Cite this article as: BMJ, doi:10.1136/bmj.38895.467130.55 (published 18 July 2006)

BMJ

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 accuracy[5] 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.

Previous research on systematic reviews of diagnostic tests noted poor methods and reporting. Irwig et al reviewed 11 meta-analyses published in 1990-1 and drew up guidelines to address key areas where reviews were deficient.[8] Schmid et al reported preliminary results on methods used for search strategies and meta-analysis in 189 systematic reviews,[9] and Whiting et al reported on the extent of quality assessment within diagnostic reviews.[10] Other research has focused on the methods of primary studies.[6 11–16]

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. We used three search strings: the Cochrane Cancer Network string to identify cancer studies[17]; a search string optimised for diagnostic studies[18]; and search strings to identify systematic reviews and meta-analyses ('meta?analysis' and the Medline systematic review filter[19]).

Inclusion criteria

Reviews were included if they assessed a diagnostic test for presence or absence of cancer or staging of cancer including metastasis and reoccurrence (screening tests and tests for risk factors for cancer such as human papillomavirus were excluded); 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. Studies limited to methods of sample collection or

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

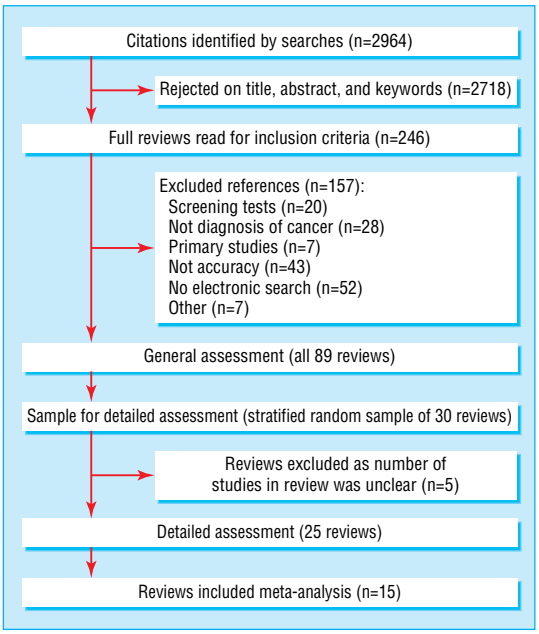

Fig 1 Flowchart of reviews

computer decision tools were excluded. English, French, and Italian reviews were included. Reviews in other languages were included when a translation was available from authors.

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 (fig 1).

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 10 11 13 18 20–22] In the general assessment of all reviews we evaluated 34 items at the review and study level. In the detailed assessment we evaluated 98 items examining questions at review, study, and individual test level. For the detailed assessment, if there was more than one index test, we selected the test reported in the largest number of studies (or the test listed first in the title or text of the review when there were an equal number of studies). Copies of forms are available on request.

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. SM, VC, SHa, SHo were the assessors. 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. Subsequently, SM assessed all seven additional reviews conducted as part of clinical guidelines in detail to determine whether the three clinical guideline reviews included in our sample of 25 reviews were typical.

Agreement between assessors

We used percentage agreement for each data item to assess reliability between raters in the detailed assessment where duplicate data extraction was undertaken. The agreement between assessors was too high for κ scores to be informative. We calculated the average agreement across reviews for individual data questions, and, when it was below 75%, we assessed and reported the reason for disagreement.

Table 1 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	12 (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.

