Parameter
Parameter uncertainty is generally handled on a qualitative basis with either univariate or multivariate sensitivity analysis or maximum–minimum analysis, or quantified using statistical techniques such as the Delta method, joint confidence intervals, bootstrapped estimates or Monte-Carlo simulation. Although it is generally agreed that the application of multivariate sensitivity analysis is necessary, there is ongoing controversy over its value.

Structural
No proven method exists to validate structural uncertainties due to either parameter values assigned or the mathematical form in which the parameter values are combined, except to compute cost-effectiveness ratios for each alternative structural assumption and examine the appropriateness of results.

Consistency
Internal consistency
Each model should be shown to demonstrate face validity.

External validity
Wherever possible, models should be validated against other data sets.

Predictive validity
Each model should be shown to demonstrate predictive validity.

ISPOR Task Force[13]

Structure
Statement of scope/perspective
The model should be structured so that the inputs and outputs are relevant to the decision-making perspective (costs and consequences). Both costs and health consequences should reflect the decision-making perspective. If a perspective narrower than societal is used, there should be discussion, at least qualitatively, of the implications of broadening the perspective.

Consistency of structure
The structure of the model should be consistent with a coherent theory of the health condition.

This does not mean that all causal links need to have been proven, rather that links used should not be contraindicated by available evidence. The structure of the model should reflect the essential features of the disease and its interventions irrespective of data availability, but it is reasonable that data availability will affect choices regarding model structure. For example, a particular staging system that is commonly used in trials may be preferable to use to define health states than another that is less used but performs better in differentiating costs and/or health-related quality of life.

Structural assumptions
If evidence regarding structural assumptions is incomplete then limitations of the evidence supporting the chosen model structure should be acknowledged. If possible, sensitivity analysis using alternative model structures should be performed.

Health states
Health states may be defined to correspond to either the underlying disease process or observable health states, or a combination of both. However, in general, structural bias is avoided by modelling underlying disease status, and then by calibrating outputs to data on observable clinical status. States should not be omitted because of lack of data. Reasons to include additional subdivisions of health states may be based on their clinical importance, their relation to mortality or to quality of life, or cost. Disease states that may not be considered clinically important may well be important to include for these other reasons. Likewise, clinical states that do not affect the model's results may be included for face validity reasons.

Cycle length
The cycle length should be short enough to ensure that multiple changes cannot occur within a single cycle. The cycle length should be justified.

Parsimony
The structure of the model should be as simple as possible, while capturing the essentials of the disease process and interventions. If simplifications are made, these should be justified on the basis that they will not materially affect the results of the model.

Strategies/comparators
Options and direct strategies should not be limited to the availability of direct evidence from trials or what is current practice. There should be

a balance between including a broad range of feasible options and the need to keep the model manageable, interpretable and research based.

Heterogeneity
Modelled populations should be disaggregated according to strata, which have different event probabilities, quality of life or costs. This is particularly important when recurrent event rates over time are correlated within subpopulations.

Time horizon
The time horizon of the model should be long enough to reflect important and valued differences. Lifetime horizons are appropriate for most models. Shorter time horizons can be justified on the basis of no differences in survival or long-term chronic sequelae between options. A lack of long-term follow-up data should not be used as a rationale for failing to extend the time horizon of the model to the period relevant for the decision.

Data
Data identification
General
A model should not be faulted because existing data fall short of ideal standards of scientific rigour. Systematic reviews of literature should be conducted on key model inputs. Evidence that this has been done, or justification for why it has not been done, should be provided. If known data sources are excluded from parameter estimation, this should be justified. Data sources and results should not be rejected solely on the basis that they do not reach thresholds that define statistical significance. Results should be reported conditional on input estimates.

Expert opinion
Expert opinion is a legitimate source for parameters provided these parameters are shown not to affect the results or a sensitivity analysis is performed on these parameters with a clear statement that the results are conditional on the subjective judgements. If expert opinion is used, and the results are sensitive to the elicitations, then the methods for obtaining the values should be disclosed in detail. Formal methods such as Delphi or nominal group techniques are preferred.

Iterative data collection
A case should be made for reasonable opportunities to obtain new additional data to have been considered before modelling. Reasonable in this context means that the cost and delay inherent in obtaining additional data are

justified by the expected value of the new information in the analysis. Although formal methods of assessing the value of information exist, it is sufficient to give a heuristic argument as to why the current body of evidence was optimal.

Models should be repeatedly updated, and sometimes abandoned and replaced, as new evidence becomes available to inform their structure or input values. As a corollary, models that have been shown to be consistent with subsequent evidence, but have not been revised to calibrate against or incorporate this new evidence, should be abandoned until such recalibration has been accomplished.

Data modelling
General
Data modelling is the mathematical steps taken to transform empirical observations into a form that is useful for decision-makers. Data modelling assumptions should be disclosed and supported by evidence of their general acceptance and of their empirical validity. Key steps in developing the model should be documented. When alternative but equally defensible modelling approaches may lead to materially different results, sensitivity analysis should be performed to assess the implications of this. The base-case analysis should relate to assumptions where there is most support and alternative assumptions should be assessed in sensitivity analysis.

Data synthesis
Data modelling methods should follow generally accepted methods for biostatistics and epidemiology, for example meta-analysis.

Treatment effect
It is often appropriate to derive relative risks (or odds ratios) between treatment options in trials and superimpose these onto baseline probabilities derived from other sources (usually population based).

Transition probabilities
Interval probabilities from the literature need to be transformed into rates, and then into transition probabilities corresponding to the time interval (cycle time) used in the model. When transition rates or probabilities depend on events or states that may have been experienced in prior periods, this dependence should be modelled. This may be done by either incorporating a clinical/treatment history into the specification of the health states or including history as a covariate in specifying the transition probabilities.

Combining disease-specific and all-cause mortality
In general, it is acceptable to derive all-cause mortality from life tables unless an alternative source can be justified. It is not generally necessary to correct all-cause mortality for disease-specific mortality unless the disease represents a significant part of all-cause mortality.

Modelling survival
For example, exponential or Weibull forms may be used. The choice of survival function should be specified and justified. In general, all-cause mortality should be modelled non-parametrically from life tables.

Modelling risk factors
Evidence supporting additive or multiplicative effects of risk factors on baseline probabilities or rates of disease incidence or mortality should be sought.

Utility
It is preferable to use validated health-related quality of life instruments with prespecified scoring systems based on SG (Standard Gamble) or Time Trade Off (TTO). The methods for transforming health-state valuations into quality of life weights should be explicit.

Adjustment should be made for inflation or purchasing power across time and among countries. Adjustment for inflation should be based on the consumer price index, its healthcare components or one or more of its subcomponents, such as services or equipment. The choice between these will depend on whether the resource being costed is better represented by a general market basket or by a healthcare market basket. Purchaser power parities are the appropriate method for adjustments between countries.

Half-cycle correction
When appropriate, and if the difference in quality-adjusted survival is less than one cycle, the half-cycle correction should be applied.

Data incorporation
Measurement units, time intervals and population characteristics should be mutually consistent throughout the model. Specification of probability distributions for input parameters based on sampling uncertainty may be incorporated into formal probabilistic sensitivity analysis, but this is not always necessary or cost-effective. The preferred method of obtaining input distributions is to use posterior distributions from formal meta-analysis

and Bayesian analysis, but practical considerations may lead to the use of expert opinion. Ranges should always accompany estimates for which sensitivity analyses are performed. Either probabilistic (first order) simulation or deterministic simulation is acceptable. If first order Monte Carlo simulation is used, evidence should be provided that random simulation errors are smaller than the effect sizes of interest. Monte Carlo simulations should be carried out with fixed random number seeds to minimise random simulation error.

Uncertainty
Parameter
All modelling should include extensive sensitivity analysis of key parameters. Either deterministic or probabilistic sensitivity analysis is appropriate. Ranges should accompany base-case estimates of all input parameters for which sensitivity analyses are performed. The choice of parameters for sensitivity analysis is a matter of judgement by the analyst. If cohort simulation is used, sensitivity analysis can be performed using probabilistic sensitivity analysis. Care should be taken to ensure that interdependence between parameters is reflected in joint distributions.

Structural
If evidence regarding structural assumptions is incomplete, then limitations of the evidence supporting the chosen model structure should be acknowledged. If possible, sensitivity analysis using alternative model structures should be performed. A structural sensitivity analysis that uses a less aggregated model can provide reassurance that the simplifications do not materially affect the results. Recommendations for conducting or designing research investigations to guide future decision-making can be based on formal value of information or informal interpretation of the implications of sensitivity analysis.

Consistency
Internal consistency
Models should only be used after careful testing to ensure that the mathematical calculations are accurate and consistent with the specifications of the model. This process should include using null or extreme values to determine whether their use produces expected results. The results of the model should make sense and be explained on an intuitive level (face validity).

Between-model consistency
Models should be developed independently from one another to permit tests of between-model corroboration. The extent to which different

models of the same decision problem reach different conclusions should be explained (cross-validation).

External consistency
Models should ensure that their inputs and outputs are consistent with available data (calibration). Calibration data should be independent from data used to develop the model. However, models should not be criticised if independent calibration data do not exist, but should be open to criticism if independent data do exist and the model has not been calibrated with them, or the two have been compared and the model is not consistent with the data.

Predictive validity
Tests of predictive validity are valuable but not essential. A model should not be criticised for failing to predict the future. However, a good model should be amenable to respecification and recalibration as new data become available. It is not necessary for every data point or structural assumption in the model to be validated in prospective studies because results should be reported conditional on input data. The criteria for determining whether, and to what degree, tests of predictive validity are required before the use of the model depend on the benefits in terms of improving the model for decision-making and the costs of delaying the flow of information while obtaining the additional data.

McCabe and Dixon[10]

Structure
Statement of decision problem
The purpose of the model must be clearly specified, as this will have a fundamental impact on several aspects of the model, such as its perspective, the comparators, the complexity, data sources and outputs.

Justification of the modelling approach
A clear justification of the need for a model must be made, together with a justification of the approach taken (model type and structure). This will require indication of the lack of any alternative information or appraisal that demonstrates its weakness.

Structural assumptions
The absence of data is not in itself a justification for simplification. The structure of the model should be compared with other models and differences explained and justified. The simplicity

of the model should be justified by stating what simplifying assumptions have been made. An explanation of why such simplifications will not have material impact on the results of the model should also be given.

Consistency of structure
It is important that the possible pathways of the model are feasible and sensible. Current practice in the decision-making context to which the model is being applied should be capable of being described by the model's structure. Not all models require absolute detail of the clinical area; simpler structures may not affect the results. The intuitively appealing notion of descriptive validity cannot be described in a clear and unambiguous manner as it is inevitably linked to the purpose of the model and knowledge about the process being modelled.

Data
Data identification
General
The relative importance of one data source over another is not constant. Hierarchy of data is a secondary consideration to identifying a hierarchy of parameters within the model. A full systematic review is not required for each parameter. Even a fully systematic approach to data identification and full critical appraisal of each data source would not preclude the need for sensitivity analysis. The issue of how different data sources relate to each other must be given emphasis. Any trade-offs that are made in terms of, for example, appropriateness of study population versus study design should be identified explicitly. All data that are relevant for the model need to be appraised in terms of their appropriateness to the purpose of the model and their comparability with other data sources. The implications of choosing one combination of data over another may be explored empirically.

Data incorporation
The inputs of a model should be compared with those of existing models, and differences explained and justified. The justification of data sources should include assessment of the importance of the parameter, in addition to assessment of the importance of the value used. All data that are relevant to the model need to be appraised in terms of their appropriateness to the purpose of the model and their comparability with other data sources. The reasons for choosing certain data over others should be justified.

Consistency
Because a gold-standard test for validity of outcomes rarely exists, assessment of the structure

and inputs, as well as the outputs to the model, is required. Four different aspects of the modelling process should be assessed: the structure of the model, the inputs to the model, the results of the model and the value of the model to the decision-maker.

External consistency
Intermediate outputs of a model should be compared with existing data if available, and final outputs of a model should be compared with prospectively collected data if appropriate. Any differences should be explained and justified. The value of the model to the decision-maker can be assessed on whether:

- the model is appropriate for the decision-making context
- the model is understandable (its level of complexity needs to be tailored to its main audience)
- the model is believable.

Internal consistency
The computational correctness of the model needs to be assured.

Between-model consistency
The structure, data and results of a model should be compared with those of existing models, and differences explained and justified.

Predictive validity
It is important that the predictive validity of the model focuses on the modelled relationships between inputs and outputs.

Nuijten and colleagues[15]

Structure
Statement of decision problem/objective
The objective of the modelling study should be stated explicitly. The description of the objectives of a modelling study will be based on the hypothesis and does not usually differ from the objectives of a prospective pharmacoeconomic study.

Statement of scope/perspective
The analytical framework for the study should be described (i.e. CEA, CUA) and justified. The choice of perspective will depend on treatment patterns, the rules for reimbursement and country-specific pharmacoeconomic guidelines. The choice of perspective should be justified and the rationale explained. For example, data availability may limit the perspective of analysis.

The study population should be specified. This will be based on the clinical trial(s) study population; hence, the most relevant inclusion and exclusion criteria should be specified. Factors that may limit the applicability of the results to a wider population should be discussed. The healthcare setting to which the model relates should be specified and justified. Clinical events that do not differ statistically between intervention and control can be omitted from the model. Any limitation of the scope of the objective due to the availability of relevant data should be mentioned.

Justification of modelling approach
Within the introduction, there should be a description of the rationale for the study. This will include the information upon which the study hypothesis is based and justification for why a modelling study is required, including the design and methods used.

Structural assumptions
All assumptions that were used in structuring the model should be listed, including the process of creating the assumptions and the validation process.

Model type
Details should be given regarding the prognosis of the patient and how the disease progresses over time. This should be used to justify the type of model chosen and the characteristics of the model. The choice of type of model used should be justified.

Health states
Health states can be determined by the nature of disease progression or local treatment patterns, or both. The choice of health states should be justified.

Cycle time
Cycle time for Markov models should be determined by the nature of disease progression or local treatment patterns, or both.

Strategies/comparators
The strategies assessed within the model will depend on local treatment patterns (current treatment options, setting of treatment) and related clinical studies (e.g. surgical trials for the disease area). Relevant international and national guidelines may also be useful in determining strategies for the model. No therapy, prevention and screening options may also be relevant.

Time horizon
Justification should be given for the analytical

horizon. With Markov models, this is usually a lifetime perspective. For decision tree models, this is usually from onset of treatment until recovery or death.

Data
Data identification
General
The selection criteria used for studies and databases should be discussed and an indication of the potential of bias and likely direction of bias in the data sources should be given. Any search strategies used should be detailed.

Expert opinion
If expert opinion is used, the methods for eliciting this should be described in detail, and include inclusion criteria for experts and the process of eliciting opinions (e.g. number of rounds, interviews, questionnaires, postal survey). The use of Delphi panels can be justified on the basis that no other sources of data are available.

Data collection
The methods of data collection, including questionnaires or data abstraction forms, should be justified and presented in an appendix. The quality process for data entry should be described.

Premodel data analysis
The methods for determining (transition) probabilities should be described. The method used to adjust probabilities to the required cycle time of the model should be explained.

Methods for modelling survival
Assumptions and methods used to extrapolate short-term results to a longer time-frame should be given. There should be evidence that the methodology has been validated.

Data incorporation
The source of all data used in the model should be detailed. The source of all model variables should be described in sufficient detail for the reader to be aware of the type of data source. For example, where data come from RCTs, details of the results of the trial and the trial design, such as follow-up period and patient characteristics, should be given. The strengths and weaknesses and possible sources of bias inherent in each data source should be described. The methods used for determination of the range for each parameter in the model should be given and justified. The methods used for the determination of healthcare

utilisation (base case and range) should be given. For Markov models, the allocation of healthcare utilisation data to health states should be described, including any assumptions made.

Costs
The source for all cost estimates should be given.

Uncertainty
Parameter
The choice of variables upon which sensitivity analysis is performed should be justified. Sensitivity analysis should be performed for the clinical variables upon which the study hypothesis is based and on the main cost drivers.

Consistency
External consistency
The results of the model should be compared with other studies or expert opinion. Any observed differences should be addressed and explained. The impact of the level of uncertainty and limitations of the modelling study should be explained. The validation methodology and methods used for quality control should be described.

Between-model consistency
The results of the model should be compared with other studies. Any observed differences should be addressed and explained.

Ramsey[61]

This paper is directed at clinicians who may be faced with evaluating the results of a decision model for their own practice. It outlines the strengths and weaknesses of decision analysis before offering some guidelines for critical appraisal. Critical appraisal essentially answers three questions: Does the study apply to my patient or patient group? Are the methods used to address the problem appropriate? Do the results help me to improve the care of my patient or patient group compared with what I would have done before I read the article?

Structure
Statement of decision problem/objective
The clinical scenario should be defined clearly and precisely.

Statement of scope/perspective
The patient group that is the subject of the decision model should be well described and as specific as possible.

Consistency of structure
The decision pathway must be clearly displayed and justified. The outcomes following a chance node should include all that are possible for the patient. The total group of terminal nodes should be mutually exclusive and represent all major end-points related to the condition of interest.

Strategies/comparators
The options included in the model should be mutually exclusive and include all options that are typical for the situation and (if applicable) the new option of interest.

Data
Data identification
General
As with other areas of evidence-based medicine, the analysts must search the literature and select data from the highest quality studies. The process for gathering information to populate the model should be described explicitly.

Expert opinion
If expert opinion is used, this should be stated explicitly.

Data incorporation
Citations should be provided for all values used in the model. The analyst should justify the choice of value used in the analysis when one value is used over another.

Data modelling
Data synthesis
If pooled results are used, the pooling should be done using a recognised technique such as meta-analysis.

Utility
The reader should be satisfied that the utility estimates associated with outcomes are credible. Credibility can be judged by:

- the size of the sample from which the utility values are derived
- whether the utility weights match the reader's own clinical experience with similar patients
- whether the estimates are conservative (are they chosen to bias the outcome away from the new option?).

Methods for transforming charges to costs
Costs must be valued credibly. Costs should reflect the value of the resources rather than what is charged for those resources.

Uncertainty
Parameter
The reader must be aware of how uncertainty has been addressed in the model. This can be an issue of the structure of the model itself, or the values that populate the model. The value of sensitivity analysis can be assessed on:

- whether all clinically important variables were subjected to sensitivity analysis
- whether the best and worst cases met or exceeded the reader's own judgement on the range of possible values for the parameter
- whether more than one form of sensitivity analysis (one-way, two-way, probabilistic sensitivity analysis) was used
- whether tables or figures display the range of results for parameters where varying the results alters the outcome of the analysis.

Structural
The reader should be made aware of how structural uncertainties have been addressed in the model.

Consistency
Internal consistency
All probabilities should sum to one.

Sculpher and colleagues[7]

Structure
Statement of decision problem/objective
At the outset there should be a clear statement about the decision problem prompting the analysis. The structure of the model should be consistent with the stated decision problem.

Consistency of structure
The theory of disease, rather than the availability of data, should dictate the structure of the model. However, it is recognised that data availability will play some role in structuring and refining the scope of a model. For example, data may not be available to facilitate subgroup analysis, so the scope of the model may relate to a single homogeneous or 'average' group. If data are incomplete, or difficult to interpret, ignoring them cannot be justified. Models should combine all the information readily available at a particular point in time with current clinical or biological theory about the disease process (this should be reflected in the structure).

Model type
The analysts should choose the simplest model type that adequately reflects the time dependence of the disease events.

Health states

Disease states should be chosen to reflect the underlying biological process of the disease in question, and the impact of interventions, rather than health service inputs. Models should reflect accepted theory/clinical classifications of disease. The number of health states should be manageable, but sensitive to changes in the underlying disease. States should not be omitted because of lack of data. Any choices that are made should be justified. The number and definition of health states is a trade-off between descriptive realism and simplicity. Analytical methods such as cluster analysis may be helpful in justifying the choices made, although there are no agreed standards. The disease process should be understood in terms of the symptoms reported by patients.

Strategies/comparators

Options and strategies should not be limited by constraints of current practice, but a balance should be struck between modelling the full range of feasible options (given the context and perspective of the model) and time and resource constraints. The assumptions and compromises that are necessary to satisfy these constraints must be transparent and justified. The inclusion of extreme strategies is useful to serve as an anchor point for other strategies and to assess consistency.

Time horizon

The time horizon of the model should be sufficient to indicate when cost and effect differences between options are stable, but it is recognised that this is an output of the model rather than an a priori specification. Lifetime time horizons will be appropriate in most longitudinal models, but shorter horizons can be justified for some disease processes and interventions.

Cycle length

The cycle length should be the minimum interval over which pathology and symptoms in patients is expected to alter. The analyst should justify the chosen cycle length. This contrasts with a choice that is based on clinical reviews, which is inappropriate because finding the optimal length of follow-up or review should be one aspect of the options evaluated by the model. Similarly, cycle length should not be determined by the availability of data. Cycle length should be driven by what is known about the underlying disease process.

Parsimony

The model should be as simple as possible while encapsulating important factors such as time and the characteristics of the disease.

Data
Data identification
General

A model should not be criticised because of a dearth of data or because existing data fall short of scientific rigour. Best available data should be referred to as optimum available data, as it is an empirical question about whether acquiring all existing evidence to determine the 'best available data' is a good use of resources. Models can be used to undertake formal value of information analysis to determine the optimal data to incorporate.

Iterative data collection

The number of available iterations of the model will help to identify whether the model is based on optimum available data. If a series of models has been developed to address a particular question, with each model incorporating further primary or secondary evidence deemed cost-effective to collect and incorporate by an earlier version of the model, the data inputs are justified by virtue of the modelling process. However, in current practice, there is rarely a series of models; most results relate to a single iteration of a model. When a model is based on a single iteration, or for the first iteration of the model, the modellers should make clear that all sources of information that are available at relatively low cost in terms of researcher time have been searched for using the most appropriate parameter values. The model should be updated as information and theory change.

Expert opinion

In the context of no data being identified, the methods used to elicit expert opinion should be fully detailed. This may include:

- inclusion criteria and sample size
- descriptions of what the experts were asked and the format for the questions
- the avoidance of an attempt to forge a group consensus.

