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 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. 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. Studies are observational in nature, prone to various biases, 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 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. 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. Schmid et al reported preliminary results on methods used for search strategies and meta-analysis in 189 systematic reviews, and Whiting et al reported on the extent of quality assessment within diagnostic reviews. Other research has focused on the methods of primary studies.

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; a search string optimised for diagnostic studies; and search strings to identify systematic reviews and meta-analyses (‘meta?analysis’ and the Medline systematic review filter).

Inclusion criteria
Reviews were included if they assessed a diagnostic test for presence or absence of cancer or staging of cancer including metastasis and recurrence (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...
procedures used in the review—guided by previous study results, graphs and meta-analysis, quality and bias, and pants and clinical setting, index test, reference test, study design, validity assessment and data abstraction

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).

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.

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 $k$ 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.

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. Only when
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. 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.

### 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 case-control, masking. Reporting of prospective/retrospective study design was poor: only 12% (3/25) of reviews reported use of prospective design exclusively.

### 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.

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
for each study, or summary statistics of test performance. Graphs are efficient tools for summarizing results and depicting variability between study results. Of the 89 reviews, 40% contained graphs of study findings, and 39% reported sensitivities and specificities, likelihood 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 dichotomized results; three reported the cutpoint used.

Meta-analysis, quality, and bias
Appropriate use of meta-analysis can effectively summarise data across studies. 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%).

Procedures in review
The reliability of a review depends partly on how it was done.25 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, or reporting that could serve as examples for new reviewers.25 26

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 (57% 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-
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. 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 finding solutions will be presented elsewhere.

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

<table>
<thead>
<tr>
<th>Assessment item</th>
<th>Percentage (No) of reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-analysis (69 reviews)</td>
<td>61 (54)</td>
</tr>
<tr>
<td>Meta-analysis included‡</td>
<td>61 (54)</td>
</tr>
<tr>
<td>Quality (56 reviews)</td>
<td>58 (52)</td>
</tr>
<tr>
<td>Quality of included studies assessed*</td>
<td>58 (52)</td>
</tr>
<tr>
<td>Formal assessment</td>
<td>32 (29)</td>
</tr>
<tr>
<td>Discussion only</td>
<td>26 (23)</td>
</tr>
<tr>
<td>Quality (25 reviews)</td>
<td>52 (13)</td>
</tr>
<tr>
<td>Quality of included studies assessed†</td>
<td>52 (13)</td>
</tr>
<tr>
<td>Published measure only</td>
<td>8 (2)</td>
</tr>
<tr>
<td>Own measure only</td>
<td>20 (5)</td>
</tr>
<tr>
<td>Own measure and published measure</td>
<td>24 (6)</td>
</tr>
<tr>
<td>Bias (25 reviews)</td>
<td>92 (23)</td>
</tr>
<tr>
<td>Type of bias discussed:</td>
<td>92 (23)</td>
</tr>
<tr>
<td>Publication</td>
<td>40 (10)</td>
</tr>
<tr>
<td>Verification</td>
<td>40 (10)</td>
</tr>
<tr>
<td>Observation</td>
<td>56 (14)</td>
</tr>
<tr>
<td>Selection</td>
<td>56 (14)</td>
</tr>
<tr>
<td>Spectrum</td>
<td>80 (20)</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>52 (13)</td>
</tr>
<tr>
<td>Median No of biases discussed per review</td>
<td>4</td>
</tr>
<tr>
<td>Meta-analysis (15 reviews)</td>
<td>100 (15)</td>
</tr>
<tr>
<td>Method of meta-analysis‡</td>
<td>100 (15)</td>
</tr>
<tr>
<td>Pool accuracy measures separately:</td>
<td>100 (15)</td>
</tr>
<tr>
<td>Sensitivity/specificity</td>
<td>33 (5)</td>
</tr>
<tr>
<td>Likelihood ratios</td>
<td>7 (1)</td>
</tr>
<tr>
<td>Summary ROC</td>
<td>54 (8)</td>
</tr>
<tr>
<td>Other</td>
<td>20 (3)</td>
</tr>
<tr>
<td>Confidence intervals for meta-analysis†</td>
<td>66 (10)</td>
</tr>
<tr>
<td>MA weighting†‡</td>
<td>66 (10)</td>
</tr>
<tr>
<td>Adjusted</td>
<td>66 (10)</td>
</tr>
<tr>
<td>Not adjusted</td>
<td>13 (2)</td>
</tr>
<tr>
<td>Both of above</td>
<td>13 (2)</td>
</tr>
<tr>
<td>Not reported/unclear method</td>
<td>33 (5)</td>
</tr>
<tr>
<td>No of patients in MA can be calculated†</td>
<td>93 (14)</td>
</tr>
<tr>
<td>*Relates to review.</td>
<td>93 (14)</td>
</tr>
<tr>
<td>‡Two reviews used more than one method.</td>
<td>93 (14)</td>
</tr>
</tbody>
</table>

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, 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.28 (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.)

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 spikes 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 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, 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.28 (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.30 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. 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).

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 included studies, as well as review methods, as will likely be required in the forthcoming Cochrane Reviews of Test Accuracy.

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 to ensure reviews are useful. We have shown the current poor quality of published reviews and indicated areas for improvement.

References