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

Figure 1 shows the 89 reviews that met the inclusion criteria. Table 1 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. Tables 2-6 show the findings across the nine assessment domains. Items are classified according to whether they relate to the review, a single test within the review, or a single study within the review. 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

A clear statement of objectives and inclusion criteria for the review are important for a systematic approach.[23] Only when

Table 2 Assessment of reviews of diagnostic tests in patients with cancer, according to objectives and search, and participants and clinical setting

Assessment item	Percentage (No) of reviews
Objectives and search (89 reviews)	
Objectives*:	
To review accuracy of diagnostic test	80 (71)
Other objectives:	
Clinical guidelines	11 (10)
General overview of disease	7 (6)
Health economic study	2 (2)
Inclusion criteria of review stated*	75 (67)
Table of study characteristics included*	49 (44)
Participants and clinical setting (89 reviews)	
Type of tumour*:	
Not reported or unclear in some studies	51 (45)
Primary tumour only	6 (5)
Metastatic tumour only	13 (12)
Recurrent only	1 (1)
Mix of types	29 (26)
Clinical setting stated*	17 (15)
Details of patient characteristics for individual studies†:	
Reported	45 (40)
Information extracted but not reported	21 (19)
Disease severity in patients for individual studies, or grade†:	
Reported	17 (15)
Information extracted but not reported	18 (16)
Not applicable (for example, metastatic tumour)	16 (14)
Participants and clinical setting (detailed assessment, 25 reviews)	
Patient demographics reported†:	
Age	24 (6)
Sex	20 (5)
Sex not applicable (single sex disease)	36 (9)
Country	12 (3)

*Relates to review.
†Relates to individual studies.

search strategies are reported can readers of the review appraise how well the review has avoided bias in locating studies.

The primary purpose of most reviews was to assess test accuracy; some did so as part of a clinical guideline or economic evaluation (table 2). Three quarters of the 89 reviews stated inclusion criteria, though the number of studies included was unclear in 15 reviews. Of the 25 reviews assessed in detail, 16 used study inclusion criteria relating to sample size or study design, and 15 discussed the appropriateness of patient inclusion criteria used by the primary studies. Nearly a third (32%, 8/25) of the 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

Clinical relevance and reliability requires reporting of information on the target condition, patients, and clinical setting.[22] Reporting severity of disease is important because, for example, the performance of many imaging techniques is related to tumour size.

Half of the 89 reviews did not report whether tumours were primary, recurrent, or metastatic (table 2). Only 17% (15/89) reported on the clinical setting, and 45% 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. Reviews sometimes omitted information that had been collected—for example, 18% (16/89) of reviews collected information on the severity of disease but did not report it.

Table 3 Assessment of reviews of diagnostic tests in patients with cancer, according to study design

Assessment item	Percentage (No) of reviews
Study design (89 reviews)	
Details of individual design assessed for individual studies (any of prospective/retrospective, consecutive/case-control, masking)*:	
Reported at least one aspect per study	40 (36)
Extracted but not reported at least one aspect per study	27 (24)
Study design (detailed assessment, 25 reviews)	
Reporting of consecutive/non-consecutive study design†:	
All study designs are not reported or unclear	80 (20)
Consecutive studies only	8 (2)
Randomised design only	0
Mix of consecutive, non-consecutive and random	12 (3)
Reporting of prospective/retrospective study design†:	
Prospective studies only	12 (3)
Report study designs include prospective and/or retrospective	64 (16)
Some study designs are not reported or unclear‡	20 (5)
All study designs are not reported or unclear	36 (9)
Test masking§:	
Test masking discussed in review	60 (15)
Type of test masking discussed:	
Masking both ways between reference and index test	24 (6)
Index test masked to reference test	24 (6)
Reference test masked to index test	0)
Other (between two index tests or unspecified)	12 (3)

*Relates to individual studies.
†Relates to single test within review.
‡May include some tests where a mixture of prospective and retrospective studies are used.
§Relates to review.

Study design

Consecutive prospective recruitment from a clinically relevant population of patients with masked assessment of index and reference tests is the recommended design to minimise bias and ensure clinical applicability of study results.

