Number XX

A comprehensive overview of the methods and reporting of meta-analyses of test accuracy

Version [#]

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

This information does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services determination or policy.

Contract No.

Prepared by:
[Redacted]

Investigators
[Redacted]

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted, for which further reproduction is prohibited without the specific permission of copyright holders.

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report.

This report is based on research conducted under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD. The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care researchers and funders of research make well-informed decisions in designing and funding research and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of scientific judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical research and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances.

This report may be used, in whole or in part, as the basis for research design or funding opportunity announcements. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

AHRQ Contacts

Elisabeth U. Kato, MD, MRP
Medical Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

Contents

Abstract 7
Background 8
Methods 10
Search strategy and eligibility criteria 10
Data extraction 10
Information extracted from meta-analyses published between 1966 and 2003 10
Information extracted from meta-analyses published after 2003 11
Data cleaning and quality control 11
Data analysis 12
Results 13
Eligible systematic reviews 13
Literature review methods in reviews of test accuracy 18
Factors affecting the number of studies included in each review 22
Quality assessment and use of checklists 23
Statistical analyses and presentation of results in reviews of test accuracy 26
Comparison of reviews in the 5 most commonly assessed medical fields 30
Comparison of reviews of the 5 most commonly assessed test categories 33
Discussion 35
Abbreviations 38
References 39

Tables

Table 1: Characteristics of eligible reviews. 15
Table 2: Literature search and study selection methods employed in reviews of test accuracy 19
Table 3: Quality assessment in systematic reviews of medical test accuracy studies 23
Table 4: Statistical analyses and presentation of results in reviews of test accuracy 26
Table 5: Comparison of meta-analyses of test accuracy conducted in the 5 medical fields where most reviews had been published 32
Table 6: Comparison of meta-analyses of test accuracy conducted in the 5 test categories assessed in most meta-analyses 34
Table 7: Summary of selected previously published overviews of systematic review of test accuracy 37

Figures

Figure 1: Literature search flow 13
Figure 2: Number of meta-analyses of test accuracy published per year 14
Figure 3: Number of primary studies included in each meta-analysis of test accuracy 16
Figure 4: Trends over time in the proportion of meta-analyses, by clinical field 17
Figure 5: Trends over time in the proportion of meta-analyses, by test category 18
Figure 6: Trends over time in the reporting of search strategies and study selection criteria 20

Figure 7: Trends over time in the proportion of meta-analyses using specific databases or other sources to identify eligible studies 21
Figure 8: Trends over time in the proportion of studies searching for or considering for inclusion studies based on language of publication criteria 22
Figure 9: Trends over time in the proportion of meta-analyses of test accuracy appraising specific quality items or using quality assessment checklists 25
Figure 10: Trends over time in the proportion of meta-analyses using each metric of test accuracy 28
Figure 11: Trends over time in the proportion of meta-analyses assessing, accounting for in the analyses and exploring heterogeneity 29
Figure 12: Trends over time in the proportion of meta-analyses using advanced statistical methods for quantitative evidence synthesis 30

Abstract

Background: Medical tests play a critical role in diagnosis or prediction of future outcomes. Meta-analyses of diagnostic or predictive test accuracy are increasingly performed and the relevant methods are continuously evolving.

Methods: We identified systematic reviews including quantitative synthesis (meta-analysis) of test accuracy for diagnostic or predictive medical tests through MEDLINE searches (1966 through to December 2009) and perusal of reference lists of eligible articles and relevant reviews. We extracted information on topics and test types covered, methods for literature synthesis and quality assessment, availability of data, and statistical analyses performed.

Results: Our searches retrieved 1225 potentially eligible reviews of which 760 (published from 1987 to 2009) were finally considered eligible for inclusion. Eligible reviews included a median of 18 primary studies and typically examined a single index test against a single reference standard. The number of publications increased per calendar year ($P < 0.001$). Most meta-analyses pertained to cardiovascular disease (21 percent) and oncology (25 percent); the most common test categories were imaging (44 percent) and biomarker tests (28 percent). Meta-analyses used multiple electronic databases (62 percent used at least one electronic database in addition to MEDLINE; P for trend over time < 0.001) to identify eligible studies. There was a striking increase in the proportion of systematic reviews that reported assessing verification bias (P for trend < 0.001), spectrum bias (P for trend $= 0.007$), blinding (P for trend < 0.001), prospective study design (P for trend < 0.001), or consecutive patient recruitment (P for trend < 0.001), over time. Improvements were associated with reporting of using quality-item checklists to guide assessment of methodological quality. In statistical analyses, sensitivity (in 77 percent), specificity (in 74 percent) and diagnostic/predictive odds ratios (in 34 percent) were the most commonly used metrics. Heterogeneity tests were used in 58 percent, and subgroup or regression analyses were used in 57 percent of meta-analyses. Random effects models were employed in 57 percent of the reviews and increasingly over time (P for trend < 0.001). Theoretically motivated methods that model sensitivity and specificity simultaneously, while accounting for between-study heterogeneity, were used in a minority of reviews (11 percent) but increasingly over time (P for trend < 0.001).

Conclusion: Meta-analyses of diagnostic or predictive tests are increasingly performed. Over time there have been substantial improvements in the literature review, quality assessment and statistical analysis methods employed. Much of the improvement in quality assessment is associated with the use of quality item checklists. Advanced statistical methods have been increasingly used over time but their use still remains limited.

Background

Diagnostic and predictive tests are an important component of medical care and clinicians rely on test results to establish diagnosis and guide patient management.[1] Despite their central role in patient care, evaluating the effectiveness of specific tests is challenging. Tests affect clinical outcomes indirectly, through the effect of test results on physicians' diagnostic thinking and subsequent management decisions, making it difficult to ascribe patient outcomes to the use of a particular test. The many existing frameworks for assessing the value of testing propose a stepwise appraisal process, moving from analytic validity (technical test performance), to clinical validity (diagnostic and predictive accuracy), clinical utility (effect on clinical outcomes) and overall cost-effectiveness assessment.[2] Primary studies that directly address all components of the assessment framework are very uncommon. Therefore, systematic reviewers are typically faced with the task of putting together the pieces of the puzzle by synthesizing studies that address each component of the framework. While the diagnostic or predictive accuracy of a medical test does not directly inform on the clinical value of testing, it is a crucial piece of the overall puzzle and one that is essential to synthesize in systematic reviews. High test accuracy is a necessary condition for tests to have any impact on clinical outcomes.[3] Meta-analysis of test accuracy can provide an estimate of average test accuracy as well as identify patient-, disease- or test-related modifiers of test performance.[4]

Meta-analyses of test accuracy present particular challenges compared to reviews of randomized trials of therapeutic interventions, not only because the studies reviewed are exclusively observational, but also because of the inherent associations among the metrics of performance. Sensitivity and specificity are likely to be correlated (between studies) because of threshold effects (i.e., because changing the diagnostic threshold affects sensitivity and specificity in opposite directions), necessitating the use of multivariate analytic methods.[5] In the presence of such correlation, univariate meta-analyses of sensitivity and specificity may produce "average" values for each metric that are incompatible and have misleading confidence intervals. Only recently have these methods penetrated into common practice and into methodological guidelines aided by their implementation in readily available software.[6-10] The large number of metrics that can be used to summarize information on test accuracy has added to the analytic complexity. In addition to sensitivity and specificity, metrics such as the odds ratio,[11] area under the receiver operating characteristic (ROC) curve,[12] and likelihood ratios have been proposed for the synthesis of studies of test accuracy.[4] Finally, clinical heterogeneity is omnipresent because the studies differ so much in their settings, patient disease spectra, and versions of the tests used. This diversity often manifests as statistical heterogeneity. Thus, meta-analyses of test accuracy need to quantify and account for the presence of heterogeneity and allow the exploration of factors that may be causing it.

Early on, it was recognized that the quality of medical test accuracy studies was often inadequate.[13,14] Many items typically considered in the appraisal of studies of therapeutic interventions, such as the use of randomization, blinding of patients to the interventions used, or allocation concealment, do not apply to studies of test accuracy. A number of studies have investigated study design and reporting items that may affect estimated test accuracy, but the evidence on which items are most important is inconclusive.[15-17] Drawing on empirical evidence and expert opinion on the quality assessment of accuracy studies, the Quality Assessment in Diagnostic Accuracy Studies (QUADAS) tool was developed and published in November

2003.[18,19] This tool has now been validated for use in systematic reviews of diagnostic tests.[20] Further, a reporting checklist, the Standards for the Reporting of Diagnostic Accuracy Studies (STARD),[21,22] was also published in January 2003. Although the checklist was primarily intended as a guide for the reporting of primary research studies on diagnostic tests, the 25 STARD items pertaining to the design, analysis, and reporting of studies are often used to guide quality assessment in systematic reviews. As these tools have now been available for some time, it is reasonable to assess their impact on quality assessment methods in meta-analyses of medical tests.

Along with the overall number of meta-analytic publications, the number of meta-analyses of test accuracy studies has skyrocketed, increasing from fewer than 10 per year in the early 1990s to almost 100 publications per year in recent years. The question therefore appears to have shifted from whether meta-analysis of medical test accuracy studies is useful,[23] to what methods are best for undertaking such analyses, in terms of study identification and selection, assessment of study quality, statistical analysis and reporting.

We therefore have performed a systematic overview of meta-analyses of medical test accuracy, to assess the current state of the literature and evaluate trends over time in the methods and reporting of such studies.