Data modelling
General

The general process of data incorporation should follow accepted methods of epidemiology and statistics such as meta-analysis.

Incorporation of transition probabilities

Interval rates from the literature should be translated into transition probabilities appropriately.

Half-cycle correction

The model should include a half-cycle correction.

Data incorporation

One way of reflecting parameter uncertainty is to make the parameters stochastic; that is, to incorporate a distribution rather than a point estimate. If a stochastic analysis is performed, the analyst should justify the distributions used.

Uncertainty

Heterogeneity

Different sources of uncertainty should be distinguished, especially sample uncertainty versus heterogeneity.

Parameter

If stochastic analysis is undertaken, the distributions for parameter values, which should reflect second order uncertainty, should be justified.

Value of information

Models can be used to undertake formal value of information analysis to determine the optimal data to incorporate.

Consistency

Internal consistency

The model should be checked and tested by the analyst during the modelling process to identify errors relating to model syntax and data incorporation. Simple tests include the movement of results in relation to sensitivity analysis and whether they confirm a priori expectations, or rebuilding the model in a second software package and ensuring that the results are the same.

External consistency

Intermediate outputs, such as the time in particular states, from the model should be checked with available data that have not been used to populate the model.

Between-model consistency

The results from previously developed models may be useful. Any discrepancies should be explained and justified. As different models are developed at different points in time, and other data may have been collected for other purposes, non-convergence of these sources should not be regarded as a reason to reject the model.

Sendi and colleagues[16]

Consistency

Internal consistency

Technical validity

This assesses the technical accuracy of the model.

One approach to ensuring technical validity is to double-implement the model in separate software. For example, a model may be developed in Excel and replicated in DATA. Any differences in model outputs indicate an error in one of the models. The source of the error should be identified and rectified. Sensitivity analysis, using extreme values, is another method of determining the technical validity of the model as it can validate the behaviour of the model. Face validity involves comparing the assumptions of what should be happening within the model to what would be expected; that is, does the modelled medical intervention produce an output that one might expect?

Predictive validity

Comparing intermediate and final outcomes of the model with observed data can assess predictive validity. The model should be assessed against independent (i.e. not used to develop the model) but comparable data. If the model does not agree with observed data, then either the input variables are faulty or the model has an invalid structure. One option is repeatedly to take a subset of the data, generate model results from this subset and compare these results with the remainder of the dataset.

Between-model consistency

This involves comparison of the model with results from other models independently addressing the same question. Any differences in results and conclusions should be justified.

Sonnenberg and colleagues[21]

Structure

Heterogeneity

To the extent possible, a decision model should reflect differences between alternative patient groups in terms of treatment and management pathways and disease progression. Reviewers of models should ensure that conclusions are stratified to reflect groups of patients that are meaningful in a clinical context.

Scope of decision model

If data are available, all factors that may be affected by a decision should be taken into account in the model. For example, alteration in the selection criteria for organ transplant may affect the waiting time for organs. If these effects can be ascertained, they should be included in the model. The perspective of the model should be considered.

83

Strategies/comparators

A decision model should include all reasonable options that are likely to be available to a clinician. Strategies included in the model should include a watchful waiting option, in addition to a 'do nothing' option, if this is a clinically viable option. The model should also include extreme strategies. Even if these are clinically unrealistic, they serve as anchor points against which all other strategies can be measured and also provide a means of assessing the fidelity of the model.

Time horizon

The model's time horizon should match that of the actual process being considered in the analysis.

Model type

The choice of model type should be appropriate for the problem. It should reflect the time dependence of events being modelled. Options are:

- a simple tree: here any event of interest may happen only once and at some prespecified time; the timing of events can only be reflected by the outcome measure
- recursive and Markov models: these are required when exposure to a risk is continuous, when the timing of an event is uncertain or when the timing of occurrence of an event varies and affects the outcome. When the number of cycles is more than ten, a cycle tree is infeasible and a Markov model is preferred.

Consistency of structure

Authors describe a conceptual framework to explain the relationship between the decision problem and the resulting model. The framework consists of four levels.

- Biological truth: this is not usually directly or completely knowable.
- Theoretical model: this represents understanding of the biological truth, analogous to scientific theory that best explains the facts known at a given point in time. It normally amounts to enumeration of important choices and events and the relationships between them.
- Practical model: this is the most detailed model that can be constructed on the basis of the data we have. This practical model may also be constrained by a variety of simplifying assumptions, reflecting not only limitations of the data, but also limitations to the size and complexity of the model.
- Implementation model: this represents the programming of the practical model in the relevant computer software.

The move from the theoretical model to the practical model will involve assessment of available data and the requirement of simplifying assumptions to limit the size and complexity of the model. The detail of the implementation model should be no more than supported by available data. All important clinical outcomes (e.g. complications) should be represented. The model should include key variables to represent important factors in the patient population.

Structural assumptions

Decision models should be documented carefully so that a reviewer can determine how the original model was constructed. This documentation should include:

- all simplifying assumptions
- sources of probabilities
- relationships among parameters
- mathematical formulae (such as Bayes' rule or life expectancy calculations)
- the software package used to construct the model.

Data

Data modelling

Utility

The utility structure should contain all of the relevant attributes.

Data incorporation

The source of probabilities used in the model should be documented. Most decision models are deterministic, that is, they incorporate a point estimate for each parameter. However, there is always uncertainty associated with this value. This has traditionally been addressed by n-way sensitivity analysis. Another type of uncertainty is statistical variation in the value of a parameter with a known mean. This uncertainty is best represented using a probability distribution. Both quantity and quality of life should be incorporated into the model.

Uncertainty

Parameter

Uncertainty due to statistical variation is best represented by using a probability distribution rather than a point estimate. Monte Carlo analysis can be used to assess the impact of this known variation on results, although the interpretation of these types of result in the context of decision analysis is not clear.

Consistency
Internal consistency
The practical model should behave as the theoretical model would suggest (face validity).

Common errors
Errors include:

- invalid model syntax
- conditioning of an action on an unobservable event
- violations of symmetry (all strategies must model prognosis in the same way, i.e. take account of the same variables)
- failure to link variables
- failure to apply consistent bias
- incorrectly modelling the results of a diagnostic test
- incorrectly modelling a treatment.

These errors can be used to form a set of rules from which the model can be critiqued from the perspective of the structure and programming of the model. Non-structural considerations include choice of model type, enumeration of strategies and events and choice of utility models (to take account or not to take account of utility). The model can be verified by determining its success in representing the practical model. This is largely achieved by performing and interpreting sensitivity analysis. The theoretical model should predict how the model should behave; if opposite behaviour is observed the model may be faulty or a new insight is derived.

Predictive validity
Validity refers to the ability of the decision model to recommend optimal decisions. Short of a clinical trial of a decision model, the validity of the recommended decision cannot be assessed because there is no gold standard for the quality of the decision. It would be desirable for the reviewers of decision analyses to have the original model available for inspection.

Between-model consistency
The validity of the model can be assessed indirectly by comparing its predictions with other published studies.

Soto[19]

Structure
Statement of decision problem/objective
The study question addressed by the model, along with reasons for its development, including the goal and objectives of the study, should be stated. They should be clear, transparent, relevant and achievable according to available data.

Statement of scope/perspective
Ideally, a model should provide a societal perspective. However, this viewpoint should be disintegrated into multiple viewpoints, including that of the primary decision-maker to whom the study is targeted.

Consistency of structure
Clinicians and decision-makers should be involved in the development of the study question to ensure its clinical relevance. The data sources for the development of the structure of the model should be given. The model should correspond as much as possible to the real-life situation of the disease in each setting.

Justification of modelling approach
There should be substantive evidence to specify why, given the study question, a modelling approach is justified. It should be clear that modelling is the most (or only) practical approach to the problem.

Model type
The type of model chosen should be justified. This will depend on the time-frame of the events being modelled and the complexity of the interaction of various consequences of each decision option. Models should involve the simplest time structure feasible for appropriately assessing the modelled disease and therapeutic options together with enough treatment complexity to have face validity. A disease simulation model should be used when details of the history of the patient are necessary to determine prognosis.

Structural assumptions
All assumptions that have been used in building the model should be listed. The validity of the model relies on the reasonableness of its assumptions. All assumptions should be logical and reflect daily medical practice in the country chosen. The limitations and biases of assumptions should be stated. An expert advisory board can be used to validate the assumptions used in the model. The outcome(s) chosen should show effectiveness rather than efficacy. In general, they should be long-term or final outcomes. The choice of outcome should be justified.

Health states
There should be a description of the health states along with a justification for their use.

Cycle time

The cycle time should be described and justified. The validation process to define the cycle length should be described.

Time horizon

This should extend far enough into the future to capture the major clinical and economic outcomes of the alternatives under assessment. The selection of period will depend on the nature of the clinical problem.

Strategies/comparators

Information should be given regarding the study options and other relevant treatments. A reasonable set of treatment alternatives should be based on a review of published literature, clinical trials and expert opinion. Local treatment patterns and practice guidelines will affect choices made. The characteristics of all comparators in the study should be given (efficacy, adverse events and compliance).

Heterogeneity

If there is evidence that results may be different for different subgroups, a separate analysis should be conducted to reflect each subgroup.

Data

Data identification

General

The estimation of costs and outcomes should be obtained from the best designed and least biased source that is relevant to the question and the population under study. Potential validity issues and biases inherent in prospective and retrospective data sources should be recognised. The validity and accuracy of data sources used should be assessed using recognised rules to assess the quality and reliability of experimental and observational studies. The use of case reports is not recommended, as these are likely to contain bias.

Outcomes

Outcomes data can come from high-quality meta-analyses. It is important to review the methodology of the meta-analysis used in the model.

Expert opinion

Expert opinion is a valid source of data when there are few or no published data available. When using expert opinion, the following steps should be taken.

- Parameter estimates should be based on majority, not individual, opinions.

- Explicit inclusion criteria should be defined to select the experts. It is important to include experts with substantial credibility in their field as well as those from a variety of practice settings and geographical locations.
- Details of methodology used should be stated clearly, along with the process of the study (questionnaires, number of iterations, Delphi panels, etc.).
- The experts chosen should be relevant to the population under consideration. A list of participants in the expert group should be provided.
- A definition of a consensus should be determined in advance of the execution of the study.
- The possible weaknesses of the information gained, and methods used to minimise these, should be explained.
- Validation of the data would be desirable.

All details for the elicitation of expert opinion should be made available on request.

Resource use/costs

The methods used to determine units of healthcare utilisation should be described. If retrospective data are used, it is important to ascertain that the patient group matches patients in general so that the data are generalisable. The costs included in the model will depend on the disease being modelled, the perspective of the model and the intended audience. To obtain information regarding resource use, real clinical practice rather than a clinical trial should be considered.

Data incorporation

All data sources for parameters used in the model should be described in full.

Data modelling

Modelling survival

Methods used to extrapolate trial-based outcomes to obtain life expectancy should be described.

Utility

Methods used to obtain utility weights should be described clearly.

Uncertainty

Parameter

Sensitivity analysis should be carried out to assess the robustness of the conclusions when the values of variables that are highly uncertain are altered. The choice of variables for sensitivity analysis should be justified and the rationale for the

interpretation of the results of the sensitivity analysis should be defined clearly. It is advisable to use best and worst case scenarios, as well as confidence intervals. A practical approach to sensitivity analysis is to perform univariate analysis in key values in the model to ascertain under which circumstances the variable has an important impact on the results. The most appropriate method for handling uncertainty of multiple parameters is to undertake multivariate sensitivity analysis.

Confidence intervals should be calculated for the incremental cost-effectiveness ratios obtained.

Consistency

Internal consistency
All probabilities should be between zero and one and the probabilities for consequences should sum to one.

External consistency
A validation of the model about face and predictive validity should be undertaken.

Between-model validation
Any model can be partially validated by comparing its results with previously created models.

Predictive validity
A validation of predictive validity should be undertaken. The best way to validate a model is to carry out a prospective study to ascertain that the results are similar to the model.

Weinstein and colleagues[20]

Structure

Structural assumptions
The assumptions, input parameters, and the logic connecting them to outputs should be stated with complete clarity. The model should be transparent and open to peer review.

Data

Identification
The model should be calibrated: inputs to the model should be consistent with available data.

Uncertainty
A balance must exist between the costs and consequences of obtaining and waiting for better

data and the costs and consequences of permitting a synthesis of the available evidence to influence decisions. The EVOI (expected value of information) method is directly applicable to this.

Consistency
The way we evaluate models should be consistent with the purpose of the model. Cost-effectiveness models are meant to be aids to decision-making.

Internal consistency
Models should only be used after testing to ensure internal accuracy. A model is often subjected to extreme conditions to test whether the outputs are as expected.

External consistency
The model should be calibrated: inputs to the model, and outputs of the model, should be consistent with available independent data. The process of calibrating a model is likely to be iterative. The calibration process should be explicit. The results of the model should make sense in relation to the theoretical considerations underpinning the model and they should be able to be explained in intuitive terms (face validity). The model should have been subjected to peer review by a dispassionate reviewer and found to be what it claims to be.

Between-model consistency
Modellers should be able to explain why their model reaches different conclusions to other models of the same disease and question which use the same input data (convergent validity).

Predictive validity
Tests for predictive validity (concordance between actual events and model outputs) are valuable but not necessarily essential. If a model is intended to aid decision-makers and not necessarily to predict the future, then it is inappropriate to expect predictive validity. The criteria for determining whether tests for predictive validity are required depends on the benefits in terms of improving the model for decision-making and the costs of delaying the flow of information while obtaining additional data. The paper advocates the use of value of information methodology. It is inappropriate to expect that a model can predict the future accurately, because it can only incorporate the data available at the time when the model is brought to bear on the decision.

Appendix 3

Quality assessment in decision-analytic models:
a suggested checklist

The following table provides a checklist for the critical appraisal of decision-analytic models developed for health technology assessment. The format is taken from the work by Sculpher and colleagues.[7]

Dimension of quality		Attributes of good practice	Questions for critical appraisal
Structure			
S1	Statement of decision problem/objective	There should be a clear statement of the decision problem prompting the analysis	Is there a clear statement of the decision problem?
		The objective of the evaluation and of the model should be defined	Is the objective of the evaluation and model specified and consistent with the stated decision problem?
		The primary decision-maker should be stated clearly	Is the primary decision-maker specified?
S2	Statement of scope/perspective	The perspective of the model (relevant costs and consequences) should be stated clearly, and the model inputs should be consistent with the stated perspective and overall objective of the model	Is the perspective of the model stated clearly?
			Are the model inputs consistent with the stated perspective?
		The scope of the decision model should be specified and justified	Has the scope of the model been stated and justified?
		The outcomes of the model should reflect the perspective and scope of the model and should be consistent with the objective of the evaluation	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?
S3	Rationale for structure	The structure of the model should be consistent with a coherent theory of the health condition under evaluation, and the treatment pathways (disease states or branches) should be chosen to reflect the underlying biological process of the disease in question and the impact of the intervention. The structure should not be dictated by current patterns of service provision	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?
			Are the sources of data used to develop the structure of the model specified?
		All sources of evidence used to develop and inform the structure of the model (i.e. the theory of disease) should be described. The structure should be consistent with this evidence	Are the causal relationships described by the model structure justified appropriately?
S4	Structural assumptions	All structural assumptions should be transparent and justified. They should be reasonable in the light of the needs and purposes of the decision-maker.	Are the structural assumptions transparent and justified?
			Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?
S5	Strategies/comparators	There should be a clear definition of the options under evaluation	Is there a clear definition of the options under evaluation?
		All feasible and practical options relating to the stated decision problem should be evaluated. Options should not be constrained by the immediate concerns of the decision-maker or data availability, or limited to current clinical practice	Have all feasible and practical options been evaluated?
			Is there justification for the exclusion of feasible options?
S6	Model type	The appropriate model type will be dictated by the stated decision problem and the choices made regarding the causal relationships within the model	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?

Dimension of quality		Attributes of good practice	Questions for critical appraisal
S7	Time horizon	A model's time horizon should extend far enough into the future for it to reflect important differences between options	Is the time horizon of the model sufficient to reflect all important differences between options?
		It is important to distinguish between the time horizon of the model, the duration of treatment and the duration of treatment effect	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?
S8	Disease states/pathways	Disease states/pathways should reflect the underlying biological process of the disease in question and the impact of interventions	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?
S9	Cycle length	For discrete time models, the cycle length should be dictated by the natural history of disease. It should be the minimum interval over which the pathology or symptoms are expected to alter	Is the cycle length defined and justified in terms of the natural history of disease?
Data			
D1	Data identification	Methods for identifying data should be transparent and it should be clear that the data identified are appropriate given the objectives of the model	Are the data identification methods transparent and appropriate given the objectives of the model?
		There should be justification of any choices that have been made about which specific data inputs are included in a model	Where choices have been made between data sources, are these justified appropriately?
		It should be clear that particular attention has been paid to identifying data for those parameters to which the results of the model are particularly sensitive	Has particular attention been paid to identifying data for the important parameters in the model?
		Where expert opinion has been used to estimate particular parameters, sources and methods of elicitation should be described	Has the quality of the data been assessed appropriately?
			Where expert opinion has been used, are the methods described and justified?
D2	Data modelling	All data modelling methodology should be described and based on justifiable statistical and epidemiological methods. Specific issues to consider include those listed under D2a–d, below	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?
D2a	Baseline data	Baseline probabilities may be based on natural history data derived from epidemiological/observational studies or relate to the control group of an experimental study	Is the choice of baseline data described and justified?
		Rates and interval probabilities should be transformed into transition probabilities appropriately. If there is evidence that time is an important factor in the calculation of transition probabilities in state transition models, this should be incorporated	Are transition probabilities calculated appropriately?
		If a half-cycle correction has not been used on all transitions in state transition model (costs and outcomes), this should be justified	Has a half-cycle correction been applied to both cost and outcome?
			If not, has this omission been justified?

92

Dimension of quality	Attributes of good practice	Questions for critical appraisal
D2b Treatment effects	Relative treatment effects derived from trial data should be synthesised using recognised meta-analytic techniques	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?
	The methods and assumptions that are used to extrapolate short-term results to final outcomes should be documented and justified. This should include justification of the choice of survival function (e.g. exponential or Weibull forms). Alternative assumptions should be explored through sensitivity analysis	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified? Have alternative assumptions been explored through sensitivity analysis?
	Assumptions regarding the continuing effect of treatment once treatment is complete should be documented and justified. If evidence regarding the long-term effect of treatment is lacking, alternative assumptions should be explored through sensitivity analysis	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified? Have alternative assumptions been explored through sensitivity analysis?
D2c Costs	Costing and discounting methods should accord with standard guidelines for economic evaluation	Are the costs incorporated into the model justified? Has the source for all costs been described? Have discount rates been described and justified given the target decision-maker?
D2d Quality of life weights (utilities)	Utilities incorporated into the model should be appropriate for the specified decision problem	Are the utilities incorporated into the model appropriate? Is the source for the utility weights referenced? Are the methods of derivation for the utility weights justified?
D3 Data incorporation	All data incorporated into the model should be described and the sources of all data should be given and reported in sufficient detail to allow the reader to be aware of the type of data that have been incorporated	Have all data incorporated into the model been described and referenced in sufficient detail?
	Where data are not mutually consistent in the model, the choices and assumptions that have been made should be explicit and justified	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?
	The process of data incorporation should be transparent. It should be clear whether data are incorporated as a point estimate or as a distribution. If data have been incorporated as distributions as part of probabilistic analysis, the choice of distribution and its parameters should be described and justified	Is the process of data incorporation transparent? If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified? If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?
D4 Assessment of uncertainty	In assessing uncertainty, modellers should distinguish between the four principal types of uncertainty	Have the four principal types of uncertainty been addressed? If not, has the omission of particular forms of uncertainty been justified?
D4a Methodological	Methodological uncertainty relates to whether particular analytical steps taken in the analysis are the most appropriate	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?

Dimension of quality	Attributes of good practice	Questions for critical appraisal
D4b Structural	There should be evidence that structural uncertainties have been evaluated using sensitivity analysis	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?
D4c Heterogeneity	It is important to distinguish between uncertainty resulting from the process of sampling from a population and variability due to heterogeneity (i.e. systematic differences between patient subgroups)	Has heterogeneity been dealt with by running the model separately for different subgroups?
D4d Parameter	Where data have been incorporated into the model as point estimates, the ranges used for sensitivity analysis should be stated and justified	Are the methods of assessment of parameter uncertainty appropriate?
	Probabilistic analysis is the most appropriate method of handling parameter uncertainty because it facilitates assessment of the joint effect of uncertainty over all parameters (see 'Data incorporation', p. 92)	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?
Consistency		
C1 Internal consistency	There should be evidence that the internal consistency of the model has been evaluated in terms of its mathematical logic	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?
C2 External consistency	The results of a model should be explicable. Either results should make intuitive sense or counterintuitive results should be fully explained	Are any counterintuitive results from the model explained and justified?
	All relevant available data should be incorporated into a model. Data should not be withheld for purposes of assessing external consistency	If the model has been calibrated against independent data, have any differences been explained and justified?
	The results of a model should be compared with those of previous models and any differences should be explained	Have the results of the model been compared with those of previous models and any differences in results explained?