Twenty of the 25 reviews assessed in detail did not report or were unclear on whether included studies used consecutive recruitment of patients (table 3). Few reviews limited inclusion to study designs less prone to bias—namely, consecutive (8%) or prospective (12%) studies. Sixty percent (15/25) discussed test masking. Poor reporting made it impossible to identify inclusion of case-control designs.[6]

Description of index and reference tests

Both index and reference tests need to be clearly described for a review to be clinically relevant and transparent and to allow readers to judge the potential for verification and incorporation biases.[6]

Only 36% (9/25) of reviews reported the definition of a positive result for the index test (table 4). In 40% (10/25) it was unclear if the included studies used the same, or different, index tests or procedures. When index tests were reported to vary between included studies, 71% (10/14) reported the index test for each study and the compatibility of different tests was discussed in 86% (12/14) of reviews.

Sixty eight percent (17/25) of reviews assessed in detail reported the reference tests used in the review; 40% reported reference tests for each included study. Six reviews reported whether reference tests were used on all, a random sample, or a select sample of patients.

Reporting of individual study results and graphical presentation

We assessed the level of detail used to report the results of individual studies. Ideally reviews should report data from 2×2 tables

Table 4 Assessment of reviews of diagnostic tests in patients with cancer, according to characteristics of index and reference test

Assessment item	Percentage (No) of reviews
Index test (89 reviews)	
Index tests*:	
Single index test reported in review	36 (32)
Multiple index tests reported in review	57 (51)
Index test(s) not reported or unclear	7 (6)
Reference test (89 reviews)	
Reference tests*:	
Single reference test reported in review	14 (13)
Multiple reference tests reported in review	53 (47)
Reference test(s) not reported or unclear	33 (29)
Index test (detailed assessment, 25 reviews)	
Definition of positive test result given†	36 (9)
Index test reported for each study‡	68 (17)
Time period between index test and reference test reported†	28 (7)
Uninterpretable test results reported†	12 (3)
Different index test procedures or manufacturers†:	
Single index test reported	4 (1)
Multiple index test procedures reported	56 (14)
Not reported or unclear	40 (10)
Single scale for test results†:	
Reported	16 (4)
Not reported or unclear	60 (15)
Multiple scales	24 (6)
Reference test (detailed assessment, 25 reviews)	
Reference test(s) used in review reported†	68 (17)
Reference test for each study‡:	
Reported	40 (10)
Extracted but not reported	40 (10)
Not reported or unclear	20 (5)
Reference test used on all or randomly selected patients†	24 (6)
Single reference test used for index test†:	
Reported	12 (3)
Not reported or unclear	32 (8)
More than one reference test used	56 (14)
Single test method, manufacturer, and results scale†:	
Reported	4 (1)
Not reported or unclear	40 (10)
Multiple test methods	56 (14)

*Relates to review.
†Relates to single test within review.
‡Relates to individual studies

for each study, or summary statistics of test performance. Graphs are efficient tools for reporting results and depicting variability between study results. Of the 89 reviews, 40% contained graphs of study findings, and 39% reported sensitivities and specificities, likelihoods ratios, or predictive values (table 5). Over half (56%, 14/25) of the reviews assessed in detail provided adequate information to derive 2×2 tables for all included studies. Four reviews included tests with continuous outcomes but presented only dichotomised results; three reported the cutpoint used.

Meta-analysis, quality, and bias

Appropriate use of meta-analysis can effectively summarise data across studies.[24] Quality assessment is important to give readers an indication of the degree to which included studies are prone to bias.

Sixty one percent (54/89) of reviews presented a meta-analysis (table 6) 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%).

Table 5 Assessment of reviews of diagnostic tests in patients with cancer according to graphical display and study results

Assessment item	Percentage (No) of reviews
Graphical display of data (89 reviews)	
Graphs*:	
No graph	60 (53)
Any graph	40 (36)
Type of graph:	
Summary ROC graph*	34 (30)
Forest plot of sensitivity/specificity*	3 (3)
Other plot*	16 (14)
Study results (89 reviews)	
Study results reported for all studies†:	
Sensitivity	52 (46)
Specificity	37 (33)
Two measures of test accuracy	39 (35)
Samples sizes	65 (58)
Prevalence	46 (41)
Study results (detailed assessment, 25 reviews)	
2×2 table results for each study†:	
Reported or can be calculated from review	56 (14)
Extracted but not reported	32 (8)
Not extracted or unclear	12 (3)
Report results on continuous scale†	0
Confidence interval or SE for individual studies	28 (7)

ROC=receiver operating characteristic curve; SE=standard error.
*Relates to review; reviews can have more than one graph.
†Relates to individual studies.

