

Systematic reviews of diagnostic tests in cancer: review of methods and reporting

Susan Mallett, Jonathan J Deeks, Steve Halligan, Sally Hopewell, Victoria Cornelius, Douglas G Altman

Abstract

Objectives To assess the methods and reporting of systematic reviews of diagnostic tests.

Data sources Systematic searches of Medline, Embase, and five other databases identified reviews of tests used in patients with cancer. Of these, 89 satisfied our inclusion criteria of reporting accuracy of the test compared with a reference test, including an electronic search, and published since 1990.

Review methods All reviews were assessed for methods and reporting of objectives, search strategy, participants, clinical setting, index and reference tests, study design, study results, graphs, meta-analysis, quality, bias, and procedures in the review. We assessed 25 randomly selected reviews in more detail.

Results 75% (67) of the reviews stated inclusion criteria, 49% (44) tabulated characteristics of included studies, 40% (36) reported details of study design, 17% (15) reported on the clinical setting, 17% (15) reported on the severity of disease in participants, and 49% (44) reported on whether the tumours were primary, metastatic, or recurrent. Of the 25 reviews assessed in detail, 68% (17) stated the reference standard used in the review, 36% (9) reported the definition of a positive result for the index test, and 56% (14) reported sensitivity, specificity, and sample sizes for individual studies. Of the 89 reviews, 61% (54) attempted to formally synthesise results of the studies and 32% (29) reported formal assessments of study quality.

Conclusions Reliability and relevance of current systematic reviews of diagnostic tests is compromised by poor reporting and review methods.

Introduction

Diagnostic accuracy is essential for good therapeutic treatment. The case for systematic reviews is now well established, enabling efficient integration of current information and providing a basis for rational decision making.¹ The methods used to conduct systematic reviews of diagnostic tests, however, are still developing.

Good methods and reporting are essential for reviews to be reliable, transparent, and relevant. For

example, systematic reviews need to report results from all included studies, with information on study design, methods, and characteristics that may affect clinical applicability, generalisability, and potential for bias.

Systematic reviews of diagnostic studies involve additional challenges to those of therapeutic studies.²⁻³ Studies are observational in nature, prone to various biases,⁴ and report two linked measures summarising the performance in participants with disease (sensitivity) and without (specificity). In addition, there is more variation between studies in the methods, manufacturers, procedures, and outcome measurement scales used to assess test accuracy⁵ than in randomised controlled trials, which generally causes marked heterogeneity in results.

Researchers have found evidence for bias related to specific design features of primary studies of diagnostic studies.⁶⁻⁷ There was evidence of bias when primary studies did not provide an adequate description of either the diagnostic (index) test or the patients, when different reference tests were used for positive and negative index tests, or when a case-control design was used.

We assessed the reliability, transparency, and relevance of published systematic reviews of evaluations of diagnostic tests in cancer with an emphasis on methods and reporting.

Methods

Literature search—Systematic literature searches used Medline, Embase, MEDION, Cancerlit, HTA, and DARE databases and the Cochrane Database of Systematic Reviews, from 1990 to August 2003. Additional searches included bibliographies of retrieved reviews and clinical guidelines for cancer identified from the web.

Inclusion criteria—We included reviews if they assessed a diagnostic test for presence or absence of

Editorial by Straus

Centre for Statistics in Medicine, University of Oxford, Wolfson College, Oxford OX2 6UD

Susan Mallett
medical statistician
Douglas Altman
professor of statistics in medicine

Department of Public Health and Epidemiology, University of Birmingham, Edgbaston, Birmingham B15 2TT

Jonathan Deeks
professor of health statistics

UK Cochrane Centre, Oxford OX2 7LG

Sally Hopewell
research scientist

Department of Specialist Radiology, University College London, London NW1 2BU

Steve Halligan
professor of gastrointestinal radiology

Drug Safety Research Unit, Southampton SO31 1AA

Victoria Cornelius
statistician

Correspondence to: Susan Mallett
susan.mallett@cancer.org.uk

BMJ 2006;333:413-6



A list of the reviews assessed in detail is on bmj.com.



This is the abridged version of an article that was posted on bmj.com on 18 July 2006: <http://bmj.com/cgi/doi/10.1136/bmj.38895.467130.55>

cancer or staging of cancer including metastasis and reoccurrence; reported accuracy of the test assessed by comparison to reference tests; reported an electronic search and listed references for included studies; and were published from 1990 onwards.

Sample selection—We assessed all identified reviews generally and selected a random sample of 30 reviews stratified by the type of index test for more detailed assessment. In five reviews, however, the number of included studies was unclear, so 25 reviews were assessed in detail (see bmj.com).