Appendix 4

Application of the checklist to three
decision models

(a) Reviewer 1

Payne and colleagues' LBC model[28]

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
S1	Is there a clear statement of the decision problem?	✓	The decision problem is to identify the likely impact of LBC screening techniques, in terms of incidence of cervical cancer and associated mortality, and in terms of costs and cost-effectiveness when compared with conventional smear testing for a typical UK population
	Is the objective of the evaluation and model specified and consistent with the stated decision problem?	✓	The question is specified clearly as "what would be the likely impact of the new liquid based cytology screening techniques, in terms of incidence of cervical cancer and associated mortality, and in terms of costs and cost-effectiveness when compared with conventional smear testing for a typical UK population?" The objective of the model is to simulate the life experience of a cohort of women aged 18–95 years with the alternative screening options
	Is the primary decision maker specified?	✗	None specified, although the report is for NICE
S2	Is the perspective of the model stated clearly?	✓	Health service perspective
	Are the model inputs consistent with the stated perspective?	✓	Inputs relate to the progression of cervical disease and associated cost, disease and non-disease-related mortality and life expectancy
	Has the scope of the model been stated and justified?	✗	Not specified explicitly, but most of the scope is detailed within the objective and decision problem
	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	✓	Outcomes relate to life-years
S3	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	✓	The model is based on three elements: a disease progression model depicting transitions through histologically recognised stages of precancer, the effect of screening on the natural history and a life table to reflect age-specific all-cause mortality
	Are the sources of data used to develop the structure of the model specified?	✓	The structure of the model is based on a previously published UK model of cervical screening
	Are the causal relationships described by the model structure justified appropriately?	✓	The structure of the model is taken directly from a previously published model, which provides justification for the causal relationships modelled. There is no reference to the primary clinical source as this was done in the original model

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
S4	Are the structural assumptions transparent and justified?	✗	Regression only occurs from CIN I. The effect of this assumption is not explored. Constant incidence of precancer, disease progression and fast growing cancers between 18 and 64 despite evidence to the contrary. The general effect of these assumptions is described. The reason why these assumptions are made is not clear, but it is likely to be because the original model from which this is based made these assumptions. Constant mortality risk from invasive cancer. This assumption is justified by referencing the primary source. The assumptions regarding the effect of LBC on test sensitivity are clear, but are not justified appropriately
	Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	✓	Given that this is a reproduction of a published model for the specific purpose of evaluating LBC in the UK
S5	Is there a clear definition of the options under evaluation?	✓	LBC compared with conventional cytology and no screening for 2-, 3- and 5-year screening intervals
	Have all feasible and practical options been evaluated?	✓	Given the specific question addressed
	Is there justification for the exclusion of feasible options?	✓	Annual screening is not assessed. Although this is potentially feasible, because the model suggested that biannual screening was too expensive, annual screening was probably rejected on this basis
S6	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	✓	Markov model of disease progression
S7	Is the time horizon of the model sufficient to reflect all important differences between options?	✓	Lifetime time horizon
	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?	✓	Lifetime time horizon, screening up to 64, treatment effect NA
S8	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	✓	Disease states of precancer and cancer are used. Screening options modelled as screen test results
S9	Is the cycle length defined and justified in terms of the natural history of disease?	✗	Cycle length is set at 6 months, presumably because this model is based on an already published model, which applies a 6-month cycle time

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
D1	Are the data identification methods transparent and appropriate given the objectives of the model?	✓	Data for the baseline model are taken directly from a previously published model. Data for LBC effects/costs and assumptions are taken from review work carried out
	Where choices have been made between data sources, are these justified appropriately?	✗	The authors are not clear why they have chosen to reproduce the Sherlaw–Johnson model as opposed to others that are available. Choices are made regarding cost assumptions, but these are not justified explicitly
	Has particular attention been paid to identifying data for the important parameters in the model?	✗	Data for the baseline model are derived directly from a previous model. The authors state that there is insufficient evidence for the sensitivity improvement associated with LBC, so make an assumption for the base-case model. The basis of this assumption is unclear. The effect of doubling/halving the progression rate, alternative sensitivity/specificity assumptions for LBC, the marginal cost of LBC, and test adequacy are shown by running the model with varying assumptions
	Has the quality of the data been assessed appropriately?	✓	Authors make assessments of cost data incorporated and remark on caveats of previous model from which this model is based
	Where expert opinion has been used, are the methods described and justified?	NA	
D2	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?	NA	No data modelling methodology are reported. Incorporated data are already in a form ready for use
D2a	Is the choice of baseline data described and justified?	✓/✗	Based on data from a previously published model, although there is no justification for using this model
	Are transition probabilities calculated appropriately?	?	Difficult to say as they are taken directly from a previous model
	Has a half-cycle correction been applied to both cost and outcome?	?	Not detailed
	If not, has this omission been justified?	NA	
D2b	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	NA	The relative effects associated with LBC are based largely on assumptions. The effects of these assumptions are evaluated in sensitivity analysis
	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified?	NA	
	Have alternative extrapolation assumptions been explored through sensitivity analysis?	NA	

Quality criterion	Question(s)	Response (✓, X or NA)	Comments
	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified?	NA	
	Have alternative assumptions regarding the continuing effect of treatment been explored through sensitivity analysis?	NA	
D2c	Are the costs incorporated into the model justified?	✓	Direct costs of screening, diagnosis and treatment are included and are taken from published sources of cervical screening/treatment costs (primary studies and Health Resource Group-based costs). All costs are up rated to 1999 values
	Has the source for all costs been described?	✓	Details of the original cost derivation (the primary source from where data are taken) are given. Costs associated with LBC are based largely on assumption but the effect of this assumption is described.
	Have discount rates been described and justified given the target decision-maker?	✓	Costs at 6%, life-years at 1.5% in accordance with NICE recommendations
D2d	Are the utilities incorporated into the model appropriate?	NA	The outcome relates to life-years
	Is the source for the utility weights referenced?	NA	
	Are the methods of derivation for the utility weights justified?	NA	
D3	Have all data incorporated into the model been described and referenced in sufficient detail?	✓	All data are referenced and described
	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	NA	No mutually inconsistent data
	Is the process of data incorporation transparent?	✓	Data incorporated as triangular distributions using the maximum and minimum range as end-points
	If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?	✓	Justification given is that triangular distributions were used for simplicity
	If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?	✓	Data based on range estimates for the mean

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
D4	Have the four principal types of uncertainty been addressed?	✗	Structural uncertainty not addressed
	If not, has the omission of particular forms of uncertainty been justified?	✗	No justification given explicitly
D4a	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	✓	The effect of alternative discount rates has been addressed
D4b	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	✗	See above
D4c	Has heterogeneity been dealt with by running the model separately for different subgroups?	NA	Model applied to a general UK population of women
D4d	Are the methods of assessment of parameter uncertainty appropriate?	✓	One-way and two-way sensitivity analysis along with probabilistic sensitivity analysis
	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?	NA	All range data are reported and are incorporated as triangular distributions. Cost-effectiveness acceptability frontiers presented as an appendix along with value of information analysis. Marginal cost of LBC and the specificity of the LBC test were found to account for the majority of opportunity loss
C1	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	✓	Before use, the model predictions were compared with those on which the model is based
C2	Are any counterintuitive results from the model explained and justified?	NA	
	If the model has been calibrated against independent data, have any differences been explained and justified?	✓	Incidence data (with 3-year screening) from the model are compared with cancer registry data: the model does not show the double-peaked incidence pattern that is shown with real data but this is because of the constant incidence assumption in the model. Test characteristics from the model are compared with actual data and differences are explained
	Have the results of the model been compared with those of previous models and any differences in results explained?	✓	Comparisons are made with the results of the original model. The conclusion is that this model reflects that of the earlier model. Age-specific incidence (without screening) is compared with data from a US model: differences are explained and justified

NA, not applicable.

Kendrick and Johnson's MS model[29]

Quality criterion	Question(s)	Response (✓, X or NA)	Comments
S1	Is there a clear statement of the decision problem?	X	It could be: Is long-term use of IFN-β cost-effective? Or to estimate the progression of relapsing–remitting MS patients receiving IFN-β or standard care. It is not clear to which type of patients the model relates
	Is the objective of the evaluation and model specified and consistent with the stated decision problem?	X	Although it can be inferred to relate to the evaluation of assumptions in previous cost-effectiveness models of MS/IFN-β and the effect of relaxing the clinical effectiveness assumption on cost-effectiveness
	Is the primary decision-maker specified?	X	None stated
S2	Is the perspective of the model stated clearly?	X	Although the model presents results that include health service costs only separately from costs that include a broader society perspective
	Are the model inputs consistent with the stated perspective?	NA	Perspective not stated
	Has the scope of the model been stated and justified?	X	Not stated, although costs include MS-associated costs only
	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	NA	The perspective and scope are not specified clearly. QALYs used, but there is no mention of mortality (disease and non-disease related)
S3	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	?	The model predicts the progression of EDSS scores over time as a proxy for the progression of MS. The use of EDSS is justified through comment that it is widely used in clinical trials of MS treatments. The EDSS is an ordinal scale from 0 (full health) to 10 (death). The model assumes a constant progression of EDSS scores over time
	Are the sources of data used to develop the structure of the model specified?	✓/X	The use of EDSS scores as a proxy for MS progression is referenced, as is the progression rate over time, although the constant progression assumption is not justified or appropriate given the ordinal nature of the scale
	Are the causal relationships described by the model structure justified appropriately?	X	The regression model was estimated from data from a trial (MSCRG trial), although details of the trial or parameters of the model are not given
S4	Are the structural assumptions transparent and justified?	X	The assumptions of treatment effect are transparent but are not justified. The model assumptions (i.e. parameters and the relationship between them) are not given
	Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	?	Cannot make any further assessment of this with the information given in the paper

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
S5	Is there a clear definition of the options under evaluation?	✓	The options under evaluation relate to the duration of treatment (2, 5, 10 and 20 years)
	Have all feasible and practical options been evaluated?	?	The duration of treatment options appears to be arbitrarily chosen. Results for no use of IFN-β are not reported, although the predicted progression of EDSS over 20 years for no treatment with IFN-β is presented in a figure
	Is there justification for the exclusion of feasible options?	✗	None given
S6	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	✗	The discussion of the disease process includes a description of the relapse–remitting nature of MS. This does not appear to be taken into account in this model. The model assumes a linear and continuous relationship between treatment and EDSS score over time. There is no justification for this
S7	Is the time horizon of the model sufficient to reflect all important differences between options?	?	The difference between options in the model relates only to the progression of EDSS score, and costs and consequences are specified according to EDSS score
	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?	✓/✗	The time horizon of the model is 20 years; duration of treatment is 2, 5, 10 and 20 years. Treatment effect is assumed to continue while the patient is on treatment. No justification is provided
S8	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	?	The model relates to the progression of EDSS over time. The parameters of the model are not specified
S9	Is the cycle length defined and justified in terms of the natural history of disease?	NA	
D1	Are the data identification methods transparent and appropriate given the objectives of the model?	✗	The data identification methods are not specified
	Where choices have been made between data sources, are these justified appropriately?	✓	The authors describe how a particular set of data was used because it matched the patient group of the MSCRG trial
	Has particular attention been paid to identifying data for the important parameters in the model?	✗	The crucial parameter of the model is the progression of EDSS with and without treatment. The data used to estimate this came from a trial, but no details are given to justify the use of these trial data
	Has the quality of the data been assessed appropriately?	✗	Not specified
	Where expert opinion has been used, are the methods described and justified?	NA	

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
D2	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?	✗	The use of linear regression to model progression of patients through an ordinal measurement scale, and to extrapolate 2-year data to progression over 20 years is questionable
D2a	Is the choice of baseline data described and justified?	✗	The standard treatment and on-treatment progressions come from trial data. Details of the trial are not given
	Are transition probabilities calculated appropriately?	NA	
	Has a half-cycle correction been applied to both cost and outcome?	NA	
	If not, has this omission been justified?	NA	
D2b	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	NA	
	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified?	✗	The extrapolation of 2-year data beyond this period is documented (a linear extrapolation is assumed) but not justified
	Have alternative extrapolation assumptions been explored through sensitivity analysis?	✗	No sensitivity analysis is performed
	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified?	✗	Trial data are reported that are used to justify the assumptions regarding the continued effect of treatment once treatment is complete. However, there is no justification for the assumptions beyond the 2-year period (duration of the trial)
	Have alternative assumptions regarding the continuing effect of treatment been explored through sensitivity analysis?	✗	No sensitivity analysis is carried out
D2c	Are the costs incorporated into the model justified?	✓	The costs incorporated into the model come from a previous economic evaluation of MS in the UK. The costs relate directly to EDSS score
	Has the source for all costs been described?	✓	The source is referenced, but the study from which the costs are derived is not described, except that the study approximated that of the MSCRG trial (which is where the data for the progression equations come from)
	Have discount rates been described and justified given the target decision-maker?	✓	

continued

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
D2d	Are the utilities incorporated into the model appropriate?	?	The utilities used in the model were based on a regression analysis of EDSS score and utility weight derived from a previous UK study. As the regression suggested a linear relationship with a slope of 0.09, every 1-point increase in EDSS score was assumed to equate to a 10% decrement in utility weight. Since EDSS is an ordinal scale and utility weights are on a ratio scale, this assumption is questionable
	Is the source for the utility weights referenced?	✓	Original utility scores were taken from a published study and manipulated for use in this model
	Are the methods of derivation for the utility weights justified?	✗	See above
D3	Have all data incorporated into the model been described and referenced in sufficient detail?	✓/✗	All incorporated data are referenced but not described in sufficient detail
	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	NA	As the data used in the model are not described, it is unclear whether mutually consistent or mutually inconsistent data have been used
	Is the process of data incorporation transparent?	✗	No details of the progression model are given
	If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?	NA	
	If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?	NA	
D4	Have the four principal types of uncertainty been addressed?	✗	Uncertainty is not addressed
	If not, has the omission of particular forms of uncertainty been justified?	NA	
D4a	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	✗	See above
D4b	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	✗	See above
D4c	Has heterogeneity been dealt with by running the model separately for different subgroups?	✗	See above

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
D4d	Are the methods of assessment of parameter uncertainty appropriate?	✗	See above
	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?	✗	See above
C1	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	✗	No consistency checks are reported
C2	Are any counterintuitive results from the model explained and justified?	✗	
	If the model has been calibrated against independent data, have any differences been explained and justified?	✗	See above
	Have the results of the model been compared with those of previous models and any differences in results explained?	✓	The differences are explained in terms of the assumption regarding treatment effect beyond 2 years

IFN-β, β-interferon; MSCRG, Multiple Sclerosis Collaborative Research Group.

Turner and colleagues' influenza prevention and treatment models[30]

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
S1	Is there a clear statement of the decision problem?	✓	The decision problem is not stated explicitly in the modelling section of the report. However, it was defined earlier in the main body as being to identify optimal prevention and treatment strategies for influenza and, in particular, the role that neuroaminidase inhibitors have to play. Specific questions addressed by the review are given
	Is the objective of the evaluation and model specified and consistent with the stated decision problem?	✓	To compare options in terms of cost-effectiveness for the treatment and prophylaxis of influenza (A and B) in four separate patient groups: healthy adults, at-risk adults, children and residential care elderly. Options under evaluation are also given
	Is the primary decision-maker specified?	✗	Although the report is written for HTA/NICE
S2	Is the perspective of the model stated clearly?	✓	Perspective of the NHS. The impact of reduced time away from work is assessed in sensitivity analysis

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	Are the model inputs consistent with the stated perspective?	✓	Health service costs and QALYs
	Has the scope of the model been stated and justified?	✓	Within the objective, we are given details of the patient groups and options under evaluation. The adverse events taken into account are given, along with justification for not including others
	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	✓	Treatment models: Outcomes in terms of QALYs and cost per flu illness day avoided. Length of illness determined from control arms of trials or Cochrane review data. Outcomes used in the prophylaxis model are generated from the n- treatment part of the treatment model
S3	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	✓	Seems to be consistent with a treatment pathway. The treatment pathway is not justified. Likely to be based on systematic reviews conducted earlier in report. The progression of disease is not modelled: not applicable
	Are the sources of data used to develop the structure of the model specified?	✗	No details are given regarding the development of the structure of the model. Perhaps it is based on treatment pathways observed in trials?
	Are the causal relationships described by the model structure justified appropriately?	✗	
S4	Are the structural assumptions transparent and justified?	✓	A list of all assumptions is provided along with the likely effect of bias
	Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	✓	Seem to be, although without knowledge of the area it is difficult to say
S5	Is there a clear definition of the options under evaluation?	✓/✗	All options are specified clearly, but there is little detail regarding dosage and model of administration
	Have all feasible and practical options been evaluated?	?	Cannot say without further knowledge of the area
	Is there justification for the exclusion of feasible options?	NA	
S6	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	✓	Tree model represents probabilities of events (probability of flu, presentation to the GP, presentation within 48 hours of onset of symptoms, of drug treatment if presenting before or after 48 hours, probability of severe complications and resultant death). Seems that the model examines a single treatment episode or prevention round. Repeat prophylaxis/treatment over time is not considered. Cannot judge the relevance of this without prior knowledge of the disease

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
S7	Is the time horizon of the model sufficient to reflect all important differences between options?	✗	Cannot say without further knowledge of the area. The treatment model length is based on the length of the trials in the area. May be a question about whether the time-frame is sufficient to take account of all complications
	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?	✗	I think the treatment model is based on 21 and 7 days only although it is not very clear in the report. The duration of treatment (6 weeks) and duration of treatment effect (none once treatment stops) are described for prophylaxis
S8	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	✓	Decision tree represents the pathways patients may follow when they have symptoms of flu
S9	Is the cycle length defined and justified in terms of the natural history of disease?	NA	
D1	Are the data identification methods transparent and appropriate given the objectives of the model?	✓	Effectiveness data come from full systematic review of the trial evidence and Cochrane reviews. Data for other parameters in the model are based on reviews of the literature and meta-analyses (not specified whether systematic), but search criteria are not specified
	Where choices have been made between data sources, are these justified appropriately?	✓	All data from trials used for outcomes. All probability estimates are referenced. A relatively old Cochrane review was updated with new evidence to reflect the current use of vaccines.
	Has particular attention been paid to identifying data for the important parameters in the model?	✗	Main treatment effects are derived from trial data; probabilities are derived from reviews of evidence. A key uncertainty was found to be associated with the QALY values associated with influenza and adverse events. QALY weights for adverse events are based on expert opinion; no further research was carried out to reduce the uncertainty about this estimate.
	Has the quality of the data been assessed appropriately?	✓	Presumably an assessment of quality was undertaken for the systematic review
	Where expert opinion has been used, are the methods described and justified?	✗	Expert opinion was used to identify EQ-5D scores for adverse events. No details are given regarding the expert panel or methods
D2	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?	✓	Random effects meta-analysis used to estimate time to symptoms across all trials of neuroaminidase inhibitors. Meta-regression used to estimate the relationship between time to symptoms alleviated and time to fever in days
D2a	Is the choice of baseline data described and justified?	✓	All data derived from control group of trials
	Are transition probabilities calculated appropriately?	NA	
	Has a half-cycle correction been applied to both cost and outcome?	NA	

108

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	If not, has this omission been justified?	NA	
D2b	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	✓	Random effects meta-analysis
	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified?	✓	Length of flu illness for amantadine was extrapolated using meta-regression techniques based on the observed relationship between length of fever and length of illness
	Have alternative extrapolation assumptions been explored through sensitivity analysis?	✓	Different time-frames for extrapolation explored (7 days, 21 days)
	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified?	✓	For the prophylaxis model no effect was assumed once treatment ceased
	Have alternative assumptions regarding the continuing effect of treatment been explored through sensitivity analysis?	✗	
D2c	Are the costs incorporated into the model justified?	✓	All come from standard sources
	Has the source for all costs been described?	✓	All cost sources referenced appropriately
	Have discount rates been described and justified given the target decision-maker?	✓	1.5% benefits, 6% costs
D2d	Are the utilities incorporated into the model appropriate?	✓	Come directly from trial evidence or expert opinion
	Is the source for the utility weights referenced?	✓	Trials referenced
	Are the methods of derivation for the utility weights justified?		QALY weights were taken from a Likert scale administered in the trials, which were manipulated to mean visual analogue scale scores from the Measurement and Validation of Health (MVH) study. These were then converted into time trade off estimates using an equation. No data existed for children, therefore adults weights were applied. The effect of using 7 days of data instead of 21 days was examined through sensitivity analysis
D3	Have all data incorporated into the model been described and referenced in sufficient detail?	✓	All sources are referenced
	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	✓	US data used to determine attack rates on the basis that over time attack rates in the USA and Europe would be the same

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	Is the process of data incorporation transparent?	✓	Data incorporated as distributions and as point estimates (two versions of the base-case model)
	If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?	✓	All distributions are described for each parameter but not justified. However, the choice does reflect the properties of the parameter
	If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?	✓	Monte Carlo simulation used to reflect second order uncertainty
D4	Have the four principal types of uncertainty been addressed?	✓	Although this is not explicitly stated.
	If not, has the omission of particular forms of uncertainty been justified?	NA	
D4a	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	✓	Effect of discount rates assessed
D4b	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	✓	The effect of structural assumptions is assessed. For example, data used to derive the length of illness varied from using 21 days' worth of data to using just 7 days
D4c	Has heterogeneity been dealt with by running the model separately for different subgroups?	✓	The model was run for four groups
D4d	Are the methods of assessment of parameter uncertainty appropriate?	✓	Probabilistic sensitivity analysis plus a series of one-way sensitivity analyses
	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?	✓	Mean value used for deterministic analysis
C1	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	✗	
C2	Are any counterintuitive results from the model explained and justified?	NA	The model results do not appear to be counterintuitive
	If the model has been calibrated against independent data, have any differences been explained and justified?	NA	
	Have the results of the model been compared with those of previous models and any differences in results explained?	✓	The results are compared with three other models. The reasons for differences between them are explained clearly

(b) Reviewer 2
Payne and colleagues' LBC model[28]

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
S1	Is there a clear statement of the decision problem?	✓	Questions addressed by the model were stated
	Is the objective of the evaluation and model specified and consistent with the stated decision problem?	✓	Objective not specified, but questions to be answered by the model stated
	Is the primary decision-maker specified?	✗	Not stated, but given HTA report is this not obvious?
S2	Is the perspective of the model stated clearly?	✓	Health service perspective for costs
	Are the model inputs consistent with the stated perspective?	✓	Direct healthcare costs included
	Has the scope of the model been stated and justified?	✗	Scope not defined
	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	✗/✓	Scope not defined, but is consistent with objectives and perspective of the analysis
S3	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	✓	States seem to be based on sound knowledge of the disease area
	Are the sources of data used to develop the structure of the model specified?	✗	Very little information on the model structure. No diagram to explain the structure included. Refer to original model, but without prior knowledge of this model it is difficult to gain all the details of the revised model
	Are the causal relationships described by the model structure justified appropriately?	✗	
S4	Are the structural assumptions transparent and justified?	✗	Six-month cycles for transitions not justified. List of assumptions (including data assumptions) provides some details about structural assumptions
	Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	✓	Given that the model aimed to look at cost-effectiveness of screening for a typical population is it justified in presuming no difference in incidence for ages 18–64 (although this is unlikely to be true)
S5	Is there a clear definition of the options under evaluation?	✓	Some detail on the strategies compared