Methods

This project reviewed meta-analyses published over two time periods: the first covered years up to 2003 and the second covered 2004-09. The search strategies, inclusion and exclusion criteria for both periods were the same. However, for studies published during the second period additional items were extracted from eligible reviews. These differences and the rationale for them are discussed in the pertinent sections below.

Search strategy and eligibility criteria

We searched the MEDLINE database (1966 through to December 2009) using a combination of key words related to test accuracy and meta-analysis. The complete search strategy is presented in **Appendix A**.

Papers were considered eligible when they reported the findings of systematic reviews (defined as reviews using explicit methods to identify, select and extract information from primary research studies) that used quantitative synthesis (meta-analysis) methods to obtain summary estimates of diagnostic or predictive accuracy of medical tests. Our definition of tests encompassed clinical signs and symptoms. We only included English-language reviews published in full text; retrieving the full-text and extracting information from non-English articles entails substantial effort and would be unlikely to affect our conclusions. We did not consider systematic reviews that did not use quantitative analysis methods because one of our key aims was to assess the temporal evolution and current status of meta-analytic methods for synthesizing test accuracy data. We excluded reviews reporting meta-analyses based on individual patient data because they are subject to different design, analysis and reporting considerations. We also excluded Health Technology Assessment documents, evidence reports produced by the Effective Health Care Program of the Agency for Healthcare Research and Quality (AHRQ), and Cochrane Reviews of diagnostic tests; these documents are substantially longer than the typical meta-analyses published in journals and are subject to reporting conventions determined by the respective entities.

Data extraction

Nine reviewers extracted data from non-overlapping sets of publications in extraction forms generated using electronic data collection forms. Forms were piloted using articles extracted independently by multiple reviewers and modifications were performed based on the pilot results. The final data extraction form is presented in **Appendix B**.

Information extracted from meta-analyses published 1966-2003

For each paper we extracted the following items: bibliographic information (first author, journal, year of publication); number of index tests, reference standard tests and the number of studies included in quantitative analyses; medical subspecialty to which tests were pertinent (cardiovascular disease, obstetrics and gynecology, gastrointestinal disease infectious disease, oncology, nephrology/urology, rheumatology, pulmonary medicine, orthopedics, psychiatry, ear-nose-throat, neurology and pediatrics); the types of test being assessed [histology/cytology/culture-based tests, clinical examination, imaging, biomarker, clinical challenge tests (e.g., pharmacological stress tests), physiologic tests (e.g., electrocardiogram, electroencephalogram) or endoscopy]; details about search strategies used; quality assessment and information extracted from each primary study considered by the meta-analysis (including

whether the reviews assessed blinding, spectrum bias, and verification bias in the primary studies they reviewed); use of STAndards for the Reporting of Diagnostic accuracy studies (STARD)[22]; statistical analysis (including assessment and exploration of heterogeneity, metrics used to assess test accuracy and statistical methods used for synthesizing study findings and graphically presenting these results); and assessment of comparative evidence on alternative index tests.

Information extracted from meta-analyses published 2004-2009

All data items extracted from studies published between 1966 and 2003 were also extracted from meta-analyses published between 2003 and 2009; however, from meta-analyses published during this period we extracted additional information on blinding (specifically whether index test or reference standard assessors were blinded); use of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS, first published in November 2003)[18] checklist to guide quality assessment of the primary studies; and whether the reviews collected information on the following variables from each eligible study: spectrum bias, selection criteria, number of withdrawals, number of indeterminate test results, independence of and timing of test results compared to the reference standard, study location participants' sex and age and type of plots for the report.

During the course of data extraction, the review team met regularly to discuss specific papers, review data items, and clarify operational definitions. The majority of investigators participating in the project attended each meeting; resolutions of specific issues were reached by consensus and were circulated to all team members in writing. Eligible meta-analyses published up to 2003 (n = 260) were extracted in duplicate. Due to the rapid increase in the number of eligible publications in more recent years, only a sample of 83 articles (17% percent of eligible studies published between 2004 and 2009) was extracted in duplicate. In all cases, discrepancies were resolved by consensus among extractors.

Data cleaning and quality control

When all reviewers completed their extractions, we merged the individual data extraction forms to generate a combined database. We queried the database to identify missing values, invalid entries (e.g., a numerical value out of the expected range) and logical inconsistencies (e.g., when a study was recorded as not using an advanced statistical method we checked that no such method was checked in the relevant fields). For every missing value identified we required the data extractor familiar with the paper to re-extract information, when necessary with the help of a second reviewer. Additionally, for continuous variables we identified entries with values differing by more than 3 standard errors from their mean value and verified them against the source documents.

To ensure inter-reviewer consistency in the data extractions of reviews published after 2004 we implemented additional quality control measures. After data cleaning, we assessed inter-reviewer consistency by performing statistical comparisons between reviewers for all extracted variables, using chi-squared tests (for categorical variables) or analysis of variance (ANOVA) for continuous variables. To avoid the confounding effect of temporal trends, all comparisons were stratified by year of publication. Variables that reached statistical significance were considered suggestive of the existence of systematic between-reviewer differences and were re-extracted following discussion between reviewers on the pertinent operational definitions.

Data analysis

We calculated descriptive statistics such as means, medians and ranges for continuous variables and proportions for categorical variables, along with appropriate measures to indicate variability around these values (standard deviations, confidence intervals or interquartile ranges). We used histograms to visualize the distributions of variables of interest and line plots to depict trends in the reporting of variables of interest over time.

We compared key methodological and reporting aspects of meta-analyses pertaining to the five most common clinical areas (cardiovascular disease, oncology, gastrointestinal disease, infectious disease, and obstetrics and gynecology) and the five most common test categories (histological/cytological/culture-based tests, aspects of the clinical examination, imaging tests, biomarkers, and physiologic tests) in our dataset. These comparisons were performed using the Fisher exact test for categorical variables or the Kruskal-Wallis test for continuous and count variables.

To detect trends over time in literature review, quality assessment, statistical analysis, and reporting characteristics of meta-analyses, we used logistic regression with each of the items of interest as the response variable and year of publication as an explanatory variable. Change in the number of studies, index and reference standard tests considered in each review were assessed using linear regression of the natural logarithm of these variables on publication year.

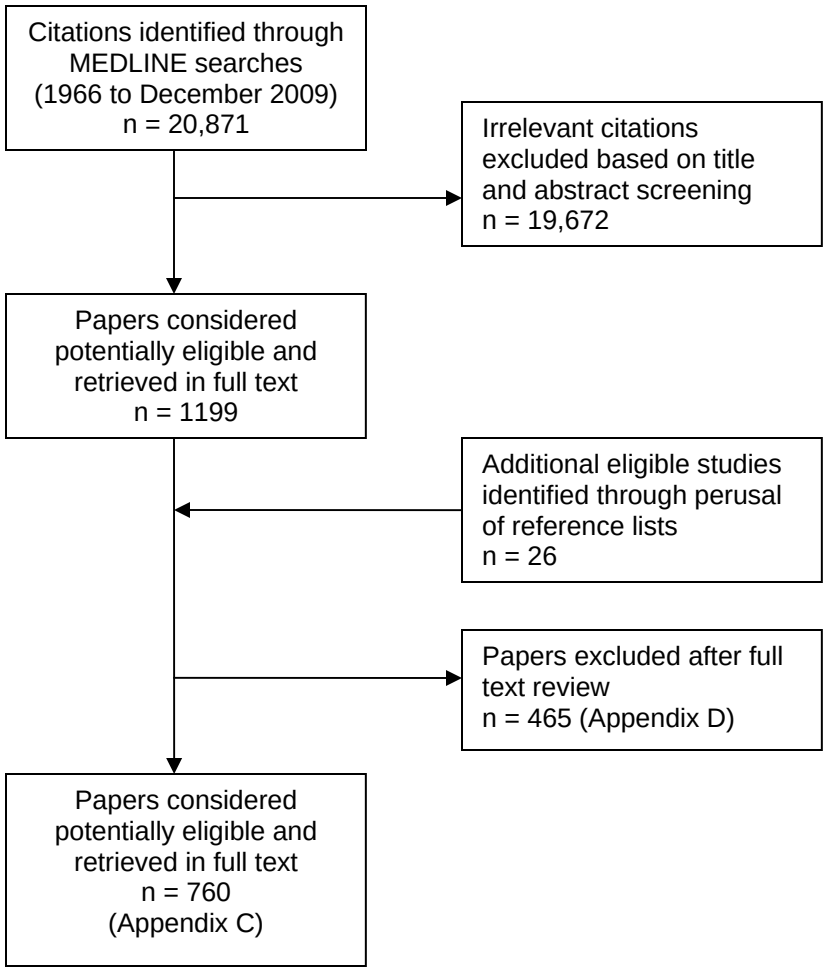
All analyses were conducted using Stata version SE/11.2 (StataCorp, College Station, TX). Statistical significance was defined as a two-sided P value < 0.05 for all comparisons; no adjustments for multiple comparisons were performed.