Validity assessment and data abstraction—We assessed the methods and reporting of each review across nine domains—review objectives and search strategy, participants and clinical setting, index test, reference test, study design, study results, graphs and meta-analysis, quality and bias, and procedures used in the review—guided by previous publications.^{6 8–15} One reviewer (SM) undertook the general assessments. In the detailed assessments, two independent assessors extracted data from each review and reached a consensus by agreement or by reference to a third party. Our results evaluate the methods and reporting of the review. Primary diagnostic studies are often poorly reported so when authors of reviews said they had sought but not found information in the included studies, we counted this as reported.

Quality score—A quality score was produced for each of the nine domains by counting question responses judged to indicate a better review. For each review, we calculated a percentage of the maximum score for each domain and plotted the data as a star plot in Stata 8.0 (StataCorp, College Station, TX). We analysed the quality of the review according to the study objective, page length, year of publication, number of diseases, number of tests, and whether the test was an imaging technology.

Results

The table summarises the characteristics of the reviews. The reviews covered a range of types of diagnostic tests and tumour sites. We could not assess five of the 30 reviews assigned for detailed review because the number of studies included in the review was unclear. Details of findings across the nine assessment domains are on bmj.com. Average agreement between duplicate data extractors was 80%, most differences occurring through reader error or from ambiguity in the reviews, particularly for the details of the reference test.

Objectives, inclusion criteria, and search—The primary purpose of most reviews was to assess test accuracy; some did so as part of a clinical guideline or economic evaluation (see bmj.com). Three quarters (67/89) of the reviews stated inclusion criteria, though the number of studies included was unclear in 15 reviews. Nearly a third (32%, 8/25) of reviews searched two or more electronic databases, 80% reported their search terms, and 84% searched bibliography lists or other non-electronic sources.

Description of target condition, patients, and clinical setting—Half of the 89 reviews did not report whether tumours were primary, recurrent, or metastatic. Only 17% (15/89) reported clinical setting, and 45%

Characteristics of included reviews (n=89)

Topic	Percentage (No) of reviews
Imaging tests*:	
PET†	20 (18)
MRI†	19 (17)
CT†	26 (23)
Other imaging	45 (40)
Non-imaging tests*:	
Laboratory test	22 (20)
Pathology/cytology	24 (21)
Clinical exam	20 (18)
More than one disease	74 (66)
Primary tumour site:	
Bone and soft tissue	5 (4)
Breast	16 (14)
Cervix	3 (3)
Colorectal	8 (7)
Endocrine	3 (3)
Endometrial	8 (7)
Head and neck	2 (2)
Lung	12 (11)
Ovarian	2 (2)
Prostate	11 (10)
Skin	121 (11)
Upper GI	7 (6)
Urological	6 (5)
More than one site	5 (4)

GI=gastrointestinal; PET=positron emission tomography; MRI=magnetic resonance imaging; CT=computed tomography.

*Reviews can contain more than one test.

†Three assay types grouped for stratified random sampling.

reported characteristics of patients for individual studies. Of 17 reviews of primary or recurrent tumours assessed in detail, 10 did not consider possible effects of tumour stage or grade on test performance. Eighteen percent (16/89) of reviews collected information on the severity of disease but did not report it.

Study design—Twenty of the 25 reviews assessed in detail did not report or were unclear on whether included studies used consecutive recruitment of patients. Few reviews limited inclusion to study designs less prone to bias—namely, consecutive (8%, n=2) or prospective (12%, n=3) studies. Sixty percent (15) of reviews discussed test masking.

Description of index and reference tests—Only 36% (9/25) of reviews reported the definition of a positive result for the index test. In 40% (10/25) of reviews it was not clear if the included studies used the same, or different, index tests or procedures. Of reviews assessed in detail, 68% (17/25) reported the reference tests used in the review; 40% reported reference tests for each included study.