Procedures in review

The reliability of a review depends partly on how it was done.[23] 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

Figure 2 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 figure 2.

We identified two reviews with good overall methods and reporting that could serve as examples for new reviewers.[25 26] 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.

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 agree-

Table 6 Assessment of reviews of diagnostic tests in patients with cancer according to meta-analysis, quality, and bias. Figures are percentage (number) of reviews or studies that had the assessment item

Assessment item	Percentage (No) of reviews
Meta-analysis (89 reviews)	
Meta-analysis included*	61 (54)
Quality (89 reviews)	
Quality of included studies assessed*	58 (52)
Formal assessment	32 (29)
Discussion only	26 (23)
Quality (25 reviews)	
Quality of included studies assessed†	52 (13)
Published measure only	8 (2)
Own measure only	20 (5)
Own measure and published measure	24 (6)
Bias (25 reviews)	
Discussion of bias*	92 (23)
Type of bias discussed:	
Publication	40 (10)
Verification	40 (10)
Observation	56 (14)
Selection	56 (14)
Spectrum	80 (20)
Reproducibility	52 (13)
Median No of biases discussed per review	4
Meta-analysis (15 reviews)	
Method of meta-analysis†‡:	
Pool accuracy measures separately:	
Sensitivity/specificity	33 (5)
Likelihood ratios	7 (1)
Summary ROC	54 (8)
Other	20 (3)
Confidence intervals for meta-analysis†	100 (15)
MA weighting†‡:	
Adjusted	66 (10)
Not adjusted	13 (2)
Both of above	13 (2)
Not reported/unclear method	33 (5)
No of patients in MA can be calculated†	93 (14)

MA=meta-analysis.
*Relates to review.
†Relates to single test within review.
‡Two reviews used more than one method.

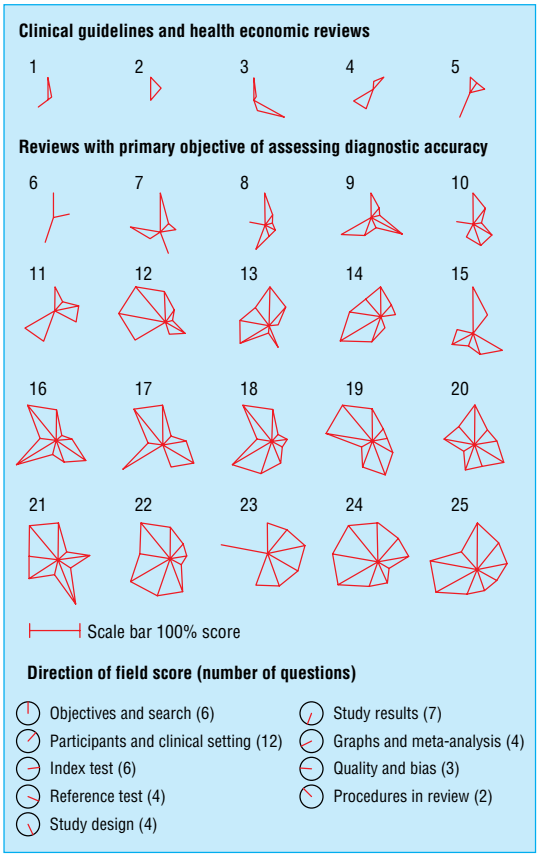

Fig 2 Star plots of methods and reporting quality of reviews. Each review assessed in detail is represented by a star plot of nine domains, indicating the percentage of a maximum score in each domain, with domain scores indicated by clockface directions. A review of high quality in all areas would correspond 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

ment (80%, interquartile range 72%-91%) between the two independent assessors of the detailed reviews. While we assessed only a sample of 30 reviews in detail, all key points were covered and similar assessment of the remaining reviews would not importantly add to this research.