Results

Eligible systematic reviews

Our searches yielded a total of 20,871 citations; after screening of titles and abstracts 1199 citations were considered potentially eligible and were retrieved in full text. An additional 26 publications were identified through perusal of reference lists of other review articles, for a total of 1225 papers reviewed in full text. Of those, 465 were excluded after full text review and 760 were considered eligible. **Figure 1** presents the search strategy flow. The list of included studies is presented in **Appendix C**; a summary of reasons for exclusion of studies reviewed in full text is presented in **Appendix D**. In the following sections we present a summary of the total database along with an assessment for trends over time for items of interest. **Appendix E** presents the regression results for trends over time for all factors assessed, both over the whole period covered (1987–2009) and for the subgroup of studies published in recent years (2005–2009).

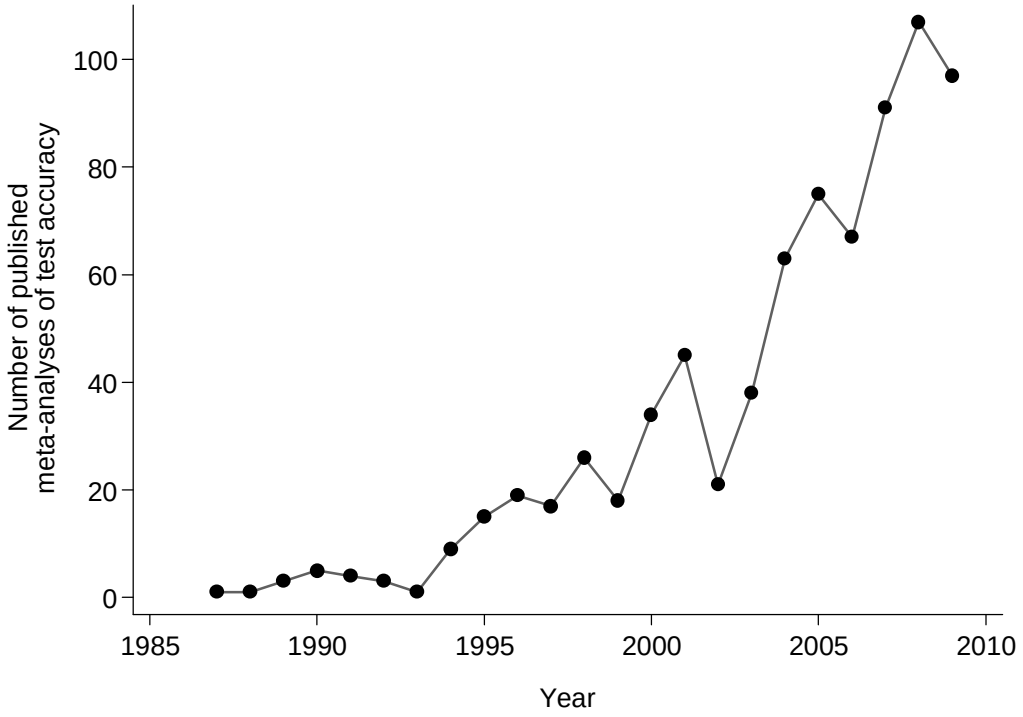
Figure 1: Literature search flow



Flow of the literature search and study selection process for this review. A list of included studies is provided in Appendix C and a list of reasons for exclusion is provided in Appendix D.

Studies were published over more than 20 years (from 1987 to 2009), and there was a clear trend in increased number of reviews over time (P for trend < 0.001; **Figure 2**).

Figure 2: Number of meta-analyses of test accuracy published per year



Line plot of the number of test accuracy meta-analyses included in this overview by year of publication.

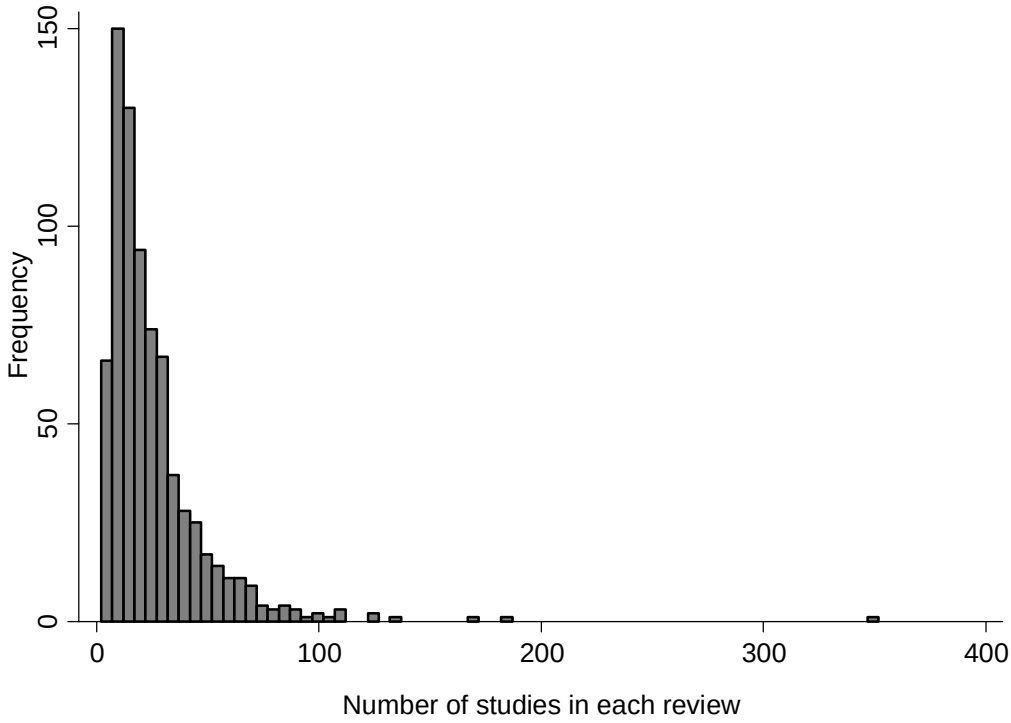
Meta-analyses had synthesized evidence from a median of 18 studies (per review), but the number of studies included varied substantially (25th–75th percentile = 11–30; minimum-maximum = 2–351). Most reviews examined imaging tests (44 percent) or molecular biomarkers (28 percent). Most tests pertained to diagnosis or prediction in oncology (25 percent), cardiology (21 percent), gastrointestinal disease (16 percent), obstetrics and gynecology (15 percent), and infectious disease (13 percent). A majority (52 percent) of meta-analyses considered a single index test and 82 percent considered a single reference standard (**Table 1**). **Figure 3** depicts the distribution of the number of studies included in the eligible systematic reviews. Comparative analyses of two or more index tests were reported in 132 reviews (17 percent).

Table 1: Characteristics of eligible reviews.

Characteristic		Numbers
Topic area	Oncology, n (%)	188 (25)
	Cardiovascular disease, n (%)	160 (21)
	Gastrointestinal, n (%)	119 (16)
	Obstetrics and gynecology, n (%)	114 (15)
	Infectious disease, n (%)	98 (13)
	Pulmonary medicine, n (%)	68 (9)
	Orthopedics, n (%)	44 (6)
	Nephrology and urology, n (%)	38 (5)
	Neurology, n (%)	37 (5)
	Pediatrics, n (%)	29 (4)
	Psychiatry, n (%)	22 (3)
	Ear-nose-throat, n (%)	19 (3)
	Rheumatology, n (%)	8 (1)
Test types	Imaging, n (%)	336 (44)
	Biomarker, n (%)	211 (28)
	Clinical exam, n (%)	112 (15)
	Histology, cytology, or culture, n (%)	103 (14)
	Physiologic or challenge test, n (%)	40 (5)
	Clinical test, n (%)	31 (4)
	Endoscopic examinations, n (%)	21 (3)
Index tests	Per review, median [25th–75th percentile] (min–max)	1 [1–3] (1–56)
	Reviews with a single index test, n (%)	396 (52)
	2 index tests, n (%)	157 (21)
	3 index tests, n (%)	68 (9)
	4 index tests, n (%)	43 (6)
	≥5 index tests, n (%)	96 (13)
Reference standard tests*	Per review, median [25th–75th percentile] (min–max)	1 [1–1] (1–7)
	Reviews with a single reference standard test, n (%)	625 (82)
	2 reference standard tests, n (%)	74 (10)
	3 reference standard tests, n (%)	38 (5)
	4 reference standard tests, n (%)	14 (2)
	≥5 reference standard tests, n (%)	9 (1)
Included studies	Per review, median [25th–75th percentile] (min–max)	18 [11–30] (2–351)
	Reviews with 2–10 studies, n (%)	180 (24)
	11–20 studies, n (%)	246 (32)
	21–30 studies, n (%)	151 (20)
	31–40 studies, n (%)	63 (8)
	41–50 studies, n (%)	45 (6)
	≥51 studies, n (%)	75 (10)
Publications per period	1985–1989, n (%)	5 (<1)
	1990–1994, n (%)	22 (3)
	1995–1999, n (%)	95 (13)
	2000–2004, n (%)	201 (26)
	2005–2009, n (%)	437 (58)

max = maximum; min = minimum. The percentages of medical topics and test types do not sum up to 100% because many test uses could be classified under multiple topics and some reviews assessed more than one test types.