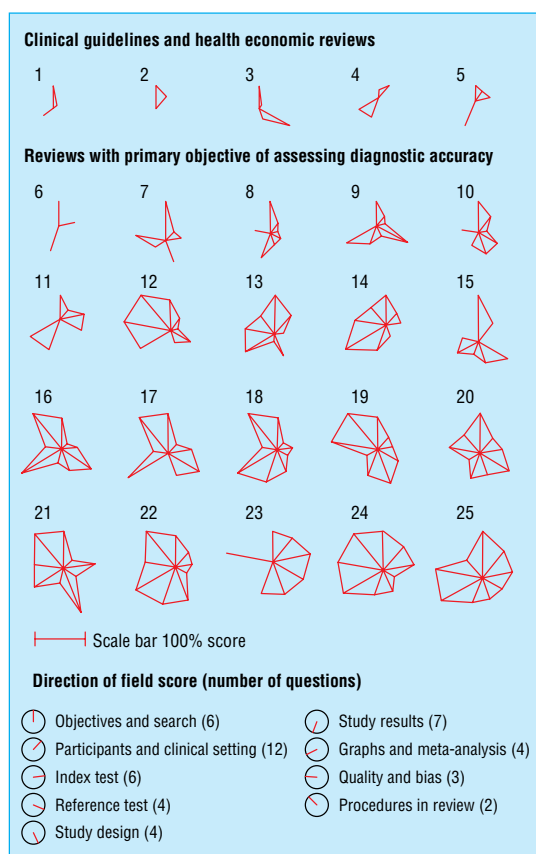
Reporting of individual study results and graphical presentation—Of the 89 reviews, 40% contained graphs of study findings, and 39% reported sensitivities and specificities, likelihoods ratios, or predictive values. Of reviews assessed in detail, 56% (14/25) provided adequate information to derive 2×2 tables for all included studies.

Meta-analysis, quality, and bias—Sixty one percent (54/89) of reviews presented a meta-analysis and 32% completed a formal assessment of quality. Twenty three of the 25 reviews assessed in detail discussed the potential for bias. Spectrum bias was most commonly

considered (80% of reviews), with verification bias and publication bias considered least (40%).

Procedures in review—Only 48% (12/25) of reviews provided information on review procedures, most reporting duplicate data extraction by two assessors (nine reviews), a method recommended to increase review reliability.

Assessment of overall review quality—The figure shows quality scores for each domain assessed by using star plots for the 25 reviews assessed in detail. Reviews of higher quality have longer spokes and larger areas within the stars. Reviews conducted for the three clinical guidelines and two health economic analyses were of particularly poor quality. Additional detailed assessment of seven further reviews of clinical practice guidelines included in our larger sample confirmed this pattern: four did not report the number of included studies, and the three remaining were of similar quality to the five in the figure. We identified two reviews with good overall methods and reporting that could serve as examples for new reviewers.^{16 17} Study quality was not related to page length, year of publication, assessment of an imaging technology, or the number of diseases or index tests assessed.



Star plots of methods and reporting quality of reviews. Each review assessed in detail is represented by a star plot of nine domains, indicating the percentage of a maximum score in each domain, with domain scores indicated by clockface directions. A review of high quality in all areas would correspond to a nonagon with all spokes at maximum length. The number of questions contributing to each domain score is listed in the key, with a scale bar. Reviews are ordered by primary objective of review to assess accuracy (or not) of diagnostic test, and within this by total quality score

Discussion

This review of reviews of diagnostic tests in cancer has highlighted the poor quality of the literature. Many reviews did not use systematic methods (37% of otherwise eligible reviews did not report an electronic search) and poor reporting was common (32% did not state the reference test used, 83% did not state the severity of the disease). The execution and reporting of systematic reviews of diagnostic tests clearly need to be improved.

Our assessment was based on all reviews we could locate of tests for cancer published between 1990 and 2003. The reliability of our assessment was good based on the high level of agreement (80%, interquartile range 72%-91%) between the two independent assessors of the detailed reviews. Few of our reviews contained large numbers of primary studies. In some specialties reviews may include 100 or more studies, making it difficult to report full information because of page limitations for journal articles. Creative use of appendix tables on journal or investigator websites should be considered. The forthcoming publication of Cochrane Reviews of Test Accuracy will also help remedy this challenge.¹⁸

Other surveys of systematic reviews have found similar problems. In a review of meta-analyses of diagnostic tests across all specialties,⁶ Lijmer et al found that a systematic search was not reported in seven of 26 reviews. Dinnes et al found 51% of reviews listed more than one reference test.⁵ (Our figure of 53% may be an underestimate as 33% of reviews were either unclear or did not report on the reference test clearly enough to examine this question.)

Reporting of details of primary studies

Interestingly, Arroll et al found that 87% of primary diagnostic studies clearly defined positive and negative test results.¹⁰ Only 40% of reviews in our study reported a definition of positive test results or reported that it was not available in the primary studies. It seems likely that key information available in primary studies is being omitted from systematic reviews.

Transparent reporting of review methods and detailed reporting of the clinical and methodological characteristics of the included studies and their results is important to enable a reader to judge the reliability of both the review and the individual studies and to assess their relevance to clinical practice and the meaning of the results reported in the review. A lack of awareness of the complexities within diagnostic studies may have led to under-reporting of critical detail of review methods and included study characteristics.