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	Have all feasible and practical options been evaluated?	✓	Two-year screen dominated. One-year screen would be even less cost-effective, so excluded as a comparator
	Is there justification for the exclusion of feasible options?	✓	
S6	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	✓	State transition model used to simulate disease process
S7	Is the time horizon of the model sufficient to reflect all important differences between options?	✓	Lifetime horizon. Model assesses cost-effectiveness in terms of cost per cancer avoided and cost per life-year gained
	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?	✓	No treatment modelled
S8	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	✓	See S3 on states included
S9	Is the cycle length defined and justified in terms of the natural history of disease?	✗	No justification for cycle length
D1	Are the data identification methods transparent and appropriate given the objectives of the model?	✗	Systematic review not undertaken for model parameters
	Where choices have been made between data sources, are these justified appropriately?	✗	No details reported
	Has particular attention been paid to identifying data for the important parameters in the model?	✗	See above
	Has the quality of the data been assessed appropriately?	✗	No mention of study quality
	Where expert opinion has been used, are the methods described and justified?	NA	
D2	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?	NA	

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
D2a	Is the choice of baseline data described and justified?	✓/✗	Baseline data described but not justified. Reference to previous model
	Are transition probabilities calculated appropriately?	?	
	Are transition probabilities calculated appropriately?	?	No details on how transition probabilities calculated
	Has a half-cycle correction been applied to both cost and outcome?		Not mentioned
	If not, has this omission been justified?	✗	
D2b	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	✗	Methods of data extraction not described. Reference to previous model
	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified?	NA	
	Have alternative extrapolation assumptions been explored through sensitivity analysis?	NA	
	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified?	NA	
	Have alternative assumptions regarding the continuing effect of treatment been explored through sensitivity analysis?	NA	
D2c	Are the costs incorporated into the model justified?	✓	
	Has the source for all costs been described?	✓	
	Have discount rates been described and justified given the target decision-maker?	✓	Chosen according to NICE recommendations
D2d	Are the utilities incorporated into the model appropriate?	NA	Authors state that there is insufficient quality of life information available to estimate a cost per QALY
	Is the source for the utility weights referenced?	NA	
	Are the methods of derivation for the utility weights justified?	NA	
D3	Have all data incorporated into the model been described and referenced in sufficient detail?	✓	Data sources are described and referenced

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	✗	
	Is the process of data incorporation transparent?	✓	Triangular distributions used for simplicity
	If data have been incorporated as distributions, has the choice of distribution for each parameter been justified?	✓	Ranges defined in a data table and referenced, but choices not justified
	If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?	✓	
	If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?	✓	
D4	Have the four principal types of uncertainty been addressed?	✗	Structure and methodological uncertainty not addressed
	If not, has the omission of particular forms of uncertainty been justified?	✗	
D4a	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	✓	Alternative discount rates in sensitivity analysis
D4b	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	✗	
D4c	Has heterogeneity been dealt with by running the model separately for different subgroups?	✗	Model ran for women 18–64 years. Age ranges not considered
D4d	Are the methods of assessment of parameter uncertainty appropriate?	✓	Presume probabilistic analysis, although not made explicit
	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?	NA	
C1	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	✓	Model validation. Comparison with earlier model and published data

113

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
C2	Are any counterintuitive results from the model explained and justified?	NA	
	If the model has been calibrated against independent data, have any differences been explained and justified?	✓	
	Have the results of the model been compared with those of previous models and any differences in results explained?	✓	Some validation with previously published model

Kendrick and Johnson's MS model[29]

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
S1	Is there a clear statement of the decision problem?	✗	Lengthy description of why literature relating to IFN-β needs to be reanalysed, but no clear objective of the model
	Is the objective of the evaluation and model specified and consistent with the stated decision problem?	NA	Decision problem not stated
	Is the primary decision-maker specified?	✗	
S2	Is the perspective of the model stated clearly?	✓	Health service and societal perspective for costs
	Are the model inputs consistent with the stated perspective?	✓	Types of costs included in each perspective are consistent
	Has the scope of the model been stated and justified?	✗	Scope not defined
	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	✗	
S3	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	?	Structure not described in detail

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	Are the sources of data used to develop the structure of the model specified?	✗	
	Are the causal relationships described by the model structure justified appropriately?	✗	Structure not described in detail
S4	Are the structural assumptions transparent and justified?	✗	
	Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	NA	No statement of objective, perspective or scope
S5	Is there a clear definition of the options under evaluation?	✓	IFN-β versus no IFN-β (standard care)
	Have all feasible and practical options been evaluated?	✓	Difficult to tell. May be alternatives to IFN-β that are not considered here
	Is there justification for the exclusion of feasible options?	✗	
S6	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	?	No real definition of the model type
S7	Is the time horizon of the model sufficient to reflect all important differences between options?	✗	20-year time horizon used. Paper suggests that after 12 years cost savings using IFN-β may be realised; therefore, a lifetime time horizon may have been the most appropriate
	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?	✓	States that beneficial effect of IFN-β may not be realised until after 12 years
S8	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	NA	Not state transition or decision model
S9	Is the cycle length defined and justified in terms of the natural history of disease?	NA	

continued

Quality criterion	Question(s)	Response ($\checkmark$, $\times$ or **NA**)	Comments
D1	Are the data identification methods transparent and appropriate given the objectives of the model?	$\times$	
	Where choices have been made between data sources, are these justified appropriately?	$\checkmark$	No definition of population in the MSCRG trial
	Has particular attention been paid to identifying data for the important parameters in the model?	$\times$	
	Has the quality of the data been assessed appropriately?	$\times$	No quality assessment
	Where expert opinion has been used, are the methods described and justified?	NA	
D2	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?	$\times$	Simple linear model used to relate EDSS scores to progression of MS
D2a	Is the choice of baseline data described and justified?	$\times$	MSCRG trial used. Very few details of this trial given
	Are transition probabilities calculated appropriately?	NA	
	Has a half-cycle correction been applied to both cost and outcome?	NA	
	If not, has this omission been justified?	NA	
D2b	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	NA	
	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified?	$\times$	Methods used to extrapolate are not described in sufficient detail
	Have alternative extrapolation assumptions been explored through sensitivity analysis?	$\times$	Not described
	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified?	$\times$	Some discussion of competing evidence on the long-term effects of IFN-β, but no justification for length of treatment effects used

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	Have alternative assumptions regarding the continuing effect of treatment been explored through sensitivity analysis?	✗	See above
D2c	Are the costs incorporated into the model justified?	✓	Murphy et al. used for costs as the study population approximated the population used in the MSCRG trial, which was used for disease progression data
	Has the source for all costs been described?	✓	Costs used in the model were from a study by Murphy et al.[62] This study was not described in detail
	Have discount rates been described and justified given the target decision-maker?	✓	6% used for costs and benefits, according to the UK Treasury rate
D2d	Are the utilities incorporated into the model appropriate?	✓	Utilities applied to stages of MS progression
	Is the source for the utility weights referenced?	✓	
	Are the methods of derivation for the utility weights justified?	✗	Regression analysis using utility values of a previous cost–utility analysis, used to predict utility values for EDSS scores from the MSCRG trial. Simple linear model used for non-ordinal data (EDSS)
D3	Have all data incorporated into the model been described and referenced in sufficient detail?	✓	
	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	✓	Some discussion of alternative estimates of treatment effects
	Is the process of data incorporation transparent?	✗	
	If data have been incorporated as distributions, has the choice of distribution for each parameter been justified?	NA	
	If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?	NA	
D4	Have the four principal types of uncertainty been addressed?	✗	Uncertainty was not addressed at all
	If not, has the omission of particular forms of uncertainty been justified?	NA	

Quality criterion	Question(s)	Response (✓, X or NA)	Comments
D4a	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	NA	
D4b	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	NA	
D4c	Has heterogeneity been dealt with by running the model separately for different subgroups?	X	Model does include three stages of disease
D4d	Are the methods of assessment of parameter uncertainty appropriate?	NA	
	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?	NA	Sensitivity analysis not conducted
C1	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	X	
C2	Are any counterintuitive results from the model explained and justified?	X	
	If the model has been calibrated against independent data, have any differences been explained and justified?	X	
	Have the results of the model been compared with those of previous models and any differences in results explained?	✓	Comparison with other models using different time horizons

Turner and colleagues' influenza prevention and treatment models[30]

Quality criterion	Question(s)	Response (✓, X or NA)	Comments
S1	Is there a clear statement of the decision problem?	✓	Objective stated separately to cost-effectiveness section of the report
	Is the objective of the evaluation and model specified and consistent with the stated decision problem?	✓	

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	Is the primary decision-maker specified?	✗	Presume NHS as HTA report
S2	Is the perspective of the model stated clearly?	✓	NHS perspective. Time away from work assessed in sensitivity analysis
	Are the model inputs consistent with the stated perspective?	✓	See above
	Has the scope of the model been stated and justified?	✓	Describes the patient groups the model will be run for and the alternative strategies evaluated
	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	✓	Incremental cost per QALY gained and cost per influenza illness day avoided
S3	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	✗	
	Are the sources of data used to develop the structure of the model specified?	✓	
	Are the causal relationships described by the model structure justified appropriately?	✗	
S4	Are the structural assumptions transparent and justified?	✓	Structural assumptions described but not justified in sufficient detail
	Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	✓	
S5	Is there a clear definition of the options under evaluation?	✓/✗	States alternative strategies but no detail on dosage or methods of administration. Introduction to the model does not make it clear how many strategies are considered by the model
	Have all feasible and practical options been evaluated?	✓	Would seem to be, but difficult given lack of knowledge of disease area
	Is there justification for the exclusion of feasible options?	NA	
S6	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	✓	Appropriate model, but could have used Markov type model where person has a probability of developing flu each period (year)

continued

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
S7	Is the time horizon of the model sufficient to reflect all important differences between options?	✗	Maximum is 21-day model. Do not consider any long-term effects of vaccination or influenza
	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?	✗	Do not distinguish between time horizon and length of treatment effect
S8	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	✓	
S9	Is the cycle length defined and justified in terms of the natural history of disease?	NA	
D1	Are the data identification methods transparent and appropriate given the objectives of the model?	✓	Data identified from systematic review
	Where choices have been made between data sources, are these justified appropriately?	✓	Cochrane data regarding the effectiveness of vaccine were updated using more recent data to reflect vaccines that are currently in use
	Has particular attention been paid to identifying data for the important parameters in the model?	✗	
	Has the quality of the data been assessed appropriately?	✓	Quality of effectiveness data assessed as part of systematic review element of HTA report. Quality of studies used not discussed in economics sections
	Where expert opinion has been used, are the methods described and justified?	✗	Expert opinion used to obtain EQ-5D scores for adverse events. Methods are not described
D2	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?	✓	
D2a	Is the choice of baseline data described and justified?	✓	Described but not justified
	Are transition probabilities calculated appropriately?	NA	
	Has a half-cycle correction been applied to both cost and outcome?	NA	
	If not, has this omission been justified?	NA	

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
D2b	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	✓	Meta-analysis used to combine estimates of treatment effect and adverse effects
	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified?	✓	
	Have alternative extrapolation assumptions been explored through sensitivity analysis?	✓	
	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified?	✓	
	Have alternative assumptions regarding the continuing effect of treatment been explored through sensitivity analysis?	✗	
D2c	Are the costs incorporated into the model justified?	✓	Relevant costs given the perspective assumed
	Has the source for all costs been described?	✓	
	Have discount rates been described and justified given the target decision-maker?	✓	NICE recommended rates used
D2d	Are the utilities incorporated into the model appropriate?	✓	
	Is the source for the utility weights referenced?	✓	Utilities were estimated from numerous trials of oseltamivir
	Are the methods of derivation for the utility weights justified?	✓	Visual analogue scale scores transformed into time trade off-type scores using a published regression equation
D3	Have all data incorporated into the model been described and referenced in sufficient detail?	✓	The types of data used and the part they played in the modelling process was described very well
	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	✓	
	Is the process of data incorporation transparent?	✓	See above
	If data have been incorporated as distributions, has the choice of distribution for each parameter been justified?	✓	

Quality criterion	Question(s)	Response (✓, ✗ or NA)	Comments
	If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?	✓	Described but not justified
	If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?	✓	Monte Carlo simulation used to reflect second order uncertainty in data
D4	Have the four principal types of uncertainty been addressed?	✗	Although not made explicit
	If not, has the omission of particular forms of uncertainty been justified?	✗	
D4a	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	✗	
D4b	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	✓	Uncertainty regarding the length of treatment effect assessed
D4c	Has heterogeneity been dealt with by running the model separately for different sub-groups?	✓	Model ran for four groups of patients: healthy adults, at-risk adults, children and residential care elderly
D4d	Are the methods of assessment of parameter uncertainty appropriate?	✓	Probabilistic analysis and a series of one-way sensitivity analyses
	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?	✓	Taken from relevant data source
C1	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	✗	No internal validation of the model
C2	Are any counterintuitive results from the model explained and justified?	NA	
	If the model has been calibrated against independent data, have any differences been explained and justified?	NA	
	Have the results of the model been compared with those of previous models and any differences in results explained?	✓	Compares results with three previous models

Appendix 5

Members of the Expert Advisory Group

Pelham Barton (Health Service Management Centre, University of Birmingham)
Colin Green (National Coordinating Centre for Health Technology Assessment, University of Southampton)
Luke Vale (Health Service Research Unit, University of Aberdeen)
Chris McCabe (School of Health and Related Research, University of Sheffield)

Suzy Paisley (School of Health and Related Research, University of Sheffield)
Alec Miners (Technology Appraisals Team, National Institute for Clinical Excellence)
Steve Palmer (Centre for Health Economics, University of York)
Elisabeth Fenwick (Centre for Health Economics, University of York)

Appendix 6

Comments from the Expert Advisory Group and amended checklist

(a) General comments on the report from the EAG

Comment	Response
Report must state if this is good modelling practice for NICE or more general guidance	Made more explicit in methods
Results section must state that the 15 papers cover general guidance and specific guidelines covering model quality	Changed in methods section
Recognise that the searches were restricted to health technology assessment and there is a wealth of literature of good modelling outside health technology assessment (Weinstein paper reviewed this in an unsystematic way). Papers relating to specific disease areas were also excluded	Changed in methods section
Provide a summary of the guidance synthesis process	Added to report
Recognise the checklist is not a substitute for critical appraisal. It is just a template that may be used during, and as part of, the critical appraisal process	Comments added to conclusions regarding the use of the checklist in Chapter 3
It would be useful to have general comments on the model quality alongside the review – this would reduce the desire for people simply to add up the number of ticks (it is up to the reviewer to come up with textual assessments)	Comments added to conclusions regarding the use of the checklist in Chapter 3

(b) Comments on the guidelines and accompanying checklist from the EAG

Comment	Response
Explain what justification means in this context	Added to first mention of justification in checklist and guidance
Don't like the term data modelling – consider changing	Changed to premodel data analysis
Add something about possible competing theories in the structure (S3) section	Added
S5 is contradictory to the notion of complexity described in the synthesised guidance	Amended to reflect description in the guidance
S5 needs a health warning. Must be used with general economic evaluation guidelines	Added
Consider adding something about presentation of results	Want to avoid duplicating what general economics evaluation guidelines say. Issue of presentation of results not included in checklist or guidance
Change biological process to biological/clinical process in S8 attributes of good practice	Added

Comment	Response
Be clear that in terms of data identification we are advocating systematic and not necessarily comprehensive searches	Term 'systematic' added
Add something to S9 about what constitutes a key model parameter, e.g. 'Process of selecting key parameters must be justified and systematic methods used to identify the most appropriate data'	Added
Add justifiable to D2b	Added
Explain what we mean by non-linear modelling	Explained in brackets in guidance
Add 'A decent model should explore alternative assumptions regarding extrapolation' to D2c attributes of good practice	Text changed in checklist
Add something more explicit about lifetime in S7, to correspond with synthesised guidelines	Has been amended to correspond with guidance recommendation of lifetime horizon
Consider dropping D2c, as this is more general to economic evaluation	Dropped
Add something about threshold analysis for extreme scenarios in D4b attributes of good practice	Added
Add the question 'Has probabilistic analysis been done, if not has this been justified' to D4b	Added
Internal consistency may be tested by two people building the model in the same package and checking if spreadsheet and model match	Methods to establish internal validity not made explicit in the checklist; however, this is likely to be a relevant option
Add questions 'Are the conclusions valid, given the data presented' to C2	Added
Make a distinction between excluding relationships that you wouldn't logically expect to differ and those that don't differ according to statistical significance	Already explicit in guidance. Added to checklist
State that the above is part of an iterative process, you may therefore add something that doesn't differ now but may do in the future (when more data become available)	Added to checklist and guidance
Predictive validity by McCabe concerns the relationships between model inputs and outcomes – check interpretation in synthesised guidance	Amended

(c) Amended checklist

Dimension of quality	Attributes of good practice	Questions for critical appraisal
Structure		
S1 Statement of decision problem/objective	There should be a clear statement of the decision problem prompting the analysis	Is there a clear statement of the decision problem?
	The objective of the evaluation and of the model should be defined	Is the objective of the evaluation and model specified and consistent with the stated decision problem?
	The primary decision-maker should be stated clearly	Is the primary decision-maker specified?
S2 Statement of scope/perspective	The perspective of the model (relevant costs and consequences) should be stated clearly, and the model inputs should be consistent with the stated perspective and overall objective of the model	Is the perspective of the model stated clearly?
		Are the model inputs consistent with the stated perspective?
	The scope of the decision model should be specified and justified.[a]	Has the scope of the model been stated and justified?
	The outcomes of the model should reflect the perspective and scope of the model, and should be consistent with the objective of the evaluation	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?
	Clinical events should not be excluded on the basis of no statistically significant difference. Clinical events may be excluded if they are not logically expected to differ. This is an iterative process, so clinical events may be included if they may be expected to differ when more data are collected	
S3 Rationale for structure	The structure of the model should be consistent with a coherent theory of the health condition under evaluation, and the treatment pathways (disease states or branches) should be chosen to reflect the underlying biological process of the disease in question and the impact of the intervention. The structure should not be dictated by current patterns of service provision	Has the evidence regarding the model structure been described?
		Is the structure of the model consistent with a coherent theory of the health condition under evaluation?
		Have any competing theories regarding model structure been considered?
	All sources of evidence used to develop and inform the structure of the model (i.e. the theory of disease) should be described. The structure should be consistent with this evidence	Are the sources of data used to develop the structure of the model specified?
		Are the causal relationships described by the model structure justified appropriately?
S4 Structural assumptions	All structural assumptions should be transparent and justified. They should be reasonable in the light of the needs and purposes of the decision-maker	Are the structural assumptions transparent and justified?
		Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?

continued

Dimension of quality		Attributes of good practice	Questions for critical appraisal
S5	Strategies/comparators	There should be a clear definition of the options under evaluation	Is there a clear definition of the options under evaluation?
		All feasible and practical options relating to the stated decision problem should be evaluated. Options should not be constrained by the immediate concerns of the decision-maker or data availability, or limited to current clinical practice. The model should, however, include one or more options that are representative of current practice	Have all feasible and practical options been evaluated? Is there justification for the exclusion of feasible options?
		The reader should refer to guidelines relating to economic evaluation good practice	
S6	Model type	The appropriate model type will be dictated by the stated decision problem and the choices made regarding the causal relationships within the model	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?
S7	Time horizon	A model's time horizon should extend far enough into the future that it can reflect important differences between options	Is the time horizon of the model sufficient to reflect all important differences between options?
		It is important to distinguish between the time horizon of the model, the duration of treatment and the duration of treatment effect	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?
		Lifetime horizons are appropriate for most models, but shorter horizons can sometimes be justified	Has a lifetime horizon been used? If not, has a shorter time horizon been justified?
S8	Disease states/pathways	Disease states/pathways should reflect the underlying biological/clinical process of the disease in question and the impact of interventions	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?
S9	Cycle length	For discrete time models, the cycle length should be dictated by the natural history of disease. It should be the minimum interval over which the pathology or symptoms are expected to alter	Is the cycle length defined and justified in terms of the natural history of disease?
Data			
D1	Data identification	Methods for identifying data should be transparent and it should be clear that the data identified are appropriate given the objectives of the model	Are the data identification methods transparent and appropriate given the objectives of the model?
		There should be justification of any choices that have been made about which specific data inputs are included in a model	Where choices have been made between data sources, are these justified appropriately?
		It should be clear that particular attention has been paid to identifying data for those parameters to which the results of the model are particularly sensitive	Has particular attention been paid to identifying data for the important parameters in the model?
		Identification of data should be carried out in a systematic, but not necessarily comprehensive, manner	Has the process of selecting key parameters been justified and systematic methods used to identify the most appropriate data?
		Where expert opinion has been used to estimate particular parameters, sources and methods of elicitation should be described	Has the quality of the data been assessed appropriately? Where expert opinion has been used, are the methods described and justified?