Figure 3: Number of primary studies included in each meta-analysis of test accuracy

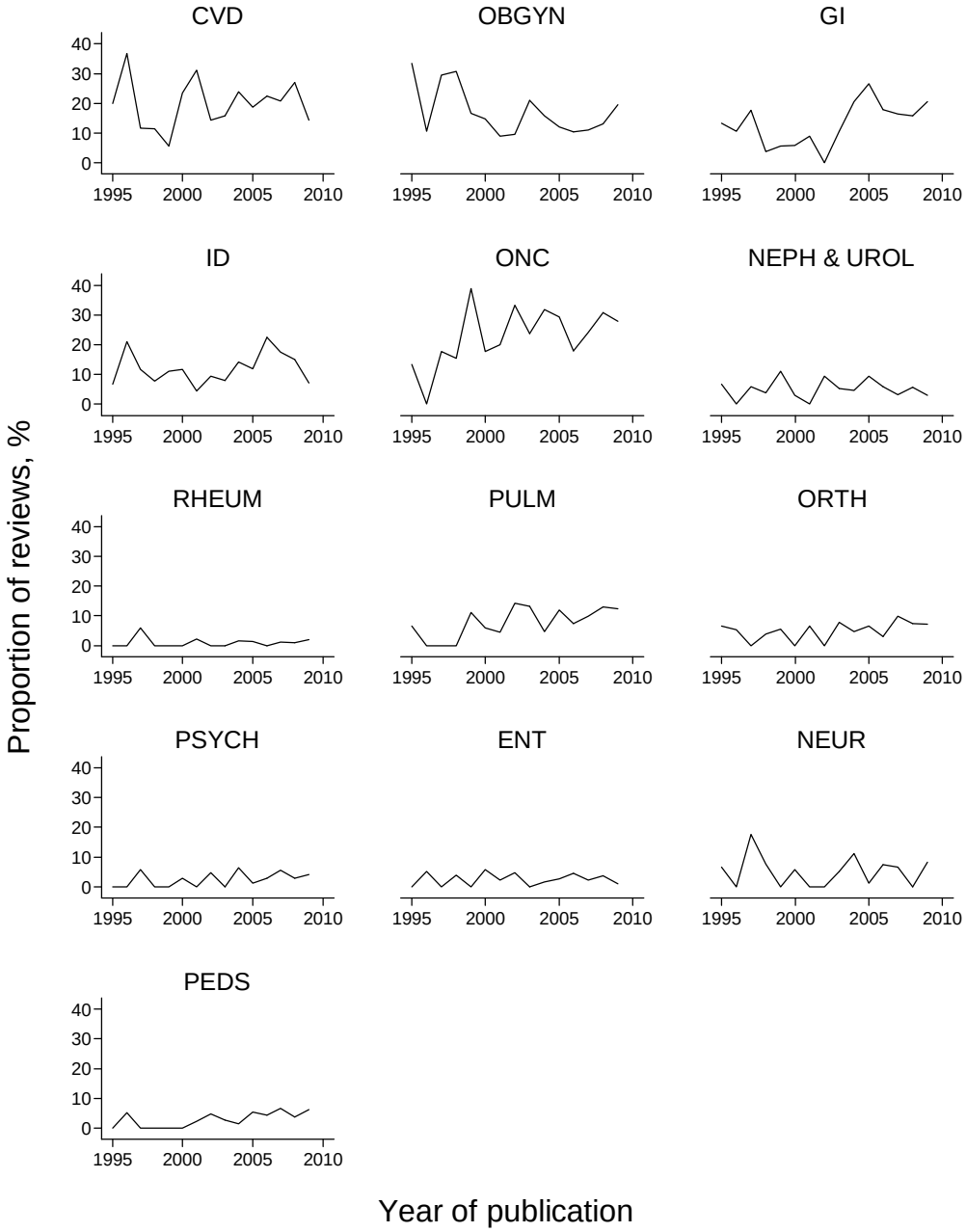


Histogram of the number of primary studies included in each meta-analysis of test accuracy included in this overview. Bin width was set at 5 studies.

Over time, there was no statistically significant change in the number of studies included in each review (-1 percent per year; 95 percent CI -2, 0.3; P = 0.147), or the number of reference standards considered (0 percent per year; 95 percent CI -0.6, 0.7; P = 0.902). However, there has been a small reduction in the number of index tests assessed (-1.2 percent per year; 95 percent CI -2.5, 0; P = 0.040). The proportion of reviews reporting comparative analyses of at least two index tests has not changed over time (per year odds ratio, OR = 0.98; 95 percent CI 0.94, 1.02; P = 0.292).

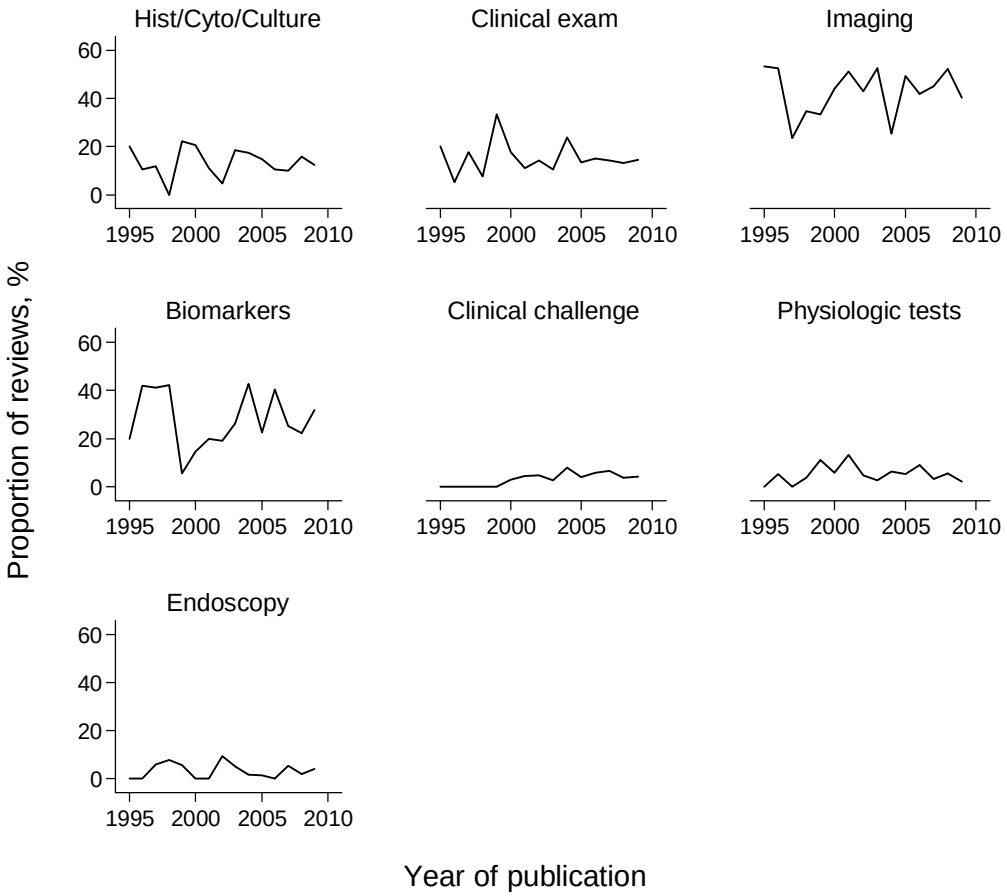
Over time there was an increase in meta-analyses pertaining to gastrointestinal disease (per year OR = 1.07; 95 percent CI 1.02, 1.13; P = 0.008), oncology (per year OR = 1.05; 95 percent CI 1.01, 1.09; P = 0.020), orthopedics (per year OR = 1.09; 95 percent CI 1.00, 1.19; P = 0.041) and pulmonary medicine (per year OR = 1.10; 95 percent CI 1.03, 1.18; P = 0.005). Changes over time were non-significant for other clinical topics (**Figure 4**). There was also a borderline increase in the proportion of meta-analyses assessing clinical challenge tests (per year OR = 1.11; 95 percent CI 1.00, 1.23; P = 0.046) (**Figure 5**).

Figure 4: Trends over time in the proportion of meta-analyses, by clinical field



Line plot of the annual proportion of meta-analyses of test accuracy in various clinical fields. CVD = cardiovascular disease; OBGYN = obstetrics and gynecology; GI = gastrointestinal disease; ID = infectious disease; ONC = oncology; NEPH & UROL = nephrology and urology; RHEUM = rheumatology; PULM = pulmonary medicine; ORTH = orthopedics; PSYCH = psychiatry; ENT = ear-nose-throat; NEUR = neurology; PEDS = pediatrics. Results are shown only after 1995 because the number of meta-analyses in previous years was too small and proportions were unstable.

Figure 5: Trends over time in the proportion of meta-analyses, by test category



Line plot of the annual proportion of meta-analyses of test accuracy in various test categories. Hist/Cyto/Culture = histological, cytological, or culture-based tests; clinical challenge = clinical challenge tests. Results are shown only after 1995 because the number of meta-analyses in previous years was too small and proportions were unstable.

Literature review methods in reviews of test accuracy

Most reviews reported searching multiple electronic databases. MEDLINE searches were nearly universal (96 percent of all reviews) and 62 percent of the reviews reported searching at least one electronic databases in addition to MEDLINE. Searches of Embase (47 percent) and the Cochrane Library (30 percent) were also common. References lists of eligible studies and relevant review articles were also considered in a large proportion of the reviews (76 percent and 28 percent, respectively). On the contrary, contacting experts in the field (17 percent) and obtaining unpublished information (12 percent) were less common.

To guide the selection of eligible studies, 19 percent of eligible reviews reported using quality criteria and 21 percent reported using a minimum cut-off sample size. These cut-offs were generally low (median = 10 participants), but some reviews excluded studies of even moderate sample size (25th–75th percentile = 10–20; 99th percentile =

100). Reviews often considered only studies published in English (36 percent); however, 31 percent of reviews explicitly reported not imposing any language restrictions.