Test methods and materials often vary between studies for both reference and index tests, but many reviews do not give details for each study. The population of patients being studied by the included studies varied so much that often different diseases were mixed together within a review.

Previous research in diagnostic studies has shown that case-control designs and non-consecutive recruitment of patients can lead to bias.^{6 7} Whether consecutive recruitment was used in primary studies was not reported or was unclear in 80% of our reviews.

What is already known on this topic

Systematic reviews of randomised controlled trials are an established way of efficiently summarising multiple studies to provide an easily accessible evidence base for making decisions about healthcare interventions

In recent years many journals have published systematic reviews on accuracy of diagnostic tests, but the quality and usefulness of these reviews has not been systematically assessed

What this study adds

The reliability and clinical relevance of published systematic reviews of diagnostic tests are compromised by poor review methods and poor reporting

Systematic reviews of diagnostic tests require improved reporting of detailed information about the design, conduct, and results of the included primary studies, as well as review methods, as will be required in the forthcoming Cochrane Reviews of Test Accuracy

Selection bias, however, was discussed in 14 reviews, 10 of which did not report or were unclear about the method of selection of patients. So, though many reviews discussed different types of bias, they did not always provide the information that would enable a reader to assess the risk of bias.

In our sample we found the quality of reviews completed for the purpose of clinical guidelines was poor, with worrying implications if these are the reviews guiding clinical practice. Reviews of diagnostic tests would be better carried out separately from the preparation of clinical guidelines.

Conclusions

Systematic reviews of diagnostic tests are complex and require reporting of detailed information about the design, conduct, and results of the included primary studies to ensure reviews are useful. We have shown the current poor quality of published reviews and indicated areas for improvement.

Contributors: See bmj.com.

Funding: SM, DGA, and VC are funded by Cancer Research UK. JJD is partially funded by a senior research fellowship in evidence synthesis from the UK Department of Health NCCRC (National Coordinating Centre for Research Capacity Development). SHo is funded from the NHS research and development programme.

Competing interests: None declared.

Ethical approval: Not required.

- Mulrow CD. Rationale for systematic reviews. *BMJ* 1994;309:597-9.
- Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323:157-62.
- Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, et al. Challenges in systematic reviews of diagnostic technologies. *Ann Intern Med* 2005;142:1048-55.
- Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;6:411-23.
- 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;9:1-128.

- Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
- Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76.
- Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;120:667-76.
- Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol* 2005;58:1-12.
- Arroll B, Schechter MT, Sheps SB. The assessment of diagnostic tests: a comparison of medical literature in 1982 and 1985. *J Gen Intern Med* 1988;3:443-7.
- Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. *JAMA* 1984;252:2418-22.
- Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002;2:9.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;138:W1-12.
- Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet* 1999;354:1896-900.
- Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.
- Gould MK, Macean CC, Kushner WG, Rydzak CE, Owens DK. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *JAMA* 2001;285:914-24.
- Harris KM, Kelly S, Berry E, Hutton J, Roderick P, Cullingworth J, et al. Systematic review of endoscopic ultrasound in gastro-oesophageal cancer. *Health Technol Assess* 1998;2:1-134.
- Deeks J, Gatsonis C, Bossuyt P, Antes G. Cochrane reviews of diagnostic test accuracy. *Cochrane News* 2004;31:Aug 2004. www.cochrane.org/newslett/ccnews31-lowres.pdf (accessed 31 May 2006).

(Accepted 31 May 2006)

doi 10.1136/bmj.38895.467130.55

Corrections and clarifications

Regulation and revalidation of doctors

Some readers might have been misled by the subtitle we added to this editorial by Mike Pringle (*BMJ* 2006;333:161-2, 22 Jul). The subtitle "England's chief medical officer's report should resolve the uncertainty" might suggest that the report (by Sir Liam Donaldson) related only to England. This is not the case. Professor Donaldson is indeed the chief medical officer for England, but the report (and the editorial) concerned medical regulation throughout the United Kingdom (the General Medical Council is the regulatory body and covers all UK countries). The same lack of clarity was evident in the first news article, by Andrew Cole, in the same issue (p 163).

Variant Creutzfeldt-Jakob disease: prion protein genotype analysis of positive appendix tissue samples from a retrospective prevalence study

An error in the electronic processing of this paper by James W Ironside and colleagues resulted in the second part of his email address being omitted (*BMJ* 2006;332:1186-8, 20 May). Correspondence about this paper should be emailed to james.ironside@ed.ac.uk.

A bipolar story

A technical editor's fumble fingered typing led to Raquel Duarte, the author of this filler (*BMJ* 2006;333:245, 29 Jul), being given an incorrect email address. Her correct address is s0126305@sms.ed.ac.uk.