Dimension of quality	Attributes of good practice	Questions for critical appraisal
D2 Premodel data analysis	All data analysis (premodel) methodology should be described and based on justifiable statistical and epidemiological methods. Specific issues to consider include those listed under D2a–d, below	Is the data analysis (premodel) methodology based on justifiable statistical and epidemiological techniques?
D2a Baseline data	Baseline probabilities may be based on natural history data derived from epidemiological/observational studies or relate to the control group of an experimental study	Is the choice of baseline data described and justified? Are transition probabilities calculated appropriately?
	Rates and interval probabilities should be transformed into transition probabilities appropriately. If there is evidence that time is an important factor in the calculation of transition probabilities in state transition models, this should be incorporated	Has a half-cycle correction been applied to both cost and outcome? If not, has this omission been justified?
	If a half-cycle correction has not been used on all transitions in a state transition model (costs and outcomes), this should be justified	
D2b Treatment effects	Relative treatment effects derived from trial data should be synthesised using recognised meta-analytic techniques	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?
	The methods and assumptions that are used to extrapolate short-term results to final outcomes should be documented and justified. This should include justification of the choice of survival function (e.g. exponential or Weibull forms). Alternative assumptions regarding extrapolation should be explored through sensitivity analysis	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified? Have alternative assumptions been explored through sensitivity analysis?
	Assumptions regarding the continuing effect of treatment once treatment is complete should be documented and justified. If evidence regarding the long-term effect of treatment is lacking, alternative assumptions should be explored through sensitivity analysis	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified? Have alternative assumptions been explored through sensitivity analysis?
D2d Quality of life weights (utilities)	Utilities incorporated into the model should be appropriate for the specified decision problem	Are the utilities incorporated into the model appropriate? Is the source for the utility weights referenced? Are the methods of derivation for the utility weights justified?
D3 Data incorporation	All data incorporated into the model should be described and the sources of all data should be given and reported in sufficient detail to allow the reader to be aware of the type of data that have been incorporated	Have all data incorporated into the model been described and referenced in sufficient detail?
	Where data are not mutually consistent in the model, the choices and assumptions that have been made should be explicit and justified	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?
	The process of data incorporation should be transparent. It should be clear whether data are incorporated as a point estimate or as a distribution. If data have been incorporated as distributions as part of probabilistic analysis, the choice of distribution and its parameters should be described and justified	Is the process of data incorporation transparent? If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified? If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?

continued

129

130

Dimension of quality		Attributes of good practice	Questions for critical appraisal
D4	Assessment of uncertainty	In assessing uncertainty, modellers should distinguish between the four principal types of uncertainty	Have the four principal types of uncertainty been addressed? If not, has the omission of particular forms of uncertainty been justified?
D4a	Methodological	Methodological uncertainty relates to whether particular analytical steps taken in the analysis are the most appropriate	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?
D4b	Structural	There should be evidence that structural uncertainties have been evaluated using sensitivity analysis	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?
D4c	Heterogeneity	It is important to distinguish between uncertainty resulting from the process of sampling from a population and variability due to heterogeneity (i.e. systematic differences between patient subgroups)	Has heterogeneity been dealt with by running the model separately for different subgroups?
D4d	Parameter	Where data have been incorporated into the model as point estimates, the ranges used for sensitivity analysis should be stated and justified	Are the methods of assessment of parameter uncertainty appropriate?
		Probabilistic analysis is the most appropriate method of handling parameter uncertainty because it facilitates assessment of the joint effect of uncertainty over all parameters (see Data incorporation)	Has probabilistic sensitivity analysis been done? If not, has this been justified?
		Threshold analysis may be sufficient in extreme circumstances	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?
Consistency			
C1	Internal consistency	There should be evidence that the internal consistency of the model in terms of its mathematical logic has been evaluated	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?
C2	External consistency	The results of a model should be explicable. Either results should make intuitive sense or counterintuitive results should be fully explained	Are the conclusions valid given the data presented? Are any counterintuitive results from the model explained and justified?
		All relevant data available should be incorporated into a model. Data should not be withheld for purposes of assessing external consistency	If the model has been calibrated against independent data, have any differences been explained and justified?
		The results of a model should be compared with those of previous models and any differences should be explained	Have the results of the model been compared with those of previous models and any differences in results explained?

[a] Have any choices or assumptions been explained sufficiently, in the context of available evidence?

Appendix 7

Bibliography

This appendix provides references relating to particular aspects of decision modelling for cost-effectiveness analysis. The list is based on references known to the authors and which were identified as part of the literature searching for the modelling guidelines.

Meta-analysis

Ades AE, Cliffe S. Markov chain Monte Carlo estimation of a multiparameter decision model: consistency of evidence and the accurate assessment of uncertainty. *Med Decis Making* 2002; **22**:359–71.

Sutton AJ, Abrams KR, Jones, DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *Health Technol Assess* 1999;**2**(19):1–276.

Sutton AJ, Abrams KR, Jones, DR, Sheldon TA, Song F. *Methods for meta analysis in medical research*. London: John Wiley; 2000.

Discounting

Department of Health. *Policy appraisal and health: a guide from the Department of Health*. London: Department of Health; 1995.

Gravelle H, Smith D. Discounting for health effects in cost-benefit and cost–effectiveness analysis. *Health Econ* 2001;**10**:587–99.

Great Britain HM Treasury Green Book: Appraisal and Evaluation in Central Government. London: The Stationery Office; 2003.
URL: http://greenbook.treasury.gov.uk/.
Accessed 19 November 2003.

Krahn M, Gafni A. Discounting in the economic evaluation of health care interventions. *Med Care* 1993;**31**:403–18.

Smith D, Gravelle H. The practice of discounting in economic evaluations of healthcare interventions. *Int J Technol Assess Health Care* 2001; **17**:236–43.

Types of model

Barton P, Bryan S, Robinson S. Modelling in the economic evaluation of health care: selecting the appropriate approach. *Journal of Health Services Research and Policy* 2004;**9**:110–18.

Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, Weintein M. *Decision making in health and medicine: integrating evidence and values*. Cambridge: Cambridge University Press; 2001. Chapter 7.

Kuntz K, Weinstein M. Modelling in economic evaluation. In Drummond M, McGuire A, editors. *Economic evaluation in health care*. Oxford: Oxford University Press; 2001.

Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Med Decis Making* 1993;**13**:322–38.

Adjusting costs between countries/for inflation

Gray AM. Are the results of cost-effectiveness studies internationally transferable? Health Economics Research Unit. Health Economics Study Group (HESG) Meeting, Aberdeen, 1–2 July 1991.

Ludbrook A, Wordsworth S. Making the results comparable in economic evaluations across countries: is purchasing power parity enough? *International Health Economics Association (IHEA) 3rd World Conference Abstract Book*; 2001. p. 159.

Adjusting absolute treatment effect for variation in baseline events

Palmer S, Sculpher M, Philips Z, Robinson R, Ginnelly L, Bakhai A, *et al*. A cost effectiveness model comparing alternative management strategies for the use of glycoprotein IIb/IIIa antagonists in non ST elevation acute coronary syndrome; 2002. URL:
http://www.nice.org.uk/pdf/glycoprotein_modelling_report.pdf. Accessed 19 November 2003.

Transition probabilities

Miller DK, Homan SM. Determining transition probabilities: confusion and suggestions. *Med Decis Making* 1994;**14**:52–8.

Modelling life expectancy

Beck JR, Kassirer JP, Pauker SG. A convenient approximation of life expectancy (the 'DEALE'). I. Validation of the method. *Am J Med* 1982; **73**:883–8.

Beck JR, Pauker SG, Gottlieb LK, *et al*. A convenient approximation of life expectancy (the 'DEALE'). II. Use in medical decision-making. *Am J Med* 1982;**73**:889–97.

Holland RR, Ellis CA, Geller BM, Plante DA, Secker-Walker RH. Life expectancy estimation with breast cancer: bias of the declining exponential function and an alternative to its use. *Med Decis Making* 1999;**19**:385–93.

Kuntz KM, Weinstein MC. Life expectancy biases in clinical decision modeling. *Med Decis Making* 1995;**15**:158–69.

Modelling survival

Billingham LJ, Abrahms KR, Jones DR. Methods of the analysis of quality-of-life and survival data in health technology assessment. *Health Technol Assess* 1999;**3**(10):1–152.

Sanders GD, Hlatky MA, Heidenreich PA, McDonald KM, Owens DK. Use of decision models to extrapolate results from clinical trials: cost effectiveness of the implantable cardioverter defibrillator (ICD). *Med Decis Making* 1999;**19**:519.

Modelling risk factors

Kuntz K, Goldie S. Assessing the sensitivity of decision-analytic results to unobserved markers of risk: defining the effects of heterogeneity bias. *Med Decis Making* 2002;**22**:218–27.

Costing issues

Brouwer WB, Koopmanschap MA. How to calculate indirect costs in economic evaluations. *Pharmacoeconomics* 1998;**13**:563–9.

Brouwer W, Rutten F, Koopmanschap M. Costing in economic evaluations. In Drummond M, McGuire A, editors. *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press, 2001. Chapter 4.

Sculpher M. The role and estimation of productivity costs in economic evaluation. In Drummond M, McGuire A, editors. *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press, 2001. Chapter 5.

Half-cycle correction

Briggs A, Sculpher M. An introduction to Markov modelling for economic evaluation. *Pharmacoeconomics* 1998;**13**:397–409.

Sonnenberg FA, Beck JR. Markov models in medical decision-making: a practical guide. *Med Decis Making* 1993;**13**:322–38.

Uncertainty

Probabilistic sensitivity analysis

Briggs A. Handling uncertainty in economic evaluation and presenting the results. In Drummond M, McGuire A, editors. *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press; 2001. Chapter 8.

Critchfield GC, Willard KE. Probabilistic analysis of decision trees using Monte Carlo simulation. *Med Decis Making* 1986;**6**:85–92.

Doubilet P, Begg CB, Weinstein MC, Braun P, McNeil BJ. Probabilistic sensitivity analysis using Monte Carlo simulation. *Med Decis Making* 1985; **5**:157–77.

General sensitivity analysis

Briggs A. Handling uncertainty in economic evaluation and presenting the results. In Drummond M, McGuire A, editors. *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press; 2001.

Briggs AH. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics* 2000; **17**:479–500.

Hay JW. Economic modeling and sensitivity analysis. *Value Health* 1998;**1**:187–93.

Hunink MG, Bult JR, de Vries J, Weinstein MC. Uncertainty in decision models analyzing cost-effectiveness: the joint distribution of incremental costs and effectiveness evaluated with a nonparametric bootstrap method. *Med Decis Making* 1998;**18**:337–46.

Moayyedi P, Soo S, Deeks J, Forman D, Delaney BC. Markov modelling with probabilistic sensitivity analysis based on meta-analysis: cost-effectiveness of *Helicobacter pylori* eradication for non-ulcer dyspepsia. *Med Decis Making* 2000; **20**:501.

Parmigiani G, Samsa GP, Ancukiewicz M, Lipscomb J, Hasselblad V, Matchar DB. Assessing uncertainty in cost-effectiveness analyses: application to a complex decision model. *Med Decis Making* 1997;**17**:390–401.

Pasta DJ, Taylor JL, Henning JM. Probabilistic sensitivity analysis incorporating the bootstrap: an example comparing treatments for the eradication of *Helicobacter pylori. Med Decis Making* 1999; **19**:353–63.

Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med Decis Making* 1998; **18**(Suppl 2):S68–S80.

Value of information

Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ* 1999;**18**:341–64.

Claxton K, Neumann PJ, Araki S, Weinstein MC. Bayesian value-of-information analysis – an application to a policy model of Alzheimer's disease. *Int J Technol Assess Health Care* 2001; **17**:38–55.

Distributions for probabilistic decision modelling

Briggs AH. Handling uncertainty in economic evaluation. In Drummond MF, McGuire A, editors. *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press; 2001.

Briggs AH, Ades AE, Price MJ. Probabilistic sensitivity analysis for decision trees with multiple branches: use of the Dirichlet distribution in a Bayesian framework. *Med Decis Making* 2003; **23**:341–50.

133

Appendix 8

Search strategies and results of the case study

(a) Search strategies used to identify data to populate the UTI model

Baseline event rates
Incidence and prevalence
Questions 1, 5, 4 (first part) and 8 (first part)
MEDLINE (1966 to January week 2 2003)
Searched 27 January 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (recurren$ or recrudescense$ or remission$ or relapse$ or reinfection$ or re-infection$).ti,ab.
8. Recurrence/
9. 7 or 8
10. 6 and 9
11. exp infants/
12. (frequency or frequencies or occurrence$ or incidence$ or prevalence$ or number$ or times or rate or rates or episode$ or natural history).ti,ab.
13. exp data collection/
14. 12 or 13
15. 10 and 11 and 14

This search strategy produced 108 records.

Questions 1, 5, 4 (first part) and 8 (first part)
EMBASE (1980 to week 6 2003)
Searched 21 February 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infection"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (recurren$ or recrudescense$ or remission$ or relapse$ or reinfection$ or re-infection$).ti,ab.

8. Recurrent Disease/
9. 7 or 8
10. 6 and 9
11. (frequency or frequencies or occurrence$ or incidence$ or prevalence$ or number$ or times or rate or rates or episode$ or natural history).ti,ab.
12. exp information processing/
13. 11 or 12
14. 10 and 13
15. exp infant/
16. 14 and 15

This search strategy produced 35 records.

Questions 2, 3, 4 (second part), 6, 7 and 8 (second part)
MEDLINE (1966 to January week 2 2003)
Searched 29 January 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. exp infants/
8. (frequency or frequencies or occurrence$ or incidence$ or prevalence$ or number$ or times or rate or rates or episode$ or natural history).ti,ab.
9. exp data collection/
10. 8 or 9
11. 6 and 7 and 10
12. Vesico-Ureteral Reflux/
13. vesicoureteral reflux.ti,ab.
14. Vesico-Ureteral Reflux.ti,ab.
15. vur.ti,ab.
16. or/12-15
17. 11 and 16
18. 17 and (mild or severe or grade$).af.

This search strategy produced 28 records.

Questions 2, 3, 4 (second part), 6, 7 and 8 (second part)
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infection"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (frequency or frequencies or occurrence$ or incidence$ or prevalence$ or number$ or times or rate or rates or episode$ or natural history).ti,ab.
8. exp information processing/
9. 7 or 8
10. Vesicoureteral Reflux/
11. vesicoureteral reflux.ti,ab.
12. Vesico-Ureteral Reflux.ti,ab.
13. vur.ti,ab.
14. or/10-13
15. exp infant/
16. 6 and 9 and 15 and 14
17. (mild or severe or grade$).af.
18. 16 and 17

This search strategy produced 18 records.

Questions 9, 12 (first part), 13 and 16 (first part)
MEDLINE (1966 to January week 2 2003)
Searched 29 January 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (recurren$ or recrudescense$ or remission$ or relapse$ or reinfection$ or re-infection$).ti,ab.
8. Recurrence/
9. 7 or 8
10. 6 and 9
11. exp child/
12. (frequency or frequencies or occurrence$ or incidence$ or prevalence$ or number$ or times or rate or rates or episode$ or natural history).ti,ab.
13. exp data collection/
14. 12 or 13
15. 10 and 11 and 14

This search strategy produced 201 records.

Questions 9, 12 (first part), 13 and 16 (first part)
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infection"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (recurren$ or recrudescense$ or remission$ or relapse$ or reinfection$ or re-infection$).ti,ab.
8. Recurrent Disease/
9. 7 or 8
10. 6 and 9
11. exp child/
12. (frequency or frequencies or occurrence$ or incidence$ or prevalence$ or number$ or times or rate or rates or episode$ or natural history).ti,ab.
13. exp information processing/
14. 12 or 13
15. 10 and 11 and 14

This search strategy produced 112 records.

Questions 10, 11, 12 (second part), 14, 15 and 16 (second part)
MEDLINE (1966 to January week 2 2003)
Searched 29 January 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. exp child/
8. (frequency or frequencies or occurrence$ or incidence$ or prevalence$ or number$ or times or rate or rates or episode$ or natural history).ti,ab.
9. exp data collection/
10. 8 or 9
11. 6 and 7 and 10
12. Vesico-Ureteral Reflux/
13. vesicoureteral reflux.ti,ab.
14. Vesico-Ureteral Reflux.ti,ab.
15. vur.ti,ab.
16. or/12-15
17. 11 and 16
18. 17 and (mild or severe or grade$).af.

This search strategy produced 30 records.

Questions 10, 11, 12 (second part), 14, 15 and 16 (second part)
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infection"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. exp child/
8. (frequency or frequencies or occurrence$ or incidence$ or prevalence$ or number$ or times or rate or rates or episode$ or natural history).ti,ab.
9. exp information processing/
10. 8 or 9
11. 6 and 7 and 10
12. Vesicoureteral Reflux/
13. vesicoureteral reflux.ti,ab.
14. Vesico-Ureteral Reflux.ti,ab.
15. vur.ti,ab.
16. or/12-15
17. 11 and 16
18. 17 and (mild or severe or grade$).af.

This search strategy produced 38 records.

Rates of occurrence
Questions 17 and 20
MEDLINE (1966 to January week 2 2003)
Searched 29 January 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (recurren$ or recrudescense$ or remission$ or relapse$ or reinfection$ or re-infection$).ti,ab.
8. Recurrence/
9. 7 or 8
10. 6 and 9
11. exp infants/
12. (proportion or proportions or rate or rates or percent$ or per cent$ or incidence or level).mp. [mp=title, abstract, cas registry/ec number word, mesh subject heading]
13. pyelonephriti$.ti,ab.
14. exp pyelonephritis/
15. 13 or 14

16. 10 and 11 and 12 and 15

This search strategy produced 25 records.

Questions 17 and 20
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infection"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (recurren$ or recrudescense$ or remission$ or relapse$ or reinfection$ or re-infection$).ti,ab.
8. Recurrent Disearse/
9. 7 or 8
10. 6 and 9
11. (proportion or proportions or rate or rates or percent$ or per cent$ or incidence or level).mp. [mp=title, abstract, subject headings, drug trade name, original title, device manufacturer, drug manufacturer name]
12. pyelonephriti$.ti,ab.
13. exp pyelonephritis/
14. 12 or 13
15. exp infant/
16. 15 and 14 and 10
17. 16 and 11

This search strategy produced eight records.

Questions 18, 19, 21, 22, 24, 25, 27 and 28
MEDLINE (1966 to January week 2 2003)
Searched 29 January 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. pyelonephriti$.ti,ab.
2. exp pyelonephritis/
3. 1 or 2
4. *"Urinary Tract Infections"/di, ep [Diagnosis, Epidemiology]
5. *bacteriuria/di, ep [Diagnosis, Epidemiology]
6. (uti or utis).ti.
7. (urinary tract adj3 infection$).ti.
8. bacteriuria.ti.
9. or/4-8
10. (proportion or proportions or rate or rates or percent$ or per cent$ or incidence or level).mp. [mp=title, abstract, cas registry/ec number word, mesh subject heading]
11. Vesico-Ureteral Reflux/
12. vesicoureteral reflux.ti,ab.

13. Vesico-Ureteral Reflux.ti,ab.
14. vur.ti,ab.
15. or/11-14
16. 3 and 9 and 15 and 10
17. (mild or severe or grade$).af.
18. 16 and 17

This search strategy produced 13 records.

Questions 18, 19, 21, 22, 24, 25, 27 and 28
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb
Gateway (http://gateway.ovid.com/athens)

1. pyelonephriti$.ti,ab.
2. exp pyelonephritis/
3. 1 or 2
4. *"Urinary Tract Infection"/di, ep [Diagnosis, Epidemiology]
5. *bacteriuria/di, ep [Diagnosis, Epidemiology]
6. (uti or utis).ti.
7. (urinary tract adj3 infection$).ti.
8. bacteriuria.ti.
9. or/4-8
10. (proportion or proportions or rate or rates or percent$ or per cent$ or incidence or level).mp. [mp=title, abstract, cas registry/ec number word, mesh subject heading]
11. Vesicoureteral Reflux/
12. vesicoureteral reflux.ti,ab.
13. Vesico-Ureteral Reflux.ti,ab.
14. vur.ti,ab.
15. or/11-14
16. 3 and 9 and 15 and 10
17. (mild or severe or grade$).af.
19. 17 and 18

This search strategy produced 13 records.

Questions 23 and 26
MEDLINE (1966 to January week 2 2003)
Searched: 29 January 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (recurren$ or recrudescense$ or remission$ or relapse$ or reinfection$ or re-infection$).ti,ab.
8. Recurrence/
9. 7 or 8
10. 6 and 9
11. exp child/

12. (proportion or proportions or rate or rates or percent$ or per cent$ or incidence or level).mp. [mp=title, abstract, cas registry/ec number word, mesh subject heading]
13. pyelonephriti$.ti,ab.
14. exp pyelonephritis/
15. 13 or 14
16. 10 and 11 and 12 and 15

This search strategy produced 40 records.

Questions 23 and 26
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb
Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infection"/di, ep [Diagnosis, Epidemiology]
2. *bacteriuria/di, ep [Diagnosis, Epidemiology]
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. or/1-5
7. (recurren$ or recrudescense$ or remission$ or relapse$ or reinfection$ or re-infection$).ti,ab.
8. Recurrent Disease/
9. 7 or 8
10. 6 and 9
11. exp child/
12. (proportion or proportions or rate or rates or percent$ or per cent$ or incidence or level).mp. [mp=title, abstract, cas registry/ec number word, mesh subject heading]
13. pyelonephriti$.ti,ab.
14. exp pyelonephritis/
15. 13 or 14
16. 10 and 11 and 12 and 15

This search strategy produced 19 records.

Relationship of events
Question 29
MEDLINE (1966 to January week 2 2003)
Searched 29 January 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. pyelonephriti$.ti,ab.
2. pyonephrosis.ti,ab.
3. *PYELONEPHRITIS/
4. or/1-3
5. renal scar$.mp.
6. kidney scar$.mp.
7. exp Kidney Diseases/
8. exp kidney/
9. *Cicatrix/
10. (7 or 8) and 9
11. renal lesion$.mp.

12. kidney lesion$.mp.
13. renal damage.mp.
14. (cause or causes or causative or relations or relationship$ or link or effect or etiology).ti,ab.
15. (subsequent or lead or leads or leading or correlated or related or complications).ti,ab.
16. 14 or 15
17. 5 or 6 or 10 or 11 or 12 or 13
18. 4 and 17 and 16
19. animal/ not (animal/ and human/)
20. 18 not 19

This search strategy produced 167 records.

Question 29
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb
Gateway (http://gateway.ovid.com/athens)

1. pyelonephriti$.ti,ab.
2. pyonephrosis.ti,ab.
3. *PYELONEPHRITIS/
4. or/1-3
5. renal scar$.mp.
6. kidney scar$.mp.
7. Kidney Scar/
8. renal lesion$.mp.
9. kidney lesion$.mp.
10. renal damage.mp.
11. (cause or causes or causative or relations or relationship$ or link or effect or etiology).ti,ab.
12. (subsequent or lead or leads or leading or correlated or related or complications).ti,ab.
13. 11 or 12
14. (animal/ or nonhuman/) not ((animal/ or nonhuman/) and human/)
15. or/5-10
16. 4 and 15 and 13
17. 16 not 14

This search strategy produced 146 records.