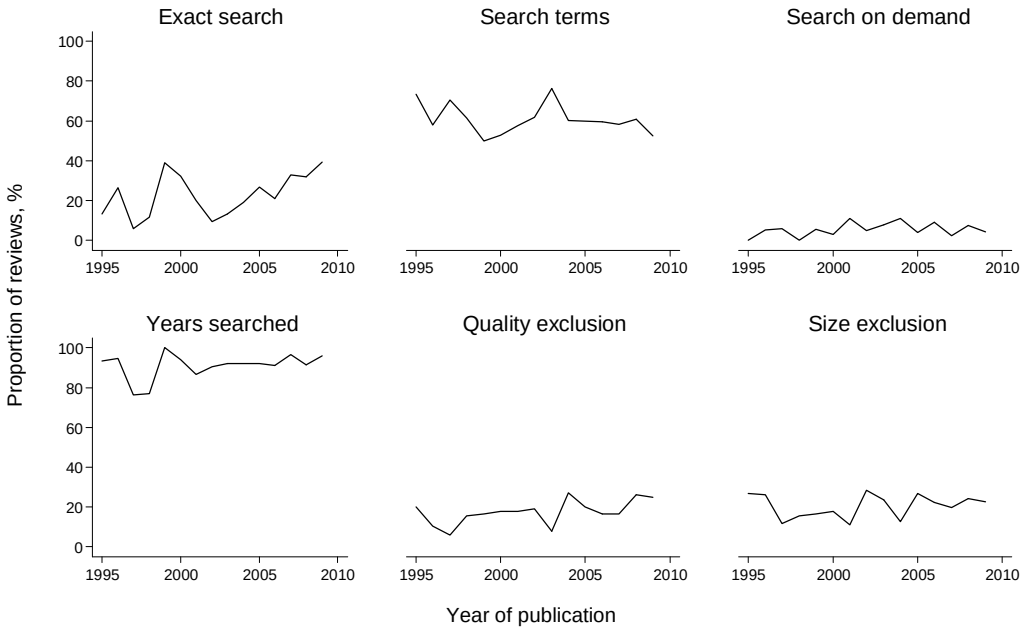
Table 2 summarizes information on the databases searched, the reporting of search strategies in the eligible reviews, and the study selection characteristics of eligible reviews.

Table 2: Literature search and study selection methods employed in reviews of test accuracy

Characteristic		Numbers
Availability of the search strategy	Reporting of the exact search string, n (%)	195 (26)
	Reporting of search terms, n (%)	445 (59)
	Search strategy available upon request, n (%)	43 (6)
	No information reported, n (%)	77 (10)
Reporting of the years searched, n (%)		697 (92)
Exclusion of studies based on quality, n (%)		146 (19)
Exclusion of studies based on sample size, n (%)		159 (21)
Median sample size cut-off [25th–75th percentile] (min–max)		10 [10–20] (4–1000)
Languages searched	English-only (or other single language), n (%)	271 (36)
	English and other specific languages, n (%)	90 (12)
	No language restrictions, n (%)	235 (31)
	No information on selection based on language, n (%)	164 (22)
Electronic bibliographic databases searched	MEDLINE, n (%)	729 (96)
	Embase, n (%)	358 (47)
	Cochrane library, n (%)	228 (30)
	CINAHL, n (%)	82 (11)
	ISI WOK/ SCI, n (%)	74 (10)
	Current Contents, n (%)	35 (5)
	Other specific database, n (%)	202 (27)
Other sources of eligible studies	Conference proceedings, n (%)	88 (12)
	Bibliographies of eligible articles, n (%)	576 (76)
	Bibliographies of relevant review articles, n (%)	213 (28)
	Experts in relevant fields, n (%)	132 (17)
	Manual/electronic searches of specific journals, n (%)	105 (14)
	Manufacturers of tests/assays, n (%)	33 (4)
	Unpublished information, n (%)	93 (12)

Unless stated otherwise, the total sample size for analyses in this table is 744. A single review considering only Korean-language articles is included here along with the "English-only" reviews, to indicate the consideration of articles in a single language only. Percentages may not sum to 100 percent due to rounding.
CINAHL = Cumulative Index to Nursing and Allied Health Literature; ISI WOK/ SCI = Institute of Scientific Information Web of Knowledge/Science Citation Index; max = maximum; min = minimum.

Figure 6: Trends over time in the reporting of search strategies and study selection criteria

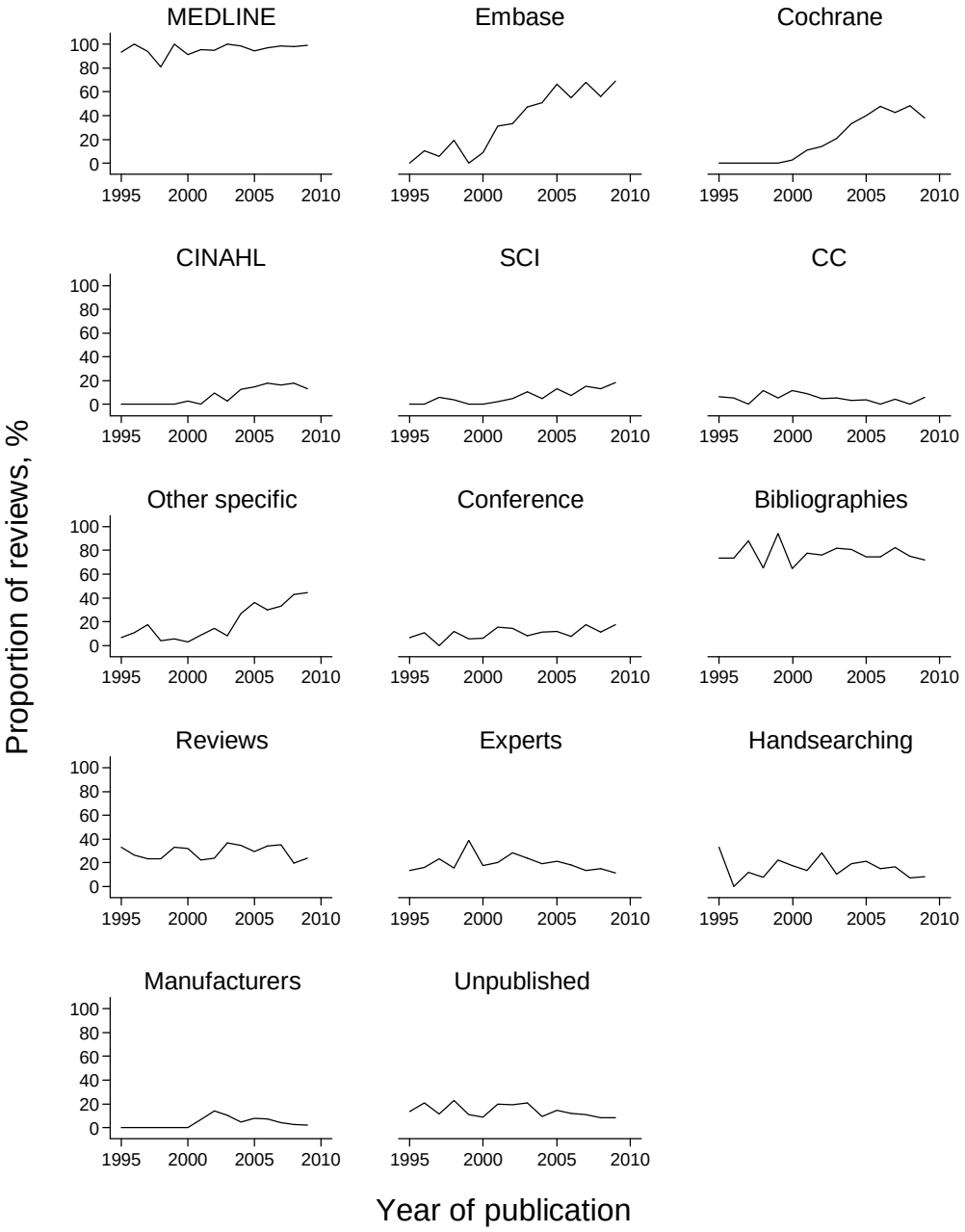


Line plots of the annual proportion of meta-analyses of test accuracy using specific reporting practices for their search strategies or study selection criteria. Exact search = reporting of the search strategy in a way that can be directly replicated; Search terms = reporting of search terms without the full search strategy or Boolean operands; Search on demand = the full search strategy is available from the authors or in a website (other than that of the journal publishing the meta-analysis); years searched = reporting of the years covered by the searches; quality exclusion = exclusion of studies from the meta-analysis based on quality criteria; size exclusion = exclusion of studies from the meta-analysis based on sample size criteria. Results are shown only after 1995 because the number of meta-analyses in previous years was too small and proportions were unstable.

Figure 6 presents trends over time in the reporting of search strategies and methods for study selection for inclusion in meta-analysis. Over time there has been a substantial increase in the number of reviews reporting the exact search strategy used (per year OR = 1.09; 95 percent CI 1.05, 1.14; P < 0.001) and the years searched (per year OR = 1.08; 95 percent CI 1.02, 1.13; P = 0.004). However, there has also been an increase in the number of reviews using quality criteria to select studies for inclusion (per year OR = 1.06; 95 percent CI 1.02, 1.11; P = 0.007).

