A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy

How heterogeneity has been examined in systematic reviews of diagnostic test accuracy

J Dinnes,[1*] J Deeks,[2] J Kirby[3] and P Roderick[4]

[1] Wessex Institute for Health Research and Development, University of Southampton, UK
[2] Centre for Statistics in Medicine, Oxford, UK
[3] Southampton Health Technology Assessments Centre, University of Southampton, UK
[4] Health Care Research Unit, University of Southampton, UK

* Corresponding author

Executive summary

Health Technology Assessment 2005; Vol. 9: No. 12

Health Technology Assessment
NHS R&D HTA Programme

Executive summary

Background

Systematic reviews of therapeutic interventions are now commonplace in many if not most areas of healthcare, and in recent years interest has turned to applying similar techniques to research evaluating diagnostic tests. One of the key parts of any review is to consider how similar or different the available primary studies are and what impact any differences have on studies' results. Between-study differences or heterogeneity in results can result from chance, from errors in calculating accuracy indices or from true heterogeneity, that is, differences in design, conduct, participants, tests and reference tests. An important additional consideration for diagnostic studies is differences in results due to variations in the chosen threshold for a positive result for either the index or reference test.

Dealing with heterogeneity is particularly challenging for diagnostic test reviews, not least because test accuracy is conventionally represented by a pair of statistics and not by a single measure of effect such as relative risk, and as a result a variety of statistical methods are available that differ in the way in which they tackle the bivariate nature of test accuracy data:

- methods that undertake independent analyses of each aspect of test performance
- methods that further summarise test performance into a single summary statistic
- methods that use statistical models that simultaneously consider both dimensions of test performance.

The validity of a choice of meta-analytical method depends in part on the pattern of variability (heterogeneity) observed in the study results. However, currently there is no empirical guidance to judge which methods are appropriate in which circumstances, and the degree to which different methods yield comparable results. All this adds to the complexity and difficulty of undertaking systematic reviews of diagnostic test accuracy.

Objectives

Our objective was to review how heterogeneity has been examined in systematic reviews of diagnostic test accuracy studies.

Methods

Systematic reviews that evaluated a diagnostic or screening test by including studies that compared a test with a reference test were identified from the Centre for Reviews and Dissemination's Database of Abstracts of Reviews of Effects. Reviews for which structured abstracts had been written up to December 2002 were screened for inclusion. Data extraction was undertaken using standardised data extraction forms by one reviewer and checked by a second.

Results

A total of 189 systematic reviews met our inclusion criteria and were included in the review. The median number of studies included in the reviews was 18 [inter-quartile range (IQR) 20]. Meta-analyses ($n = 133$) have a higher number with a median of 22 studies (IQR 20) compared with 11 (IQR 13) for narrative reviews ($n = 56$).

Identification of heterogeneity

Graphical plots to demonstrate the spread in study results were provided in 56% of meta-analyses; in 79% of cases these were in the form of plots of sensitivity and specificity in the receiver operating characteristic (ROC) space (commonly termed 'ROC plots').

Statistical tests to identify heterogeneity were used in 32% of reviews: 41% of meta-analyses and 9% of reviews using narrative syntheses. The χ^2 test and Fisher's exact test to assess heterogeneity in individual aspects of test performance were most commonly used. In contrast, only 16% of meta-analyses used correlation coefficients to test for a threshold effect.

Type of syntheses used

A narrative synthesis was used in 30% of reviews. Of the meta-analyses, 52% carried out statistical pooling alone, 18% conducted only summary receiver operator characteristic (SROC) analyses and 30% used both methods of statistical synthesis. Of the reviews that pooled accuracy indices, most pooled each aspect of test performance separately with only a handful

producing single summaries of test performance such as the diagnostic odds ratio. For those undertaking SROC analyses, the main differences between the models used were the weights chosen for the regression models. In fact, in 42% of cases (27/64) the use of, or choice of, weight was not provided by the review authors.

The proportion of reviews using statistical pooling alone has declined over time from 67% in 1995 to 42% in 2001, with a corresponding increase in the use of SROC methods, from 33% to 58%. However, two-thirds of those using SROC methods also carried out statistical pooling rather than presenting only SROC models. Reviews using SROC analyses also tended to present their results as some combination of sensitivity and specificity rather than using alternative, perhaps less clinically meaningful, means of data presentation such as diagnostic odds ratios.

Investigation of heterogeneity sources

Three-quarters of meta-analyses attempted to investigate statistically possible sources of variation, using subgroup analysis (76) or regression analysis (44). The median number of variables investigated was four, ranging from one variable in 20% of reviews to over six in 27% of reviews. The ratio of median number of variables to median number of studies was 1:6.

The impact of clinical or socio-demographic variables was investigated in 74% of these reviews and test- or threshold-related variables in 79%. At least one quality-related variable was investigated in 63% of reviews. Within this subset, the most commonly considered variables were the use of blinding (41% of reviews), sample size (33%), the reference test used (28%) and the avoidance of verification bias (25%).

Conclusions

The emphasis on pooling individual aspects of diagnostic test performance and the under-use of statistical tests and graphical approaches to identify heterogeneity perhaps reflect the uncertainty in the most appropriate methods to use and also greater familiarity with more traditional indices of test accuracy. This is an indication of the level of difficulty and complexity of carrying out these reviews. It is strongly suggested that in such reviews meta-analyses are carried out with the involvement of a statistician familiar with the field.

Recommendations for further research

The following areas are suggested for further research.

- Further methodological work on the statistical methods available for combining diagnostic test accuracy studies is needed.
- Sufficiently large, prospectively designed primary studies of diagnostic test accuracy that compare two or more tests for the same target disorder are needed so that sources of heterogeneity are minimised and comparative accuracy can be established in a wide spectrum of patients.
- Use of individual patient data meta-analysis in diagnostic test accuracy reviews should be explored to allow heterogeneity to be considered in more detail.

Publication

Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess* 2005;**9**(12).

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 by the HTA Programme as project number 02/31/01. As funder, by devising a commissioning brief, the HTA Programme specified the research question and study design. 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 or the Department of Health.

ISSN 1366-5278

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.