Question 30
MEDLINE (1966 to January week 2 2003)
Searched 29 February 2003 on OvidWeb
Gateway (http://gateway.ovid.com/athens)

1. renal scar$.mp.
2. kidney scar$.mp.
3. exp Kidney Diseases/
4. exp kidney/
5. *Cicatrix/
6. (3 or 4) and 5
7. (cause or causes or causative or relations or relationship$ or link or effect or etiology).ti,ab.

8. (subsequent or lead or leads or leading or correlated or related or complications).ti,ab.
9. 7 or 8
10. animal/ not (animal/ and human/)
11. Kidney Failure, Chronic/
12. esrd.ti,ab.
13. end-stage renal disease$.ti,ab.
14. end-stage kidney disease$.ti,ab.
15. renal insufficiency.ti,ab.
16. kidney insufficiency.ti,ab.
17. renal failure.ti,ab.
18. or/11-17
19. 1 or 2 or 6
20. 9 and 18 and 19
21. 20 not 10

This search strategy produced 64 records.

Question 30
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb
Gateway (http://gateway.ovid.com/athens)

1. Kidney Failure, Chronic/
2. esrd.ti,ab.
3. end-stage renal disease$.ti,ab.
4. end-stage kidney disease$.ti,ab.
5. renal insufficiency.ti,ab.
6. kidney insufficiency.ti,ab.
7. renal failure.ti,ab.
8. or/1-7
9. renal scar$.mp.
10. kidney scar$.mp.
11. Kidney Scar/
12. (cause or causes or causative or relations or relationship$ or link or effect or etiology).ti,ab.
13. (subsequent or lead or leads or leading or correlated or related or complications).ti,ab.
14. 12 or 13
15. (animal/ or nonhuman/) not ((animal/ or nonhuman/) and human/)
16. or/9-11
17. 16 and 8 and 14
18. 17 not 15

This search strategy produced 74 records.

Question 31
MEDLINE (1966 to January week 2 2003)
Searched 29 January 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

((consequence$ or complication$ or outcome$ or sequelae or long-term effects or long-term effect or impact or impacts) and (renal scar$ or kidney scar$ or kidney lesion$ or renal lesion$ or renal damage or kidney damage)).ti.

This search strategy produced ten records.

Question 31
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb
Gateway (http://gateway.ovid.com/athens)

((consequence$ or complication$ or outcome$ or sequelae or long-term effects or long-term effect or impact or impacts) and (renal scar$ or kidney scar$ or kidney lesion$ or renal lesion$ or renal damage or kidney damage)).ti.

This search strategy produced six records.

Question 32
MEDLINE (1966 to January week 2 2003)
Searched 29 January 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

((consequence$ or complication$ or outcome$ or sequelae or long-term effects or long-term effect or impact or impacts) adj (esrd or renal insufficiency or kidney insufficiency or end-stage kidney disease or end-stage renal disease or end-stage renal failure or end-stage kidney failure)).ti.

This search strategy produced 41 records.

Question 32
EMBASE (1980 to week 6 2003)
Searched 24 March 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

((consequence$ or complication$ or outcome$ or sequelae or long-term effects or long-term effect or impact or impacts) adj (esrd or renal insufficiency or kidney insufficiency or end-stage kidney disease or end-stage renal disease or end-stage renal failure or end-stage kidney failure)).ti.

This search strategy produced 26 records.

Questions 1–32
IPD (1994 to 1 February 2003)
Searched 25 February 2003 on DialogWeb
(http://library.dialog.com/)

s urinary tract(3w)infection?
s bacteriuria
s uti or utis
s s1:s3
s recurren?
s recrudescense?
s remission?
s relapse?
s reinfection?
s re(w)infection?
s s5:s10
s infant?
s s4 and s11 and s12
s child?
s s4 and s11 and s14
s pyelonephriti?
s pyonephrosis
s s16:s17
s s4 and s11 and s18
s vesico(w)ureteral(w)reflux
s vesicoureteral(w)reflux
s vur
s s20:s22
s severe or mild or grade?
s s24 and s4 and s23 and s12
s s24 and s4 and s23 and s14
s s4 and s23 and s18 and s24
s renal(w)scar?
s kidney(w)scar?
s renal(w)lesion?
s kidney(w)lesion?
s renal(w)damage
s cicatrix
s s28:s33
s kidney(w)failure
s esrd
s end(w)stage(w)renal(w)disease?
s end(w)stage(w)kidney(w)disease?
s renal(w)insufficiency
s kidney(w)insufficiency
s renal(w)failure
s s35:s41
s consequence?
s complication?
s outcome?
s sequelae
s long(w)term(w)effects
s long(w)term(w)effect
s impact
s impacts
s s43:s50
s s34 and s51
s s42 and s51
s cause
s causes
s causative
s relations
s relationship?
s link
s effect
s etiology
s subsequent
s lead
s leads
s leading
s correlated

s related
s complications
s s54:s68
s s69 and s34 and s42
s s69 and s34 and s18
s13 or s15 or s25 or s26 or s27 or s52 or s53 or s70 or s71

This yielded 947 hits, of which 25 were selected by the information officer as being of potential relevance.

Health-related quality of life and its valuation
First set of health-related quality of life searches
Question 33 (using search filter by Paisley *et al.*, 2002[35])
MEDLINE (1966 to January week 2 2003)
Searched 30 January 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/
2. *bacteriuria/
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. pyelonephriti$.ti,ab.
7. exp pyelonephritis/
8. pyonephrosis.ti,ab.
9. or/1-8
10. sf36.ti.
11. (ed5d or euroqol).ti.
12. (hrql or hrqol).ti.
13. health related quality of life.ti.
14. rosser.ti.
15. standard gamble.ti.
16. (utilities or utility).ti.
17. qaly$.ti,ab.
18. 16 and 17
19. qwb$.ti.
20. (quality of wellbeing or quality of well being).ti.
21. (or/10-15) or (or/18-20)
22. 9 and 21
23. exp child/ or exp infants/
24. 22 and 23

This search strategy produced no records.

Question 33 (using brief outcome filter by Brettle *et al.*, 1998,[34] with duration of symptoms terms added)
MEDLINE (1966 to January week 2 2003)
Searched 30 January 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/
2. *bacteriuria/

3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. pyelonephriti$.ti,ab.
7. exp pyelonephritis/
8. pyonephrosis.ti,ab.
9. or/1-8
10. "health status indicators"/
11. "outcome and process assessment (health care)"/
12. "outcome assessment (health care)"/
13. quality of life/
14. outcome measure$.tw.
18. health measure$.tw.
19. or/10-18
20. 9 and 19
21. exp child/ or exp infant/
22. 20 and 21
23. ((duration or length or time or period) adj2 symptoms).tw.
24. 9 and 21 and 23
25. 22 or 24

This search strategy produced 28 records.

Question 33 (using comprehensive outcome filter by Brettle *et al.*, 1998,[34] with duration of symptoms terms added)
MEDLINE (1966 to February week 2 2003)
Searched 20 February 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/
2. *bacteriuria/
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. pyelonephriti$.ti,ab.
7. exp pyelonephritis/
8. pyonephrosis.ti,ab.
9. or/1-8
10. "health status indicators"/
11. "outcome and process assessment (health care)"/
12. "outcome assessment (health care)"/
13. quality of life/
14. health status/
15. severity of illness index/
16. "Self Assessment (Psychology)"/
17. outcome measure$.tw.
18. health status.tw. or health outcome$.tw
19. quality of life.tw.
20. (endpoint$ or end point$ or end-point$).tw.
21. (self-report$ or self report$).tw.
22. functional outcome$.tw.
23. outcome$.ti.
24. outcome$.tw.

25. measure$.tw.
26. assess$.tw.
27. (score$ or scoring).tw.
28. index.tw.
29. indices.tw.
30. scale$.tw.
31. monitor$.tw.
32. or/10-23
33. or/25-31
34. 24 and 33
35. 32 or 34
36. 35 and 9
37. exp child/ or exp infant/
38. 36 and 37
39. ((duration or length or time or period) adj2 symptoms).tw.
40. 9 and 37 and 39
41. 38 or 40

This search strategy produced 143 records.

Revised Health-related quality of life searches
Question 33
MEDLINE (1966 to April week 3 2003)
Searched 1 May 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. (sf36 or sf 36).tw.
2. (eq5d or eq 5d or euroqol or euro qol).tw.
3. (short form 36 or shortform 36 or sf thirtysix or sf thirty six or shortform thirtysix or shortform thirty six or short form thirtysix or short form thirty six).tw.
4. (hrql or hrqol or h qol or hql or hqol).tw.
5. (hye or hyes or health$ year$ equivalent$ or health utilit$).tw.
6. health related quality of life.tw.
7. rosser.tw.
8. (standard gamble$ or time trade off or time tradeoff or tto or willingness to pay).tw.
9. (disutilities or disutility or daly or disability adjusted life).tw.
10. "Quality of Life"/
11. health status indicators/
12. quality adjusted life year/
13. (qaly$ or quality adjusted life or quality of life or life quality).tw.
14. qwb$.tw.
15. (quality of wellbeing or quality of well being).tw.
16. factor analysis.tw.
17. preference based.tw.
18. health status.tw.
19. (state adj2 (value or values or valuing or valued)).tw.
20. hspv.tw.
21. *"Urinary Tract Infections"/

22. *bacteriuria/
23. (uti or utis).ti.
24. (urinary tract adj3 infection$).ti.
25. bacteriuria.ti.
26. pyelonephriti$.ti,ab.
27. exp pyelonephritis/
28. pyonephrosis.ti,ab.
29. or/21-28
30. or/1-20
31. 29 and 30
32. life expectancy/
33. life expectancy.tw.
34. 32 or 33
35. 29 and 34
36. ((duration or length or period of time or lasting or last or lasted) adj4 symptom$).ti,ab.
37. 29 and 36

This yielded 51 papers from the quality of life filter, 13 from life expectancy and 35 from duration of symptoms. In total, after deduplication, this combination produced 95 records.

Question 33
EMBASE (1980 to week 17 2003)
Searched 1 May 2003 on OvidWeb Gateway (http://gateway.ovid.com/athens)

1. (sf36 or sf 36).tw.
2. (eq5d or eq 5d or euroqol or euro qol).tw.
3. (short form 36 or shortform 36 or sf thirtysix or sf thirty six or shortform thirtysix or short form thirtysix or short form thirty six).tw.
4. (hrql or hrqol or h qol or hql or hqol).tw.
5. (hye or hyes or health$ year$ equivalent$ or health utilit$).tw.
6. health related quality of life.tw.
7. rosser.tw.
8. (standard gamble$ or time trade off or time tradeoff or tto or willingness to pay).tw.
9. (disutilities or disutility or daly or disability adjusted life).tw.
10. "Quality of Life"/
11. health status indicators/
12. quality adjusted life year/
13. (qaly$ or quality adjusted life or quality of life or life quality).tw.
14. qwb$.tw.
15. (quality of wellbeing or quality of well being).tw.
16. factor analysis.tw.
17. preference based.tw.
18. health status.tw.
19. (state adj2 (value or values or valuing or valued)).tw.
20. hspv.tw.

21. *"Urinary Tract Infection"/
22. *bacteriuria/
23. (uti or utis).ti.
24. (urinary tract adj3 infection$).ti.
25. bacteriuria.ti.
26. pyelonephriti$.ti,ab.
27. exp pyelonephritis/
28. pyonephrosis.ti,ab.
29. or/21-28
30. or/1-20
31. 29 and 30
32. life expectancy/
33. life expectancy.tw.
34. 32 or 33
35. 29 and 34
36. ((duration or length or period of time or lasting or last or lasted) adj4 symptom$).ti,ab.
37. 29 and 36

This yielded 85 papers from the quality of life filter, 14 from life expectancy and 25 from duration of symptoms. In total, after deduplication, this combination produced 122 records.

Question 34
MEDLINE (1966 to April week 3 2003)
Searched 1 May 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. animal/ not (animal/ and human/)
2. *Kidney Failure, Chronic/
3. esrd.ti.
4. end-stage renal disease$.ti.
5. end-stage kidney disease$.ti.
6. renal insufficiency.ti.
7. kidney insufficiency.ti.
8. renal failure.ti.
9. or/1-9
10. life expectancy.ti.
11. *life expectancy/
12. 10 or 11 or quality adjusted life expectancy.mp. [mp=title, abstract, cas registry/ec number word, mesh subject heading]
13. 9 and 12
14. 13 not 1

This search strategy produced nine records.

Question 34
EMBASE (1980 to week 17 2003)
Searched 1 May 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. life expectancy.ti.
2. *life expectancy/
3. 1 or 2 or quality adjusted life expectancy.tw.

4. (animal/ or nonhuman/) not ((animal/ or nonhuman/) and human/)
5. *Kidney Failure, Chronic/
6. esrd.ti.
7. end-stage renal disease$.ti.
8. end-stage kidney disease$.ti.
9. renal insufficiency.ti.
10. kidney insufficiency.ti.
11. renal failure.ti.
12. or/5-11
13. 12 and 3
14. 13 not 4

This search strategy produced five records.

Question 35
MEDLINE (1966 to April week 3 2003)
Searched 1 May 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. renal scar$.mp.
2. kidney scar$.mp.
3. exp Kidney Diseases/
4. exp kidney/
5. *Cicatrix/
6. (3 or 4) and 5
7. renal lesion$.mp.
8. kidney lesion$.mp.
9. renal damage.mp.
10. 1 or 2 or 6 or 7 or 8 or 9
11. animal/ not (animal/ and human/)
12. life expectancy.tw.
13. Life Expectancy/
14. 12 or 13
15. 10 and 14
16. 15 not 11

This search strategy produced four records.

Question 35
EMBASE (1980 to week 17 2003)
Searched 1 May 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. life expectancy.tw.
2. Life Expectancy/
3. 1 or 24. renal scar$.mp.
5. kidney scar$.mp.
6. Kidney Scar/
7. renal lesion$.mp.
8. kidney lesion$.mp.
9. renal damage.mp.
10. (animal/ or nonhuman/) not ((animal/ or nonhuman/) and human/)
11. or/4-9
12. 11 and 3
13. 13 not 10

This search strategy produced three records.

Resource use and costs
Question 36
NHS EED
Searched 17 February 2003 on CRD Internal
Administration version using CAIRS software

> S (uti OR utis OR urinary(w)tract(3w)infection$
>   OR pyonephrosis OR pyelonephritis OR
>   bacteriuria)/til

This search strategy produced 58 records.

HEED (Issue February 2003)
Searched 26 February 2003 on CD-ROM

> TI = uti
> TI = utis
> TI = 'urinary tract infection'
> TI = 'urinary tract infections'
> TI = pyonephrosis
> TI = pyelonephritis
> TI = bacteriuria
> CS = 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7

This search strategy produced 64 records.

Department of HES England. *Financial year*
2001/02. **London: Department of Health**
(http://www.doh.gov.uk/hes/)
Searched 26 February 2003

> This gave the number of day cases and hospital
> admissions for kidney or UTIs and the
> number of main operations carried out in
> NHS hospitals in England, 2001–2002.

BNF September 2002 (http://www.bnf.org)
Searched 26 February 2003

> This gave the costs of the antibiotics for UTIs.

McCormick A, Fleming D, Charlton J. Morbidity
statistics from general practice fourth national
study 1991–1992. Series MB5 No. 3. London:
HMSO
Searched 26 February 2003

> This gave the UTI prevalence rates broken
> down by gender and age (0–4, 5–15 years,
> etc.) by new and first ever episodes, patients
> consulting and consultations with doctor
> broken down by age

NHS Reference Costs 2002
(http://www.doh.gov.uk/nhsexec/refcosts.htm)
Searched 26 February 2003

> This gave the costs of kidney or urinary tract
> infections for Primary Care Trusts and the
> cost for dialysis.

Relative treatment effect of interventions
Question 39
CDSR (2003 Issue 1)
Searched 31 January 2003 on CD-ROM

> #1. BACTERIURIA single term (MeSH) 372
> #2. URINARY TRACT INFECTIONS single
>   term (MeSH) 1254
> #3. (uti:ti or utis:ti) 23
> #4. ((urinary:ti next tract:ti) and infection*:ti)
>   964
> #5. bacteriuria:ti 173
> #6. (#1 or #2 or #3 or #4 or #5)

This search retrieved ten Cochrane reviews and
four protocols, which were sifted for relevance by
the information officer. Two reviews and one
protocol were then sent to the modeller.

Question 39
DARE
Searched 31 January 2003 on CRD Public
Administration version using CAIRS software

> S (uti OR utis OR urinary(w)tract(3w)infection$
>   OR pyonephrosis OR pyelonephritis OR
>   bacteriuria)/ttl
> S (urinary-tract-infection$ OR bacteriuria)/kwo
> S s1 OR s2

This produced 30 records, which were checked for
relevance by the information officer. Six of the 30
references were then sent to the modeller.

Question 39
HTA Database
Searched 31 January 2003 on CRD Public
Administration version using CAIRS software

> S (uti OR utis OR urinary(w)tract(3w)infection$
>   OR pyonephrosis OR pyelonephritis OR
>   bacteriuria)/ttl
> S (urinary-tract-infection$ OR bacteriuria)/kwo
> S s1 OR s2

This produced four records, which were checked
by the information officer for potential relevance.
None of the references was of any relevance and
therefore the results were not sent to the modeller.

Question 40
CDSR (2003 Issue 1)
Searched 31 January 2003 on CD-ROM

> #1. VESICO-URETERAL REFLUX explode all
>   trees (MeSH)

#2. vur
#3. (vesico next ureteral next reflux)
#4. (vesicoureteral next reflux)
#5. (#1 or #2 or #3 or #4)

This resulted in six Cochrane reviews and two protocols after an initial sift by the information officer. One protocol was sent to the modeller.

DARE and HTA Database
Searched 31 January 2003 on CRD Public
Administration version using CAIRS software

S vesico(w)ureteral(w)reflux or
vesicoureteral(w)reflux

This yielded no hits, so the search was continued using CENTRAL.

CENTRAL (2003 Issue 1)
Searched 31 January 2003 on CD-ROM

#1. VESICO-URETERAL REFLUX explode all trees (MeSH)
#2. vur
#3. (vesico next ureteral next reflux)
#4. (vesicoureteral next reflux)
#5. (#1 or #2 or #3 or #4)

This resulted in 72 hits, 32 of which were sent to the health economist after an initial sift by the information officer.

Other parameters: antibiotic resistance
Question 41
MEDLINE (1966 to February week 2 2003)
Searched 21 February 2003 on OvidWeb
Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/
2. *bacteriuria/
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. pyelonephriti$.ti,ab.
7. exp pyelonephritis/
8. pyonephrosis.ti,ab.
9. or/1-8
10. *"Trimethoprim Resistance"/
11. *"Nitrofurantoin"/
12. (proloprim or trimethoprim).ti.
13. (trimpex or monotrim or trimopan).ti.
14. (macrodantin or furadantin or macrobid).ti.
15. furadoine.ti.
16. furadonine.ti.
17. furantoin.ti.

18. Nitrofurantoin.ti.
19. (Cephalexin or ceporex or keflex).ti.
20. (ceporexine or cefalexin).ti.
21. palitrex.ti.
22. *"cephalexin"/ or *"cefaclor"/ or *"cefadroxil"/ or *"cefatrizine"/ or *"cephaloglycin"/ or *"cephradine"/
23. *"trimethoprim"/ or *"trimethoprim-sulfamethoxazole combination"/
24. *"drug resistance"/ or *"drug resistance, microbial"/ or *"drug resistance, bacterial"/ or *"drug resistance, multiple"/ or *"drug resistance, multiple, bacterial"/ or *"drug tolerance"/ or *"tachyphylaxis"/
25. (resistance or resistant).ti.
26. or/10-23
27. 24 or 25
28. 9 and 26 and 27
29. exp child/ or exp infant/
30. 28 and 29

This search strategy produced 78 records.

Question 41
EMBASE (1980 to week 6 2003)
Searched 24 February 2003 on OvidWeb
Gateway (http://gateway.ovid.com/athens)

1. *"Urinary Tract Infections"/
2. *bacteriuria/
3. (uti or utis).ti.
4. (urinary tract adj3 infection$).ti.
5. bacteriuria.ti.
6. pyelonephriti$.ti,ab.
7. exp pyelonephritis/
8. pyonephrosis.ti,ab.
9. or/1-8
10. *"Nitrofurantoin"/
11. (proloprim or trimethoprim).ti.
12. (trimpex or monotrim or trimopan).ti.
13. (macrodantin or furadantin or macrobid).ti.
14. furadoine.ti.
15. furadonine.ti.
16. furantoin.ti.
17. Nitrofurantoin.ti.
18. (Cephalexin or ceporex or keflex).ti.
19. (ceporexine or cefalexin).ti.
20. palitrex.ti.
21. *"cefalexin"/ or *"cefaclor"/ or *"cefadroxil"/ or *"cefatrizine"/ or *"cephaloglycin"/ or *"cephradine"/
22. *"trimethoprim"/ or *"TRIMETHOPRIM SULFATE"/ or *"SULFADOXINE PLUS TRIMETHOPRIM"/ or *"TRIMETHOPRIM DERIVATIVE"/
23. *"drug resistance"/ or *"antibiotic resistance"/ or *"cross resistance"/ or *"multidrug

(b) Results by database from searches for the case study

Search strategy	Question number(s)	Total records retrieved for the search	Deduplication within each set of questions	No. of records identified per database					
				MEDLINE	EMBASE	Other database and no. of records	Other database and no. of records	Other database and no. of records	Other database and no. of records
Baseline events data									
1	1,4,5,8	222	199	108	35	IPD 79			
2	2,3,4,6,7,8	50	39	28	18	IPD 4			
3	9,12,13,16	315	242	201	112	IPD 2			
4	10–16	73	49	30	38	IPD 5			
5	17,20	92	86	25	8	IPD 59			
6	18–22,24,25,27	30	21	13	13	IPD 4			
7	23,26	118	105	40	19	IPD 59			
8	29	323	224	167	146	IPD 10			
9	30	155	111	64	74	IPD 17			
10	31	32	30	10	6	IPD 16			
11	32	759	742	41	26	IPD 692			
Quality of life									
12a first search	33	171	143	143+28+0	122				
12b revised search	33	217	173	95	5				
13	34	14	12	9	5				
14	35	7	6	4	3				
Resource use									
15	36	121	99			NHS EED 57	HEED 64		
Effectiveness of interventions									
16	39	48	34			CDSR 14	DARE 30		HTA 4
17	40	78	78			CDSR 6	DARE 0 HTA 0		CENTRAL 72
Resistance to drugs									
18	41	284	242	78	206				
Total		3109	2631	1056	831				

The resulting Endnote Library with the results of all the searches (except for those from IPD) contained 2162 records before deduplication. After deduplication, however, the number of records in the Endnote Library was reduced considerably to 1213 [some of the duplication was between databases for the same search (478 records) and some duplication of the records was from the results of the searches for the different questions (471 records)]. The extra 25 records resulting from the searches in IPD were given to the modeller in printed format. Of these 25 records, three were duplicates from previous searches.

resistance"/ or *"drug tolerance"/ or *"drug cross tolerance"/
24. (resistance or resistant).ti.
25. 23 or 24
26. exp child/ or exp infant/

27. or/10-22
28. 25 and 26 and 27 and 9
29. 25 and 27 and 9

This search strategy produced 206 records.