Figure 7 presents trends over time in the use of electronic databases and other literature sources by systematic reviews of test accuracy. Overall, there was an increasing trend in the use of electronic databases other than MEDLINE (per year OR = 1.30; 95 percent CI 1.24, 1.35; P < 0.001); this appeared to be due to the increasing use of Embase (per year OR = 1.28; 95 percent CI 1.23, 1.35; P < 0.001); the Cochrane libraries (per year OR = 1.30; 95 percent CI 1.23, 1.38; P < 0.001); the Science Citation Index (or other ISI databases, per year OR = 1.18; 95 percent CI 1.09, 1.28; P < 0.001); CINAHL (per year OR = 1.25; 95 percent CI 1.15, 1.36; P < 0.001); or other specific electronic databases (per year OR = 1.25; 95 percent CI 1.18, 1.32; P < 0.001).

Figure 7: Trends over time in the proportion of meta-analyses using specific databases or other sources to identify eligible studies

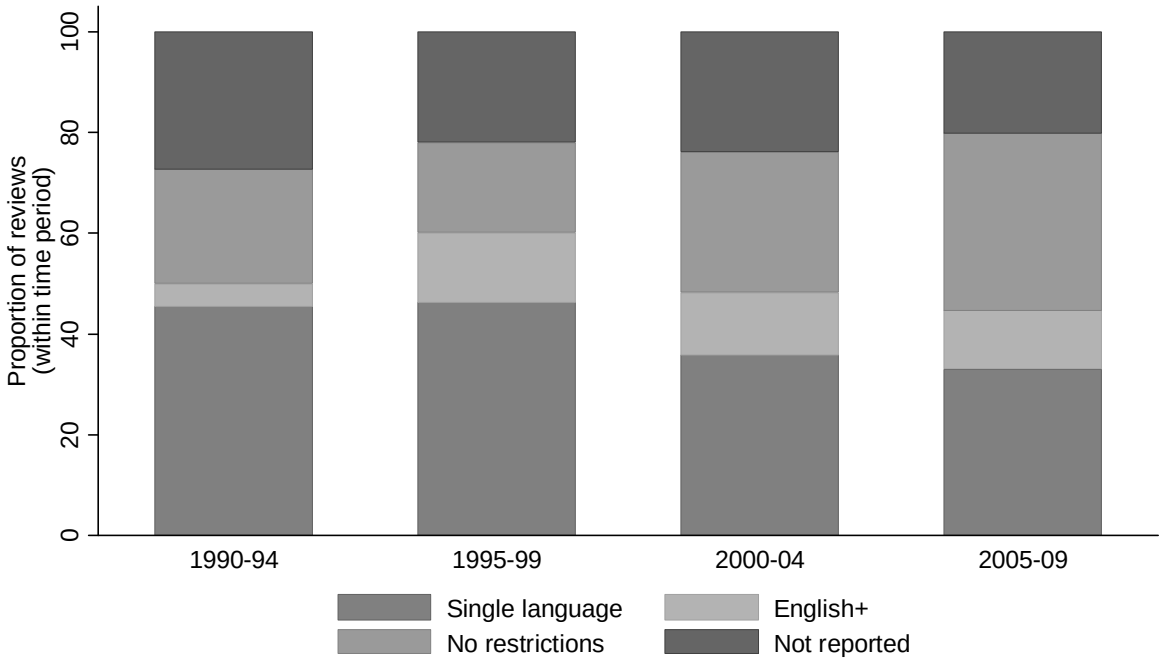


Line plots of the annual proportion of meta-analyses of test accuracy using specific databases or other sources to identify eligible studies. Results are shown only after 1995 because the number of meta-analyses in previous years was too small and proportions were unstable. Bibliographies = perusal of reference lists of included studies; CC = current contents; CINAHL = Cumulative Index to Nursing and Allied Health Literature; Cochrane = searching of databases maintained by the Cochrane Collaboration; Experts in the field consulted to provide additional studies; Handsearching = searching manually (or electronically) the contents of selected journals; Other specific = searching of other specific electronic databases; Reviews = perusal of the reference lists of relevant review articles; SCI = Science Citation Index or other Institute of Scientific

Information databases; Unpublished = search for studies not published in the peer-reviewed literature.

Figure 8 presents trends over time in the handling of languages other than English in reviews of test accuracy. There has been an increase in the proportion of studies that explicitly reported considering non-English language articles (considering studies published in at least one language other than English, per year OR = 1.05; 95 percent CI 1.02, 1.09; P = 0.003). This increase is mostly due to an increasing number of reviews not imposing any language restrictions. As expected, there has been a concomitant decrease in reviews considering English language studies only (per year OR = 0.96; 95 percent CI 0.93, 0.99; P = 0.028).

Figure 8: Trends over time in the proportion of studies searching for or considering for inclusion studies based on language of publication criteria



Stacked bar graph of the annual proportion of meta-analyses of test accuracy using different language criteria to identify eligible studies. Single language = reviews only considering studies published in a single language (in all but one case the language used was English, a single review considered only studies published in Korean); English+ = reviews considering studies published in English and at least one more specific Language; No restrictions = reviews not using language restrictions to select eligible studies; Not reported = studies not providing information on the languages considered. Results are shown only after 1990 because the number of meta-analyses in previous years was too small and proportions were unstable.

Factors affecting the number of studies included in each review

The number of studies included in each meta-analysis did not appear to be affected by the search of electronic databases in addition to MEDLINE (0.2 percent difference between reviews that included at least one additional electronic database; 95 percent CI -12.7, 12.1 percent; P = 0.967) or the inclusion of languages other than

English (8.3 percent more studies in reviews including non-English language studies; 95 percent CI -2.7, 21 percent; P = 0.139), after adjusting for publication year, clinical topic, and test category.

Quality assessment and use of checklists

The majority of reviews performed some qualitative assessment of the studies they included. This assessment was based on the QUADAS checklist and the STARD guidelines in 20 percent (27 percent after 2004) and 9 percent (12 percent after 2003) of reviews, respectively. Commonly assessed items included blinding (65 percent), prospective recruitment of patients (59 percent), verification bias (48 percent), and the description of the reference standard used (88 percent). Blinding of test assessors was examined in more detail for articles published since 2004: blinding of the index test assessor to the reference standard results was reported in 53 percent, and blinding of the reference standard assessor to the index test results in 50 percent of the 500 studies published between 2004 and 2009.

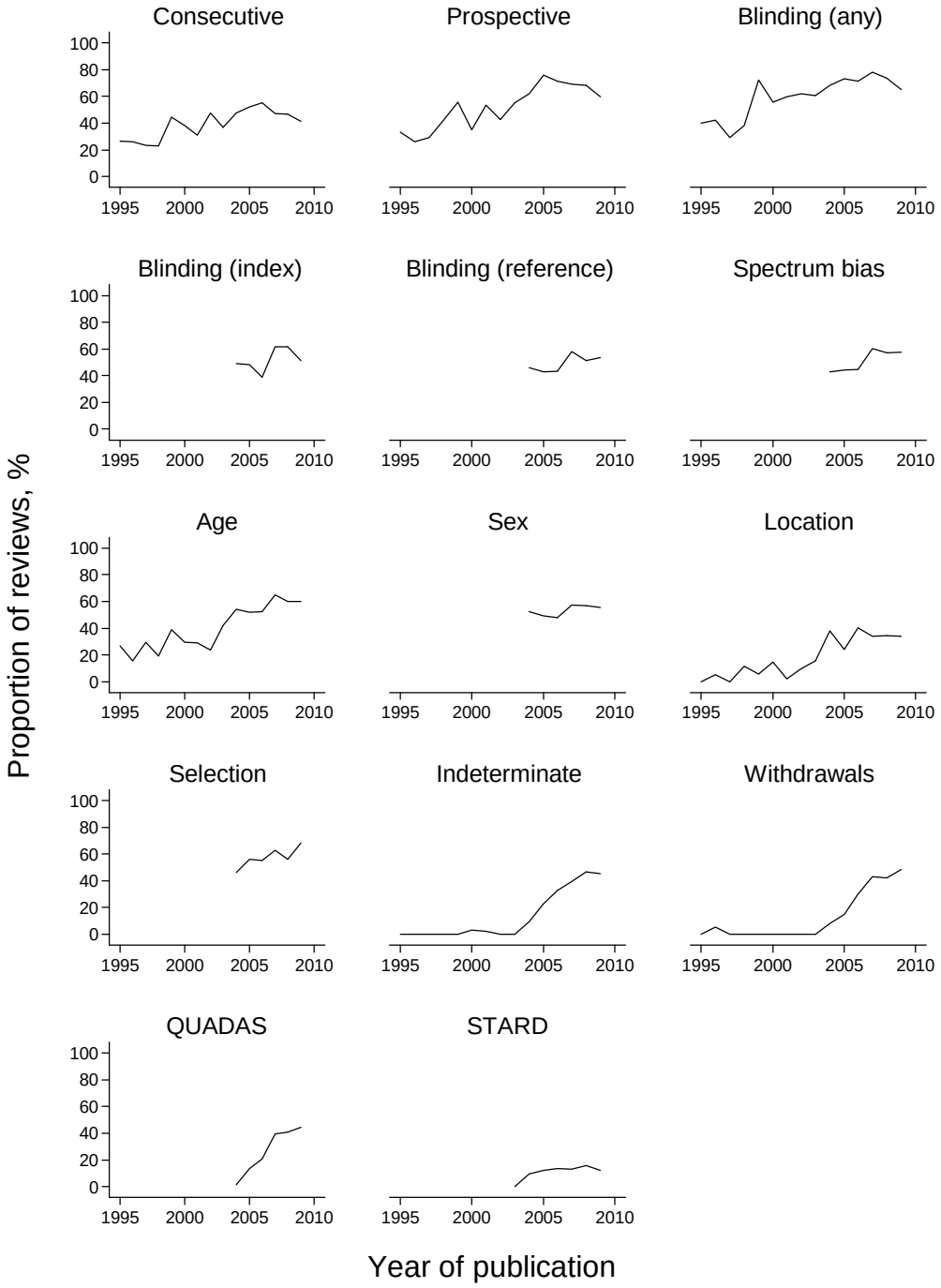
Table 3: Quality assessment in systematic reviews of medical test accuracy studies