Though this research relates to reviews on the diagnosis of cancer, we believe that the results are likely to be typical of other specialties. Most of our assessment questions could be applied in any medical topic. Similar problems have been found in other medical topics for systematic reviews[5 8 11] and primary studies.[10 12] We also found that the types of tests used in cancer were similar to those in a recent survey of meta-analyses of diagnostic tests of all specialties, where about half of diagnostic tests were imaging tests, one third were laboratory tests, and the rest were clinical tests.[8] The treatment of the variability between studies, and the assessment of how much of this is due to heterogeneity rather than variation due to chance alone, is a complex issue and findings will be presented elsewhere.

Reporting in 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.[27]

Other surveys of systematic reviews have found a similar prevalence of reporting 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.

Other research has also found variation of methods within 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 primary study details

Interestingly, Arroll et al found that 87% of primary diagnostic studies clearly defined positive and negative test results.[11] 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 are 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. Index tests would probably have a different accuracy in patients with primary and metastatic tumours. At least 20% of reviews estimated accuracy of a specific test with a mixture of studies covering patients with primary and metastatic tumours, and in a further 48% of reviews the type of tumour was unclear.

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. 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.

We found that only 12% of reviews reported that only one reference test was used, indicating the potential for heterogeneity due to difference reference tests in a large proportion of reviews. Nearly three quarters of reviews did not report the time interval between index and reference tests, so the extent of bias due to disease progression in our reviews was unclear. The time interval between reference and index test was reported only in reviews that used a specified time of follow-up for the reference test. In these reviews, the time between index and reference test was several years, giving concern about interval cancer cases. Previous studies have found that follow-up is used as reference in 11% (21/189) of all reviews but is used in 24% of cancer reviews (10/42).[5]

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: SHa, SHo, and VC completed duplicate data extraction to SM for a third of the data each. SM, JJD, and DGA contributed to analysis and presentation of data and drafting the article. SHo and SHa contributed to editing. SM and DGA are guarantors.

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 NCCRCD (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.

1 Mulrow CD. Rationale for systematic reviews. *BMJ* 1994;309:597-9.
2 Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323:157-62.
3 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.
4 Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;6:411-23.
5 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.
6 Lijmer JG, Mol BW, Heisterkamp S, Bonsel 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.
7 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.
8 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.
9 Schmid C, Chung M, Chew P, Lau J. Survey of diagnostic test meta-analyses. In: 12th Cochrane Colloquium, Ottawa, 2-6th Oct 2004.
10 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.
11 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.
12 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;274:645-51.
13 Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. *JAMA* 1984;252:2418-22.
14 Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, et al. Quality of reporting of diagnostic accuracy studies. *Radiology* 2005;235:347-53.
15 Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess* 2004;8:1-234.
16 Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189-202.
17 Lodge M. The Cochrane cancer network. *Cochrane Library*. 2005, Issue 1.
18 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.
19 Shojania KG, Bero LA. Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy. *Eff Clin Pract* 2001;4:157-62.
20 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.
21 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.
22 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.
23 Oxman AD, Guyatt GH. The science of reviewing research. *Ann NY Acad Sc* 1993;703:125-33.
24 Egger M, Ebrahim S, Smith GD. Where now for meta-analysis? *Int J Epidemiol* 2002;31:1-5.
25 Gould MK, Maclean CC, Kuschner 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.
26 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.
27 Deeks J, Gatsonis C, Bossuyt P, Antes G. Cochrane reviews of diagnostic test accuracy. *Cochrane News*. Issue 31;Aug 2004. www.cochrane.org/newslett/ccnews31-lowres.pdf (accessed 31 May 2006).

(Accepted 31 May 2006)

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 reporting

Systematic reviews of diagnostic tests require 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

doi 10.1136/bmj.38895.467130.55

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