(c) Relevant records for each question set

Question area	MEDLINE	EMBASE	IPD	
Baseline events data	0	5	6	
	HEED	NHS EED		
Resource use and unit costs	2	2		
	MEDLINE			
Health-related quality of life using Paisley and Brettle's filters	6			
	MEDLINE	EMBASE		
Revised health-related quality of life searches	2	1		
Life expectancy	3	3		
	CDSR	DARE	HTA	CCTR
Relative treatment effects	2	7	0	10
	MEDLINE	EMBASE		
Other parameters: antibiotic resistance	2	2		

Appendix 9

Searches for the effects of bias on treatment outcomes

(a) First searches: scoping search

MEDLINE 1966 (March week 2 2003)
Searched 26 March 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. "bias (epidemiology)"/ or selection bias/
2. "publication bias"/
3. "confounding factors (epidemiology)"/
4. ((exaggerat$ or precision or bias or biased or biases or accurat$ or inaccurat$ or underexaggerat$ or imprecise or larger or smaller) adj estimate$ adj2 (treatement or effect or intervention or benefit or harm or efficacy)).mp. [mp=title, abstract, cas registry/ ec number word, mesh subject heading]
5. (underestimat$ adj2 (treatement or effect or intervention or benefit or harm or efficacy)).mp.
6. (bias or biases or biased).ti.
7. exp research design/
8. exp epidemiologic study characteristics/ or exp epidemiologic research design/
9. odds ratio/
10. or/1-6
11. or/7-9
12. 10 and 11
13. limit 12 to yr= 1990-2003

(b) Second set searches: focused search strategies

MEDLINE (1966 to April week 4 2003)
Searched 12 May 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. *"bias (epidemiology)"/ or *selection bias/
2. *"publication bias"/
3. *"confounding factors (epidemiology)"/
4. ((overestimat$ or shrink$ or magnitude or exaggerat$ or understimat$ or precision or accurat$ or inaccurat$ or underexaggerat$ or overexaggerat$ or imprecis$ or larger or smaller) adj2 (effect estimate$ or treatment effect$ or treatment size$ or effect size$ or bias$ or benefit or harm or efficacy)).ti,ab.

5. ((quantif$ or amount or measur$ or correct$) adj2 bias$).ti,ab.
6. ((deal$ or calibrat$ or impact$ or effect$ or adjust$) adj2 bias$).ti,ab.
7. (bias or biased or biases).ti.
8. *research design/mt, st, sn or *control groups/mt, st, sn or *double-blind method/mt, st, sn or *patient selection/mt, st, sn or *random allocation/mt, st, sn or *sample size/mt, st, sn
9. *epidemiologic studies/mt, st, sn or *case-control studies/mt, st, sn or *retrospective studies/mt, st, sn or *cohort studies/mt, st, sn or *longitudinal studies/mt, st, sn or *follow-up studies/mt, st, sn or *prospective studies/mt, st, sn or *cross-sectional studies/mt, st, sn or *seroepidemiologic studies/mt, st, sn or *hiv seroprevalence/mt, st, sn or *clinical trials/mt, st, sn or *clinical trials, phase i/mt, st, sn or *clinical trials, phase ii/mt, st, sn or *clinical trials, phase iii/mt, st, sn or *clinical trials, phase iv/mt, st, sn or *controlled clinical trials/mt, st, sn or *multicenter studies/mt, st, sn or *feasibility studies/mt, st, sn or *intervention studies/mt, st, sn or *pilot projects/mt, st, sn or *sampling studies/mt, st, sn or *twin studies/mt, st, sn or *epidemiologic research design/mt, st, sn or *control groups/mt, st, sn or *cross-over studies/mt, st, sn or *double-blind method/mt, st, sn or *matched-pair analysis/mt, st, sn or *random allocation/mt, st, sn or *reproducibility of results/mt, st, sn or *sample size/mt, st, sn or *"sensitivity and specificity"/mt, st, sn or *predictive value of tests/mt, st, sn or *roc curve/mt, st, sn or *single-blind method/mt, st, sn
10. *Cost-Benefit Analysis/st, sn, mt or *odds ratio/ or *markov chains/ or *meta-analysis/ or *randomized controlled trials/
11. 8 or 9 or 10
12. 1 or 2 or 3 or 4 or 5 or 6 or 7
13. 12 and 11
14. limit 13 to yr= 1990-2003

This yielded 388 records.

EMBASE (1980 to week 19 2003)
Searched 13 May 2003 on OvidWeb Gateway
(http://gateway.ovid.com/athens)

1. ((overestimat$ or shrink$ or magnitude or exaggerat$ or understimat$ or precision or accurat$ or inaccurat$ or underexaggerat$ or overexaggerat$ or imprecis$ or larger or smaller) adj2 (effect estimate$ or treatment effect$ or treatment size$ or effect size$ or bias$ or benefit or harm or efficacy)).ti,ab.
2. ((quantif$ or amount or measur$ or correct$) adj2 bias$).ti,ab.
3. ((deal$ or calibrat$ or impact$ or effect$ or adjust$) adj2 bias$).ti,ab.
4. (bias or biased or biases).ti.
5. bias.sh.
6. (or/1-4) or 5
7. exp *economic evaluation/
8. *types of study/ or *feasibility study/ or *theoretical study/ or *drug comparison/ or *controlled study/ or *case control study/ or *randomized controlled trial/ or *meta analysis/ or *outcomes research/ or *stochastic model/ or *placebo/
9. 7 or 8
10. 6 and 9
11. limit 10 to yr=1990-2004

This yielded 133 records.

HTA Database
Searched 13 May 2003
(http://nhscrd.york.ac.uk/welcome.htm)

Bias or biased or biases

This yielded 47 records.

Cochrane Database of Methodology Reviews (2003 Issue 2) (web version)
Browsed 13 May 2003 (http://www.update-software.com/clibng/cliblogon.htm)

This database is still relatively small, so the nine completed and seven ongoing reviews were browsed for potentially relevant reviews.

CMR (2003 Issue 2) (web version)
Searched 13 May 2003 (http://www.update-software.com/clibng/cliblogon.htm)

BIAS (EPIDEMIOLOGY) explode all trees (MeSH)
PUBLICATION BIAS single term (MeSH)
(bias:ti or biased:ti or biases:ti)
(#1 or #2 or #3)

This yielded 256 records.

DARE
Searched 13 May 2003 on CRD administration database using CAIRS software

S m/st1
S bias or biased or biases/til
S bias/kwo
S (overestimat$ or shrink$ or magnitude or exaggerat$ or underestimat$ or precision or accurat$ or underexaggerat$ or overexaggerat$ or imprecis$ or larger or smaller)(2w)(estimat$ or treatment(W)effect$ or treatment(w)size$ or effect(w)size$ or bias$ or benefit or harm or efficacy)
S S2:S4
S S1 and S5

This yielded 175 records.

NHS EED
Searched 13 May 2003 on CRD administration database using CAIRS software S methodology/st1

S bias or biased or biases/til
S bias/kwo
S (overestimat$ or shrink$ or magnitude or exaggerat$ or underestimat$ or precision or accurat$ or underexaggerat$ or overexaggerat$ or imprecis$ or larger or smaller)(2w)(estimat$ or treatment(W)effect$ or treatment(w)size$ or effect(w)size$ or bias$ or benefit or harm or efficacy)
S S2:S4
S S1 and S5

This yielded 26 records.

(c) Data extraction table for studies that explore quantification of bias

Study	Objective and Design	Results	Conclusion
Balk et al., 2002[46]	**Objective:** To determine whether quality measures are associated with treatment effect size in RCTs. **Design:** Quality measures from published quality assessment scales were evaluated in RCTs included in meta-analyses from four medical areas (cardiovascular disease, infectious disease, paediatrics and surgery). For each quality measure, a relative OR for treatment effect was calculated, defined as the ratio of the strength of the treatment effect in studies in which the quality measure was present to the strength of the effect in studies in which it was absent	24 quality measures were analysed for 276 RCTs from 26 meta-analyses. Relative ORs of high- versus low-quality studies for these quality measures ranged from 0.83 to 1.26; none was statistically significantly associated with treatment effect. The proportion of studies fulfilling specific quality measures varied widely in the four medical areas. In analyses limited to specific medical areas, placebo control, multicentre studies, study country, caregiver blinding, and statistical methods were significantly associated with treatment effect on seven occasions. These relative ORs ranged from 0.40 to 1.74. However, the directions of these associations were not consistent Quality items related to randomisation: • 'Randomisation methods described' (OR 1.03, 95% CI 0.89 to 1.20) • 'Central randomisation site' (OR 1.01, 95% CI 0.85 to 1.18)) • 'Allocation concealment' (OR 1.05, 95% CI 0.91 to 1.21)	Individual quality measures are not reliably associated with the strength of treatment effect across studies and medical areas. Although use of specific quality measures may be appropriate in specific well-defined areas in which there is pertinent evidence, findings of associations with treatment effect cannot be generalised to all clinical areas or meta-analyses
Benson and Hartz, 2000[47]	**Objective:** To compare the results of observational studies with those of randomised controlled trials **Design:** Literature search (1985–1998) to identify observational studies that compared two or more treatments or interventions for the same condition. Subsequently to identify all RCTs and observational studies comparing the same treatments for these conditions	136 reports (53 observational studies and 83 RCTs) about 19 diverse treatments were found. In most cases, the estimates of the treatment effects from observational studies and RCTs were similar. In two of the 19 analyses of treatment effects, did the combined magnitude of the effect in observational studies lie outside the 95% CI for the combined magnitude in the RCTs?	Little evidence was found that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different to those obtained in RCTs Comment: Qualification 'similar' is very broad

Study	Objective and Design	Results	Conclusion
Britton *et al.*, 1998[48]	**Objective:** To explore those issues related to the process of randomisation that may affect the validity of conclusions drawn from the results of RCTs and non-randomised studies. Four research questions were addressed. • Do non-randomised studies differ systematically from RCTs in terms of treatment effect? • Are there systematic differences between included and excluded individuals and do these influence the measured treatment effect? • To what extent is it possible to adjust for baseline differences between study groups? • How important is patient preference in terms of outcome? **Design:** The review was based on a series of systematic reviews involving structured searches of databases	18 papers that directly compared the results of RCTs and prospective non-randomised studies were found and analysed. No obvious patterns emerged; neither the RCTs nor the non-randomised studies consistently gave larger or smaller estimates of the treatment effect. The type of intervention did not appear to be influential, although more comparisons need to be conducted before definite conclusions can be drawn. The number of eligible subjects included in the RCTs ranged from 1 to 100%. Reasons for exclusions may be medical (e.g. high risk of adverse events in certain groups) or scientific (selecting only small homogeneous groups to increase the precision of estimated treatment effects). Blanket exclusions (e.g. the elderly, women of childbearing potential) are also common in RCTs. Large clinical databases containing detailed information on patient severity and prognosis have been used instead of RCTs, and where database subjects are selected according to the same inclusion criteria as RCTs, the treatment effects of the two methods are similar. Most RCTs failed to document adequately the characteristics of eligible individuals who did not participate in trials. However, RCTs were more likely than non-randomised trials to include university and teaching centres, and this may have exaggerated the treatment effect measured in the RCTs. Participation in RCTs differed between studies of treatment interventions (subjects tended to be less affluent, less educated and more severely ill, and therefore had greater capacity to benefit from treatment) and those evaluating preventive interventions (more affluent, better educated and generally healthier, and therefore had less potential to benefit than eligible subjects who declined to participate). Adjustment for differences in baseline prognostic factors in non-randomised studies often changed the treatment effect size but not significantly; importantly, the direction of change was inconsistent. Most of the case studies were too small to draw conclusions, but where this was possible, the superiority of one treatment over another was probably a function of the patients' clinical characteristics	Results of RCTs and non-randomised studies do not inevitably differ, and the available evidence suffers from many limitations. It does, however, suggest that it may be possible to minimise any differences by ensuring that subjects included in each type of study are comparable. The effect of adjustment for baseline differences between groups in non-randomised studies is inconsistent but, where it is done, it should involve rigorously developed formulae. Existing studies have generally been too small to assess the impact of such adjustment

Study	Objective and Design	Results	Conclusion
Concato *et al.*, 2000[49]	**Objective:** To identify RCTs and observational studies that examined the same clinical topic through published meta-analyses; and to compare the results of the original reports according to the type of research design **Design:** MEDLINE search for articles published in five major medical journals from 1991 to 1995	For the five clinical topics and 99 original articles evaluated, the average results of the observational studies were remarkably similar to those of the RCTs: • bacille Calmette–Guérin vaccine and tuberculosis: 13 RCTs: RR = 0.49 (95% CI 0.34 to 0.70); ten case–control studies: RR = 0.50 (95% CI 0.39 to 0.65) • screening mammography and mortality from breast cancer: eight RCTs: RR = 0.79 (95% CI 0.71 to 0.88); four case–control studies: RR = 0.61 (95% CI 0.49 to 0.77) • cholesterol levels and death due to trauma: six RCTs: RR = 1.42 (95% CI 0.94 to 2.15); 14 cohort studies: RR = 1.40 (95% CI 1.14 to 1.66) • treatment of hypertension and stroke: 14 RCTs: RR = 0.58 (95% CI 0.50 to 0.67); seven cohort studies: RR = 0.62 (95% CI 0.60 to 0.65) • treatment of hypertension and coronary heart disease: 14 RCTs: RR = 0.86 (95% CI 0.78 to 0.96); nine cohort studies: RR = 0.77 (95% CI 0.75 to 0.80)	The results of well-designed observational studies (with either a cohort or case–control design) do not systematically overestimate the magnitude of the effects of treatment compared with those in RCTs on the same topic
Ioannidis *et al.*, 2001[41]	**Objective:** To compare results of randomised and non-randomised studies that evaluated medical interventions and to examine characteristics that may explain discrepancies between randomised and non-randomised studies **Design:** MEDLINE (1966 to March 2000), the Cochrane Library (Issue 3 2000), and major journals were searched	45 diverse topics were identified for which both randomised trials ($n = 240$) and non-randomised studies ($n = 168$) had been performed and had been considered in meta-analyses of binary outcomes. Very good correlation was observed between the summary ORs of randomised and non-randomised studies ($r = 0.75$; $p < 0.001$); however, non-randomised studies tended to show larger treatment effects (28 vs 11, $p = 0.009$). Between-study heterogeneity was frequent among randomised trials alone (23%) and very frequent among non-randomised studies alone (41%). The summary results of the two types of design differed beyond chance in seven cases (16%). Discrepancies beyond chance were less common when only prospective studies were considered (8%). Occasional differences in sample size and timing of publication were also noted between discrepant randomised and non-randomised studies. In 28 cases (62%), the natural logarithm of the OR differed by at least 50%, and in 15 cases (33%), the OR varied at least two-fold between non-randomised studies and randomised trials	Despite good correlation between randomised trials and non-randomised studies – in particular, prospective studies – discrepancies beyond chance do occur and differences in estimated magnitude of treatment effect are very common

Study	Objective and Design	Results	Conclusion
Ioannidis and Karassa, 2001[38]	**Objective:** To assess the prevalence of biases from selection of patients with extreme characteristics in recent uncontrolled therapeutic studies in rheumatology **Design:** A handsearch of four major rheumatology journals for uncontrolled trials published in 1997 or 1998 that measured therapeutic efficacy by comparing one or more variables at follow-up versus at baseline. Evaluation of the susceptibility to bias from random measurement error and natural variability for variables used for defining eligibility that overlap with those used for defining outcomes	25 studies were analysed. In 22 studies, the eligibility criteria were related to the outcome criteria and defined a patient population with extreme characteristics. Only three studies clearly reported that they had performed a baseline measurement separate from the screening (eligibility) measurement. The remaining 19 reports (76%) might be susceptible to bias: in seven, identical variables were used for eligibility criteria and outcomes, three used outcome variables that were also used for characterising eligibility along with other criteria, two used specific eligibility variables that were part of composite outcome scores and seven selected patients on the basis of vague descriptors of disease severity, while disease severity was also the outcome	Several recent uncontrolled trials of therapeutic interventions in rheumatology are subject to biases stemming from the selection of patients with extreme characteristics. Separate baseline evaluations from the screening measurements should be performed and eligibility criteria and outcomes should be carefully defined The extent of the bias is not reported
Kunz and Oxman, 1998[42]	**Objective:** To summarise comparisons of randomised clinical trials and non-randomised clinical trials, trials with adequately concealed random allocation versus inadequately concealed random allocation, and high-quality trials versus low-quality trials where the effect of randomisation could not be separated from the effects of other methodological manoeuvres **Design:** Systematic review	RCTs versus non-RCTs of the same intervention: eight studies found; two underestimated the effect, one found similar effects, five overestimated the effect RCTs versus non-RCTs of the same intervention: three studies found; two inconclusive, one found similar effects Trials with adequately concealed allocation versus inadequately concealed allocation: two studies found; two overestimated the effect High-quality versus low-quality trials: five studies found; two underestimated the effect, one found similar effects, one overestimated the effect, one reversal of effect Failure to use random allocation and concealment of allocation was associated with relative increases in estimates of effects of 150% or more, relative decreases of up to 90%, inversion of the estimated effect and, in some cases, no difference. On average, failure to use randomisation or adequate concealment of allocation resulted in larger estimates of effect owing to a poorer prognosis in non-randomly selected control groups compared with randomly selected control groups	Failure to use adequately concealed random allocation can distort the apparent effects of care in either direction, causing the effects to seem either larger or smaller than they really are. The size of these distortions can be as large as or larger than the size of the effects that are to be detected

Study	Objective and Design	Results	Conclusion
MacLehose *et al.*, 2000[50] See also: Reeves *et al.*, 1998[51]	**Objective:** To investigate the association between methodological quality and the magnitude of estimates of effectiveness by systematically comparing estimates of effectiveness derived from RCTs QEO studies. Quantifying any such association should help healthcare decision-makers to judge the strength of evidence from non-randomised studies. Two strategies were used: • a comparison of the RCT and QEO study estimates of effectiveness of any intervention, where both estimates were reported in a single paper • a comparison of the RCT and QEO study estimates of effectiveness for specified interventions, where the estimates were reported in different papers **Design:** Relevant literature was identified from: the Cochrane Library, MEDLINE, EMBASE, DARE and SCI, references of relevant papers already identified and experts. Electronic searches were very difficult to design and yielded few papers for the first strategy and when identifying study designs	Strategy 1: 14 papers were identified, yielding 38 comparisons between RCT and QEO study estimates; 25 were classified as low and 13 as high quality. Discrepancies between RCT and QEO study estimates of effect size and outcome frequency for intervention and control groups were smaller for high- than low-quality comparisons. For high-quality comparisons, no tendency was observed for QEO study estimates of effect size to be more extreme than RCT ones, but this tendency was seen with low-quality comparisons Strategy 2: 34 papers were identified, 17 evaluating mammographic screening and 17 folic acid supplementation; eight and four papers, respectively, were individually or cluster assigned RCTs, five and six were non-randomised trials or cohort studies, and three and six were matched or unmatched case–control studies. Two studies, one of mammographic screening and one of folic acid supplementation, used some other study design. κ statistics for most items were < 0.4, although the percentage agreement usually exceeded 60%. Cronbach α values for different aspects of quality were < 0.5, suggesting that the instrument had limited ability to differentiate aspects of quality. Regression analyses showed that both cohort and case–control studies had lower total quality scores than RCTs; cohort studies also had significantly lower scores than case–control studies. The latter, counterintuitive finding may reflect a general tendency for quasi-experimental studies (which must use cohort designs) to have lower quality than observational studies. Meta-regression of study attributes against relative risk estimates showed no association between effect size and study quality. Estimates from RCTs and cohort studies were not significantly different, but case–control studies gave significantly different estimates for both mammographic (greater benefit) and folic acid supplementation (less benefit)	The findings of strategy 1 suggest that QEO study estimates of effectiveness may be valid if important confounding factors are controlled for. The small size of discrepancies for high-quality comparisons also implies that psychological factors (e.g. treatment preferences or willingness to be randomised) had a negligible effect on outcome. However, the authors caution against generalising their findings to other contexts, for three main reasons: • Few papers were reviewed, and the findings may depend on the specific interventions evaluated. • Most high-quality comparisons studied RCT and QEO study populations that met the same eligibility criteria, which may have reduced the importance of controlling for confounding. • The literature reviewed is likely to have been subject to some form of publication bias. Authors of papers appeared to have strong a priori views about the usefulness of evidence from QEO studies, and the findings of papers appeared to support these views Strategy 2 found no association between study quality and effect size for either intervention, after taking account of study design. The lack of association between quality and effect size could have arisen for a variety of reasons, the most likely being that study quality is not associated with relative risk in a predictable way or that the instrument failed to characterise methodological quality adequately The primary aim of quantifying any association between methodological quality and effect size was thwarted by several obstacles. For objective 1, the authors were unable to draw strong conclusions because of the paucity of evidence, and the potentially unrepresentative nature of the evidence they reviewed. For objective 2, the authors were unable adequately to distinguish and measure the variations in different aspects of quality between studies. The authors' recommendations relate directly to these obstacles