Characteristic		Studies, n (%)
Settings of test use		295 (39)
Recruitment of consecutive patients		322 (42)
Prospective/retrospective design		446 (59)
Exact description of the reference standard		667 (88)
Expertise of the test readers		116 (15)
Any blinding (all eligible studies)		496 (65)
Blinding (500 studies published since 2004)	Blinding assessors of the index test to the reference standard	265 (53)
	Blinding assessors of the reference standard test to the index test	250 (50)
	Blinding unspecified	74 (15)
	No blinding	141 (28)
Demographic characteristics of participants	Age	361 (48)
	Sex (500 studies published since 2004)	272 (54)
Location of studies		190 (25)
Spectrum bias (500 studies published since 2004)		266 (53)
Selection bias (500 studies published since 2004)		291 (58)
Time between the performance of the index and reference standard tests		201 (40)
Test independence		190 (38)
Indeterminate results		178 (23)
Withdrawals		168 (22)
Verification bias		367 (48)
Quality assessment based on checklists	QUADAS	148 (19)
	STARD	65 (9)

QUADAS = Quality Assessment of Diagnostic Accuracy Studies; STARD = Standards for Reporting of Diagnostic Accuracy.

Over time, assessment of specific quality items has generally increased. (**Figure 9**). For example, there has been an increase in the number of reviews appraising verification bias (per year OR = 1.17; 95 percent CI 1.13, 1.22; P < 0.001); spectrum bias (per year OR = 1.16; 95 percent CI 1.04, 1.29; P = 0.007); test assessor blinding (per year OR = 1.09; 95 percent CI 1.05, 1.13; P < 0.001); whether study design was prospective (per year OR = 1.12; 95 percent CI 1.09, 1.16; P < 0.001); and whether patients were recruited consecutively (per year OR = 1.07; 95 percent CI 1.04, 1.11; P < 0.001). These improvements in recent years in the quality assessment of primary studies have been associated with the increasing use of quality assessment checklists: since 2004, the QUADAS checklist has increasingly been used to guide quality assessment (per year OR = 1.62; 95 percent CI 1.41, 1.86; P < 0.001). Similarly, since 2003, the STARD checklist has also been increasingly (but not quite statistically significantly) used to guide quality assessment (per year OR = 1.15; 95 percent CI 0.99, 1.33; P = 0.059).

Figure 9: Trends over time in the proportion of meta-analyses of test accuracy appraising specific quality items or using quality assessment checklists



Line plot of the annual proportion of meta-analyses of test accuracy appraising selected quality items among the primary studies they included. Information for the following variables was only collected for studies published after 2003: Blinding (index), Blinding (reference), Spectrum bias, Sex, and Selection bias. The QUADAS tool and the STARD checklist were published in November 2003 and January 2003, respectively. Results for all variables are shown only after 1995 because the number of meta-analyses in previous years was too small and proportions were unstable. Consecutive = reviews assessing whether the primary studies enrolled patients consecutively; Prospective = reviews assessing whether the primary studies had a prospective

design; Blinding = any assessment of blinding in the primary studies; Blinding (index) = assessment of whether the index test assessor was blinded to the reference standard results; Blinding (reference) = assessment of whether the reference standard assessor was blinded to the index test results; Age = reviews that extracted information on participant age from the primary studies; Sex = reviews that extracted information on participant sex from the primary studies; Location = reviews that extracted information on primary study locations; Selection = reviews that extracted information on the selection criteria of the primary studies; Indeterminate = reviews that examined the handling of indeterminate test results in the primary studies; Withdrawals = reviews that examined the handling of withdrawals in the primary studies; QUADAS = Quality Assessment of Diagnostic Accuracy Studies; STARD = Standards for Reporting of Diagnostic Accuracy.

Statistical analyses and presentation of results in reviews of test accuracy

The most popular test accuracy metrics used in meta-analysis were sensitivity (77 percent) and specificity (74 percent). Diagnostic odds ratios (34 percent) and likelihood ratios (31 percent) were also commonly used. Quantitative results were often presented in forest plots (39 percent) or ROC curves (53 percent); other graphical displays were uncommon. Heterogeneity tests were performed in 58 percent of the available studies; potential causes of underlying heterogeneity were explored in 57 percent of analyses (33 percent using exclusively subgroup analyses exclusively and 24 percent using meta-regression with or without subgroup analyses). Random effects models were used in the majority (57 percent) of the studies.

Statistical analyses most often used univariate (one outcome at a time) meta-analyses (87 percent) and the fixed effects summary receiver operating curve characteristics method as described by Moses and Littenberg[12,24] (86 percent of the studies performing ROC analyses). More theoretically motivated methods, such as bivariate random effects[5,25] or hierarchical summary ROC curve models[8,9], were rarely used (11 percent), although this is changing (see below for time trend).

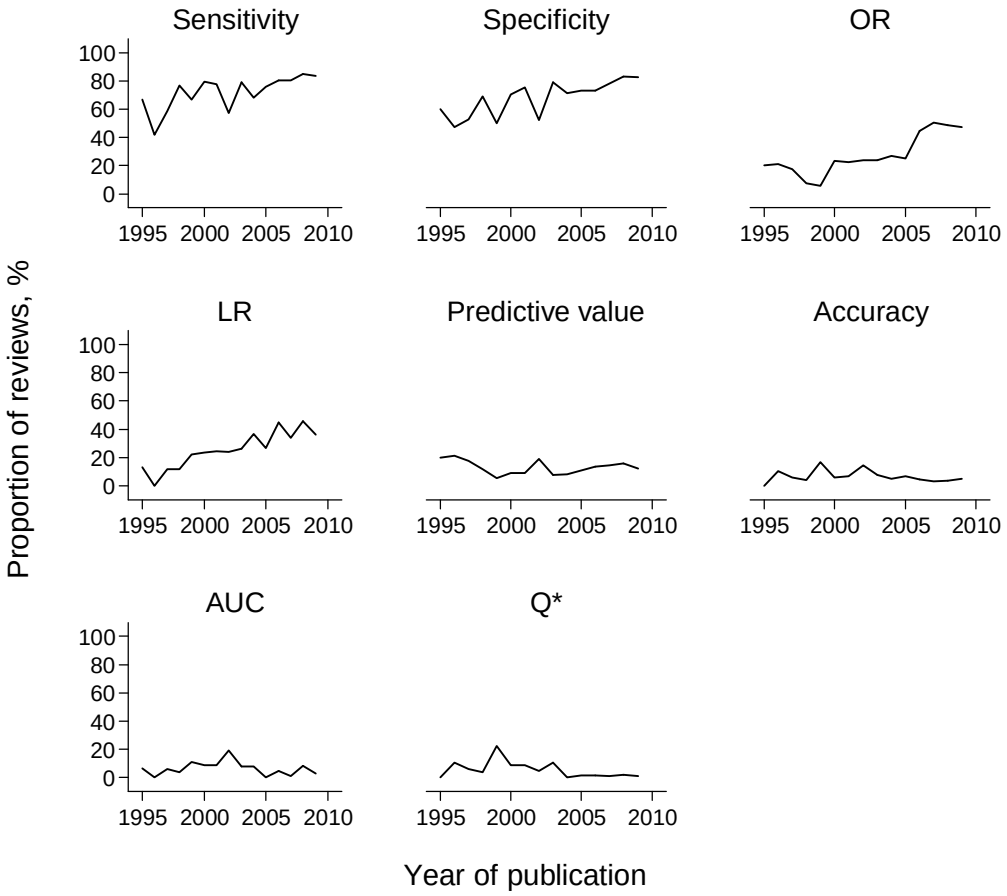
Table 4: Statistical analyses and presentation of results in reviews of test accuracy

Characteristics		Studies, n (%)
Use of random effects models		436 (57)
Metrics used in quantitative analyses	Sensitivity	582 (77)
	Specificity	560 (74)
	OR	257 (34)
	Likelihood ratios	236 (31)
	Predictive values	97 (13)
	Accuracy	42 (6)
	AUC	40 (5)
	Q*	26 (3)
Graphical displays of synthesis results	Forest plots	300 (39)
	ROC curves	403 (53)
Heterogeneity testing		439 (58)
Exploration of heterogeneity	Meta-regression analyses (+/- subgroup analyses)	180 (24)
	Subgroup analyses only	247 (33)
	No exploration	333 (44)
Statistical analyses	Univariate meta-analysis	660 (87)
	ROC-based methods (including sROC and	379 (50)

	hsROC)	
	Advanced statistical methods – BREM/hsROC	70 (9)
	Other advanced models	11 (1)
sROC model (only among studies using ROC analyses)	Moses-Littenberg	326 (86)
	Rutter-Gatsonis	24 (6)
	Other	29 (7)
Comparative analyses of index tests		132 (17)
Type of comparative analyses of index tests (only among the 131 studies reporting on test comparisons)	Direct	33 (25)
	Indirect	98 (75)
Bayesian statistical analyses		17 (2)
Reporting of data to replicate analyses	Counts reported or can be calculated	448 (59)
	Data not available	312 (41)

AUC = area under the curve; BREM = bivariate random effects meta-analysis; OR = diagnostic/predictive odds ratio; HsROC = hierarchical summary receiver operating characteristic; ROC = receiver operating characteristic; sROC = summary receiver operating characteristic.