Study	Objective and Design	Results	Conclusion
Marcus, 1997[39]	**Objective:** In some randomised clinical trials, a large proportion of patients eligible for randomisation may withhold consent to be randomised. When the subjects in the randomised trial differ from the eligible population with respect to characteristics that are associated with the magnitude of the treatment effect, there may be non-consent bias, i.e. the treatment effect for those in the randomised trial may not reflect the treatment effect for the eligible population. In response to this problem, some investigators have conducted, in addition to the randomised trial, a separate non-randomised but otherwise identical trial consisting of those patients who are eligible for randomisation, but instead choose their own treatment. This paper clarifies the ways in which parallel randomised and non-randomised clinical trials can provide useful information to assess non-consent bias **Design:** A parallel randomised and non-randomised trial which compares adenoidectomy versus identical treatment for children with recurrent otitis media is used as an illustration	Observed baseline covariate data can be used to adjust for differences between the randomised population and the eligible population when estimating the treatment effect for the eligible population. After adjusting, different outcomes for the randomised versus non-randomised treated groups and/or the randomised versus non-randomised control groups reflect the presence of hidden non-consent bias resulting from differences between the trial population and the eligible population with respect to unobserved covariates. A sensitivity analysis can display how hidden non-consent bias can account for an imbalance in the treatment groups with respect to an unobserved covariate In summary, there does not appear to be much non-consent bias in the estimate of the treatment effect due to imbalance with respect to measured covariates. Moreover, there is no evidence of hidden non-consent bias. However, since this study is sensitive to small hidden non-consent bias, it is possible that the estimate of the treatment effect derived from the randomised trial slightly overestimates the treatment effect for the eligible population if the parents of those children who were severely affected at baseline refused randomisation for their children	If there is concern that those who enter the randomised trial may differ from the entire population who meet the eligibility requirements, the addition of the non-randomised trial gives valuable information for assessing non-consent bias. It is worthwhile to collect baseline characteristics, etc. The extent of the bias is not reported.
Moyer and Finney, 2002[52]	**Objective:** To compare the participants, methodological features and post-treatment functioning in randomised and non-randomised studies of alcohol treatment. **Design:** Using systematic literature search techniques, 232 randomised and 92 non-randomised trials published between 1970 and 1998 were identified. Each investigation was scored on a methodological quality scale, participant characteristics were recorded and participants' follow-up abstinence and improvement rates were calculated	Randomised investigations were more likely to use participant selection criteria, to use recognised diagnostic criteria to characterise participants and stringently to implement treatment and assess outcomes. Non-randomised investigations were more likely to assess outcomes in higher proportions of participants over longer follow-up periods and to have greater statistical power to detect treatment effects. Abstinence and improvement rates following active treatment were similar for the two types of design (proportion abstinent: 0.35 and 0.39; proportion improved: 0.52 and 0.54, for randomised and non-randomised trials, respectively), even when differences in study features were controlled	The contrasting strengths and weaknesses of randomised and non-randomised studies suggest that they should be considered as complementary forms of treatment evaluation, in the alcohol treatment field and perhaps more generally

Study	Objective and Design	Results	Conclusion
Sackett, 1979[40]	**Objective:** Description of biases that affect the specification and selection of the study sample and the measurement of exposures and outcomes and their consequences **Design:** Nine biases described (five sampling biases and four measurement biases). Sampling biases: prevalence–incidence bias, admission rate bias, unmasking bias, non-respondent bias and membership bias	Prevalence–incidence bias: a late look at those exposed (or affected) early will miss fatal and other short episodes, plus mild or 'silent' cases and cases in which evidence of exposure disappears with disease onset. This bias may distort relative odds in either direction Admission rate bias: if hospitalisation rates differ for different exposure/disease groups, the relation between exposure and disease will become distorted in hospital-based studies. Relative odds may be spuriously increased or reduced by the admission rate bias Unmasking bias: an innocent exposure may become suspect if, rather than causing a disease, it causes a sign or symptom, which precipitates a search for the disease. Unmasking bias may lead to spuriously increased estimates of relative odds Non-respondent bias: non-respondents (or latecomers) from a specified sample may exhibit exposures or outcomes, which differ from those of respondents (or early-comers); the antithetical bias is called volunteer bias. The effect of the non-respondent bias upon relative odds is obvious and serves as the basis for repeated admonitions both to achieve response rates of at least 80% and to compare responders and non-responders Membership bias: membership in a group (the employed, joggers, etc.) may imply a degree of health which differs systematically from that of the general population. This bias may affect cohort as well as case–control analytical studies	The author believes that several biases can be both prevented and measured, although this mainly applies to measurement biases Case–control studies: • prevalence–incidence bias: not preventable; partially measurable • admission rate bias: not preventable without sacrificing its value; not measurable • unmasking bias: preventable, but overmatch; measurable • non-respondent bias: preventable; measurable. • membership bias: not preventable, partially measurable Cohort studies: • prevalence–incidence bias: preventable; measurable • admission rate bias: does not apply • unmasking bias: preventable; measurable • non-respondent bias: preventable; measurable • membership bias: not preventable, partially measurable The extent of the bias is not reported

Study	Objective and Design	Results	Conclusion
Sacks *et al.*, 1982[43]	**Objective:** To compare the use of RCTs and HCTs for clinical trials **Design:** Search of the literature for therapies studied by both methods	Six therapies were found, for which 50 RCTs and 56 HCTs were reported. 44 of 56 HCTs (79%) found the therapy better than the control regimen, but only ten of 50 RCTs (20%) agreed. For each therapy, the treated patients in RCTs and HCTs had similar outcomes. The difference between RCTs and HCTs of the same therapy was largely due to differences in outcome for the control group, with the HCT control patients generally doing worse than the RCT control groups. Adjustment of the outcomes of HCTs for prognostic factors, when possible, did not appreciably change the results	The data suggest that biases in patient selection may irretrievably weight the outcome of HCTs in favour of new therapies. RCTs may miss clinically important benefits because of inadequate attention to sample size. The predictive value of each might be improved by reconsidering the use of $p < 0.05$ as the significance level for all types of clinical trials, and by the use of confidence intervals around estimates of treatment effects
Schulz *et al.*, 1995[44]	**Objective:** To determine whether inadequate approaches to RCT design and execution are associated with evidence of bias in estimating treatment effects **Design:** An observational study in which the methodological quality of 250 controlled trials from 33 meta-analyses was assessed and then analysed, using multiple logistic regression models, the associations between those assessments and estimated treatment effects. Data sources: meta-analyses from the Cochrane Pregnancy and Childbirth Database	Main outcome measures: the associations between estimates of treatment effects and inadequate allocation concealment, exclusions after randomisation and lack of double-blinding. Compared with trials in which authors reported adequately concealed treatment allocation, trials in which concealment was either inadequate or unclear (did not report or incompletely reported a concealment approach) yielded larger estimates of treatment effects ($p < 0.001$). ORs were exaggerated by 41% for inadequately concealed trials and by 30% for unclearly concealed trials (adjusted for other aspects of quality). Trials in which participants had been excluded after randomisation did not yield larger estimates of effects, but that lack of association may be due to incomplete reporting. Trials that were not double-blind also yielded larger estimates of effects ($p = 0.01$), with ORs being exaggerated by 17%	This study provides empirical evidence that inadequate methodological approaches in controlled trials, particularly those representing poor allocation concealment, are associated with bias. Readers of trial reports should be wary of these pitfalls, and investigators must improve their design, execution and reporting of trials
Shadish and Ragsdale, 1996[45]	**Objective:** Psychotherapy meta-analyses commonly combine results from controlled experiments that use random and non-random assignment without examining whether the two methods give the same answer. Results from this article call this practice into question **Design:** A systematic search for studies of marital or family psychotherapy ($n = 84$) or enrichment ($n = 16$)	With the use of outcome studies of marital and family therapy, 64 experiments using random assignment yielded consistently higher mean post-test effects and less variable post-test effects than 36 studies using non-random assignment. This difference was reduced by about half by taking into account various covariates, especially pre-test effect size levels and various characteristics of control groups. The importance of this finding depends on (1) whether one is discussing meta-analysis or primary experiments, (2) how precise an answer is desired, and (3) whether some adjustment to the data from studies using non-random assignment is possible	It is concluded that studies using non-random assignment may produce acceptable approximations to results from randomised experiments under some circumstances, but that reliance on results from randomised experiments as the gold standard is still well founded

OR, odds ratio; RR, relative risk.

Health Technology Assessment Programme

Prioritisation Strategy Group

Members

Chair,
Professor Tom Walley,
Director, NHS HTA Programme,
Department of Pharmacology &
Therapeutics,
University of Liverpool

Professor Bruce Campbell,
Consultant Vascular & General
Surgeon, Royal Devon & Exeter
Hospital

Professor Shah Ebrahim,
Professor in Epidemiology
of Ageing, University of
Bristol

Dr John Reynolds, Clinical
Director, Acute General
Medicine SDU, Radcliffe
Hospital, Oxford

Dr Ron Zimmern, Director,
Public Health Genetics Unit,
Strangeways Research
Laboratories, Cambridge

HTA Commissioning Board

Members

Programme Director,
Professor Tom Walley,
Director, NHS HTA Programme,
Department of Pharmacology &
Therapeutics,
University of Liverpool

Chair,
Professor Shah Ebrahim,
Professor in Epidemiology of
Ageing, Department of Social
Medicine, University of Bristol

Deputy Chair,
Professor Jenny Hewison,
Professor of Health Care
Psychology, Academic Unit of
Psychiatry and Behavioural
Sciences, University of Leeds
School of Medicine

Dr Jeffrey Aronson
Reader in Clinical
Pharmacology, Department of
Clinical Pharmacology,
Radcliffe Infirmary, Oxford

Professor Ann Bowling,
Professor of Health Services
Research, Primary Care and
Population Studies,
University College London

Professor Andrew Bradbury,
Professor of Vascular Surgery,
Department of Vascular Surgery,
Birmingham Heartlands
Hospital

Professor John Brazier, Director
of Health Economics,
Sheffield Health Economics
Group, School of Health &
Related Research,
University of Sheffield

Dr Andrew Briggs, Public
Health Career Scientist, Health
Economics Research Centre,
University of Oxford

Professor Nicky Cullum,
Director of Centre for Evidence
Based Nursing, Department of
Health Sciences, University of
York

Dr Andrew Farmer, Senior
Lecturer in General Practice,
Department of Primary Health
Care, University of Oxford

Professor Fiona J Gilbert,
Professor of Radiology,
Department of Radiology,
University of Aberdeen

Professor Adrian Grant,
Director, Health Services
Research Unit, University of
Aberdeen

Professor F D Richard Hobbs,
Professor of Primary Care &
General Practice, Department of
Primary Care & General
Practice, University of
Birmingham

Professor Peter Jones, Head of
Department, University
Department of Psychiatry,
University of Cambridge

Professor Sallie Lamb, Research
Professor in Physiotherapy/Co-
Director, Interdisciplinary
Research Centre in Health,
Coventry University

Professor Julian Little,
Professor of Epidemiology,
Department of Medicine and
Therapeutics, University of
Aberdeen

Professor Stuart Logan,
Director of Health & Social
Care Research, The Peninsula
Medical School, Universities of
Exeter & Plymouth

Professor Tim Peters, Professor
of Primary Care Health Services
Research, Division of Primary
Health Care, University of
Bristol

Professor Ian Roberts, Professor
of Epidemiology & Public
Health, Intervention Research
Unit, London School of
Hygiene and Tropical Medicine

Professor Peter Sandercock,
Professor of Medical Neurology,
Department of Clinical
Neurosciences, University of
Edinburgh

Professor Mark Sculpher,
Professor of Health Economics,
Centre for Health Economics,
Institute for Research in the
Social Services, University of York

Professor Martin Severs,
Professor in Elderly Health
Care, Portsmouth Institute of
Medicine

Dr Jonathan Shapiro, Senior
Fellow, Health Services
Management Centre,
Birmingham

Ms Kate Thomas,
Deputy Director,
Medical Care Research Unit,
University of Sheffield

Professor Simon G Thompson,
Director, MRC Biostatistics
Unit, Institute of Public Health,
Cambridge

Ms Sue Ziebland,
Senior Research Fellow,
Cancer Research UK,
University of Oxford

Current and past membership details of all HTA 'committees' are available from the HTA website (www.ncchta.org)

Diagnostic Technologies & Screening Panel

Members

Chair,
Dr Ron Zimmern, Director of the Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

Ms Norma Armston, Freelance Consumer Advocate, Bolton

Professor Max Bachmann Professor Health Care Interfaces, Department of Health Policy and Practice, University of East Anglia

Professor Rudy Bilous Professor of Clinical Medicine & Consultant Physician, The Academic Centre, South Tees Hospitals NHS Trust

Dr Paul Cockcroft, Consultant Medical Microbiologist/Laboratory Director, Public Health Laboratory, St Mary's Hospital, Portsmouth

Professor Adrian K Dixon, Professor of Radiology, Addenbrooke's Hospital, Cambridge

Dr David Elliman, Consultant in Community Child Health, London

Professor Glyn Elwyn, Primary Medical Care Research Group, Swansea Clinical School, University of Wales Swansea

Dr John Fielding, Consultant Radiologist, Radiology Department, Royal Shrewsbury Hospital

Dr Karen N Foster, Clinical Lecturer, Dept of General Practice & Primary Care, University of Aberdeen

Professor Antony J Franks, Deputy Medical Director, The Leeds Teaching Hospitals NHS Trust

Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London

Dr Edmund Jessop, Medical Adviser, National Specialist Commissioning Advisory Group (NSCAG), Department of Health, London

Dr Jennifer J Kurinczuk, Consultant Clinical Epidemiologist, National Perinatal Epidemiology Unit, Oxford

Dr Susanne M Ludgate, Medical Director, Medical Devices Agency, London

Dr William Rosenberg, Senior Lecturer and Consultant in Medicine, University of Southampton

Dr Susan Schonfield, CPHM Specialised Services Commissioning, Croydon Primary Care Trust

Dr Margaret Somerville, Director of Public Health, Teignbridge Primary Care Trust

Professor Lindsay Wilson Turnbull, Scientific Director, Centre for MR Investigations & YCR Professor of Radiology, University of Hull

Professor Martin J Whittle, Head of Division of Reproductive & Child Health, University of Birmingham

Dr Dennis Wright, Consultant Biochemist & Clinical Director, Pathology & The Kennedy Galton Centre, Northwick Park & St Mark's Hospitals, Harrow

Pharmaceuticals Panel

Members

Chair,
Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Oxford Radcliffe Hospital

Professor Tony Avery, Professor of Primary Health Care, University of Nottingham

Professor Stirling Bryan, Professor of Health Economics, Health Services Management Centre, University of Birmingham

Mr Peter Cardy, Chief Executive, Macmillan Cancer Relief, London

Dr Christopher Cates, GP and Cochrane Editor, Bushey Health Centre

Professor Imti Choonara, Professor in Child Health, University of Nottingham, Derbyshire Children's Hospital

Mr Charles Dobson, Special Projects Adviser, Department of Health

Dr Robin Ferner, Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Dr Karen A Fitzgerald, Pharmaceutical Adviser, Bro Taf Health Authority, Cardiff

Mrs Sharon Hart, Managing Editor, *Drug & Therapeutics Bulletin*, London

Dr Christine Hine, Consultant in Public Health Medicine, Bristol South & West Primary Care Trust

Professor Stan Kaye, Professor of Medical Oncology, Consultant in Medical Oncology/Drug Development, The Royal Marsden Hospital

Ms Barbara Meredith, Project Manager Clinical Guidelines, Patient Involvement Unit, NICE

Dr Frances Rotblat, CPMP Delegate, Medicines Control Agency, London

Professor Jan Scott, Professor of Psychological Treatments, Institute of Psychiatry, University of London

Mrs Katrina Simister, New Products Manager, National Prescribing Centre, Liverpool

Dr Richard Tiner, Medical Director, Association of the British Pharmaceutical Industry

Dr Helen Williams, Consultant Microbiologist, Norfolk & Norwich University Hospital NHS Trust

Current and past membership details of all HTA 'committees' are available from the HTA website (www.ncchta.org)

Expert Advisory Network

Members

Professor Douglas Altman,
Director of CSM & Cancer
Research UK Med Stat Gp,
Centre for Statistics in
Medicine, University of Oxford,
Institute of Health Sciences,
Headington, Oxford

Professor John Bond,
Director, Centre for Health
Services Research,
University of Newcastle upon
Tyne, School of Population &
Health Sciences,
Newcastle upon Tyne

Mr Shaun Brogan,
Chief Executive, Ridgeway
Primary Care Group, Aylesbury

Mrs Stella Burnside OBE,
Chief Executive,
Office of the Chief Executive.
Trust Headquarters,
Altnagelvin Hospitals Health &
Social Services Trust,
Altnagelvin Area Hospital,
Londonderry

Ms Tracy Bury,
Project Manager, World
Confederation for Physical
Therapy, London

Mr John A Cairns,
Professor of Health Economics,
Health Economics Research
Unit, University of Aberdeen

Professor Iain T Cameron,
Professor of Obstetrics and
Gynaecology and Head of the
School of Medicine,
University of Southampton

Dr Christine Clark,
Medical Writer & Consultant
Pharmacist, Rossendale

Professor Collette Mary Clifford,
Professor of Nursing & Head of
Research, School of Health
Sciences, University of
Birmingham, Edgbaston,
Birmingham

Professor Barry Cookson,
Director,
Laboratory of Healthcare
Associated Infection,
Health Protection Agency,
London

Professor Howard Stephen Cuckle,
Professor of Reproductive
Epidemiology, Department of
Paediatrics, Obstetrics &
Gynaecology, University of
Leeds

Professor Nicky Cullum,
Director of Centre for Evidence
Based Nursing, University of York

Dr Katherine Darton,
Information Unit, MIND – The
Mental Health Charity, London

Professor Carol Dezateux,
Professor of Paediatric
Epidemiology, London

Mr John Dunning,
Consultant Cardiothoracic
Surgeon, Cardiothoracic
Surgical Unit, Papworth
Hospital NHS Trust, Cambridge

Mr Jonothan Earnshaw,
Consultant Vascular Surgeon,
Gloucestershire Royal Hospital,
Gloucester

Professor Martin Eccles,
Professor of Clinical
Effectiveness, Centre for Health
Services Research, University of
Newcastle upon Tyne

Professor Pam Enderby,
Professor of Community
Rehabilitation, Institute of
General Practice and Primary
Care, University of Sheffield

Mr Leonard R Fenwick,
Chief Executive, Newcastle
upon Tyne Hospitals NHS Trust

Professor David Field,
Professor of Neonatal Medicine,
Child Health, The Leicester
Royal Infirmary NHS Trust

Mrs Gillian Fletcher,
Antenatal Teacher & Tutor and
President, National Childbirth
Trust, Henfield

Professor Jayne Franklyn,
Professor of Medicine,
Department of Medicine,
University of Birmingham,
Queen Elizabeth Hospital,
Edgbaston, Birmingham

Ms Grace Gibbs,
Deputy Chief Executive,
Director for Nursing, Midwifery
& Clinical Support Servs,
West Middlesex University
Hospital, Isleworth

Dr Neville Goodman,
Consultant Anaesthetist,
Southmead Hospital, Bristol

Professor Alastair Gray,
Professor of Health Economics,
Department of Public Health,
University of Oxford

Professor Robert E Hawkins,
CRC Professor and Director of
Medical Oncology, Christie CRC
Research Centre, Christie
Hospital NHS Trust, Manchester

Professor F D Richard Hobbs,
Professor of Primary Care &
General Practice, Department of
Primary Care & General
Practice, University of
Birmingham

Professor Allen Hutchinson,
Director of Public Health &
Deputy Dean of ScHARR,
Department of Public Health,
University of Sheffield

Dr Duncan Keeley,
General Practitioner (Dr Burch
& Ptnrs), The Health Centre,
Thame

Dr Donna Lamping,
Research Degrees Programme
Director & Reader in Psychology,
Health Services Research Unit,
London School of Hygiene and
Tropical Medicine, London

Mr George Levvy,
Chief Executive, Motor
Neurone Disease Association,
Northampton

Professor James Lindesay,
Professor of Psychiatry for the
Elderly, University of Leicester,
Leicester General Hospital

Professor Rajan Madhok,
Medical Director & Director of
Public Health, Directorate of
Clinical Strategy & Public
Health, North & East Yorkshire
& Northern Lincolnshire Health
Authority, York

Professor David Mant,
Professor of General Practice,
Department of Primary Care,
University of Oxford

Professor Alexander Markham,
Director, Molecular Medicine
Unit, St James's University
Hospital, Leeds

Dr Chris McCall,
General Practitioner,
The Hadleigh Practice,
Castle Mullen

Professor Alistair McGuire,
Professor of Health Economics,
London School of Economics

Dr Peter Moore,
Freelance Science Writer,
Ashtead

Dr Andrew Mortimore,
Consultant in Public Health
Medicine, Southampton City
Primary Care Trust

Dr Sue Moss,
Associate Director, Cancer
Screening Evaluation Unit,
Institute of Cancer Research,
Sutton

Professor Jon Nicholl,
Director of Medical Care
Research Unit, School of Health
and Related Research,
University of Sheffield

Mrs Julietta Patnick,
National Co-ordinator, NHS
Cancer Screening Programmes,
Sheffield

Professor Robert Peveler,
Professor of Liaison Psychiatry,
University Mental Health
Group, Royal South Hants
Hospital, Southampton

Professor Chris Price,
Visiting Chair – Oxford,
Clinical Research, Bayer
Diagnostics Europe,
Cirencester

Ms Marianne Rigge,
Director, College of Health,
London

Dr Eamonn Sheridan,
Consultant in Clinical Genetics,
Genetics Department,
St James's University Hospital,
Leeds

Dr Ken Stein,
Senior Clinical Lecturer in
Public Health, Director,
Peninsula Technology
Assessment Group,
University of Exeter

Professor Sarah Stewart-Brown,
Director HSRU/Honorary
Consultant in PH Medicine,
Department of Public Health,
University of Oxford

Professor Ala Szczepura,
Professor of Health Service
Research, Centre for Health
Services Studies, University of
Warwick

Dr Ross Taylor,
Senior Lecturer,
Department of General Practice
and Primary Care,
University of Aberdeen

Mrs Joan Webster,
Consumer member, HTA –
Expert Advisory Network

Current and past membership details of all HTA 'committees' are available from the HTA website (www.ncchta.org)

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.