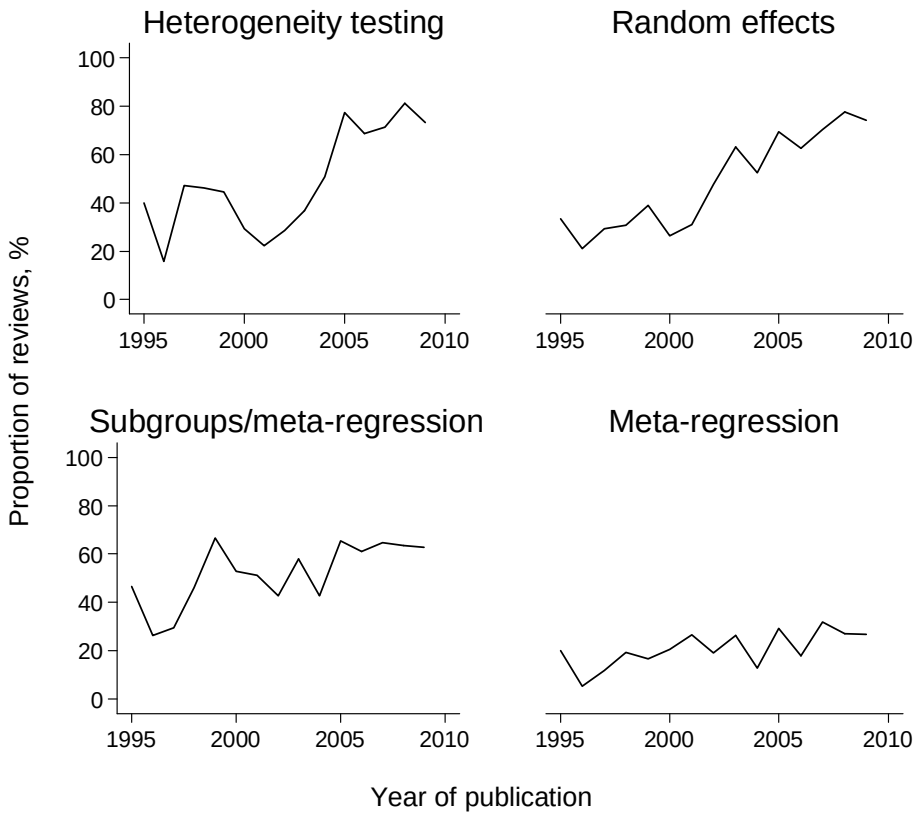
Figure 10: Trends over time in the proportion of meta-analyses using each metric of test accuracy



Line plot of the annual proportion of meta-analyses of test accuracy using each metric for quantitative evidence synthesis. OR = odds ratio; LR = likelihood ratio; AUC = area under the receiver operating characteristic curve; Q* = point where sensitivity equals specificity on SROC curve). Results are shown only after 1995 because the number of meta-analyses in previous years was too small and proportions were unstable.

Figure 10 presents trends over time in the proportion of studies using specific metrics for meta-analysis of test accuracy information. Over time there has been increasing use of the diagnostic OR (per year OR = 1.17; 95 percent CI 1.12, 1.22; P < 0.001), sensitivity (per year OR = 1.07; 95 percent CI 1.03, 1.11; P < 0.001), specificity (per year OR = 1.08; 95 percent CI 1.04, 1.12; P < 0.001), and likelihood ratios (per year OR = 1.13; 95 percent CI 1.08, 1.18; P < 0.001), as metrics for meta-analyses of test accuracy.

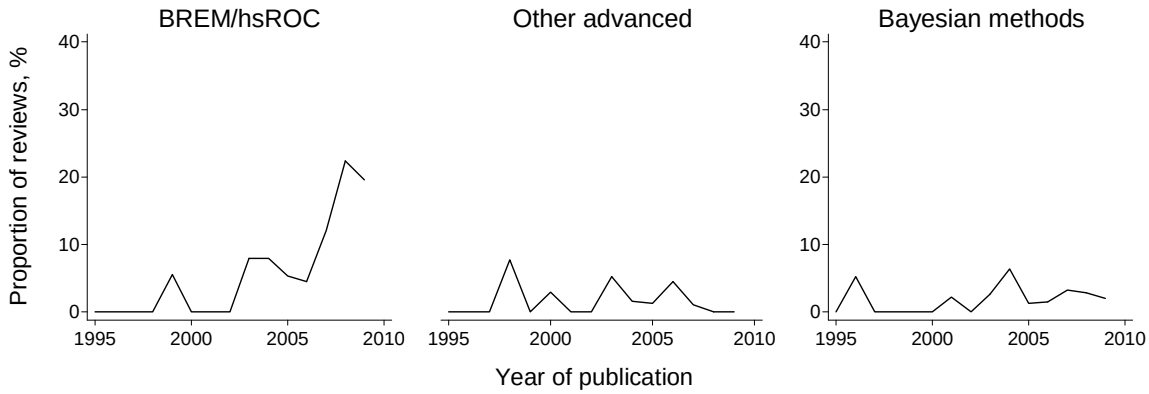
Figure 11: Trends over time in the proportion of meta-analyses assessing, accounting for in the analyses and exploring heterogeneity



Line plot of the annual proportion of meta-analyses of test accuracy using tests for heterogeneity, employing random effects meta-analysis models, exploring heterogeneity (using subgroup or regression analyses), and using meta-regression analyses. Results are shown only after 1995 because the number of meta-analyses in previous years was too small and proportions were unstable.

Figure 11 presents trends over time in the proportion of reviews assessing, accounting for, and exploring heterogeneity. There has been a clear increase in the number of studies assessing heterogeneity using statistical tests (per year OR = 1.21; 95 percent CI 1.17, 1.26; P < 0.001), and exploring the underlying reasons leading to heterogeneity using subgroup or meta-regression methods (per year OR = 1.08; 95 percent CI 1.04, 1.12; P < 0.001). This increase has been mostly due to the use of subgroup analyses, as the proportion of reviews performing meta-regression analyses has not changed significantly over time; per year OR = 1.03 (95 percent CI 0.95, 1.11; P = 0.484). Use of random effects models also increased over time (per year OR = 1.21; 95 percent CI 1.16, 1.26; P < 0.001).

Figure 12: Trends over time in the proportion of meta-analyses using advanced statistical methods for quantitative evidence synthesis



Line plot of the annual proportion of meta-analyses of test accuracy using advanced meta-analysis methods, such as the bivariate random effects meta-analysis or the hierarchical summary receiver operating characteristic curve models; other advanced methods (in most cases, random effects variants of the Lindberg-Moses sROC method); or Bayesian analysis methods. Results are shown only for years after 1995 because the statistical methods of interest were practically not used at all during earlier years and the number of meta-analyses was relatively small leading to instability of the estimated annual proportions. BREM = bivariate random effects meta-analysis; hsROC = hierarchical receiver operating characteristic curve; sROC = summary receiver operating characteristic method.

Figure 12 presents trends in the proportion of studies using advanced statistical methods. Overall, the proportion of studies using advanced meta-analysis methods has increased over time (per year OR = 1.27; 95 percent CI 1.16, 1.40; P < 0.001). Specifically, the bivariate random effects meta-analysis model and the hierarchical sROC model are increasingly used: per year OR = 1.42; 95 percent CI 1.26, 1.60; P < 0.001). These methods were used about 20 percent of the time in 2008 and 2009 following publication of several articles that recommended their use.[5,25-27] Bayesian statistical methods have been rarely utilized.

Comparison of reviews in the 5 most commonly assessed medical fields

The five most commonly assessed medical fields in meta-analyses of test accuracy were oncology (25 percent), cardiovascular disease (21 percent), gastrointestinal disease (16 percent), obstetrics and gynecology (15 percent), and infectious disease (13 percent). For the comparison between fields we excluded 92 studies (16 percent of those relevant to the top 5 fields) that were considered relevant to more than one of these fields. There were few differences in the methods or reporting of meta-analyses in these fields (**Table 5**). There was a higher rate of exclusion of studies using quality criteria in meta-analyses in gastrointestinal (28 percent) and cardiovascular (25 percent) disease compared to those in obstetrics and gynecology (12 percent), infectious disease (13 percent), or oncology (17 percent); overall P = 0.022. Additionally, there was a higher rate of exclusion of studies based on sample size cut-offs in reviews in oncology (31 percent), cardiovascular disease (26 percent), and infectious disease (23 percent), compared to those in obstetrics and gynecology (9 percent) or gastrointestinal

disease (17 percent); overall P = 0.003. Due to the large number of comparisons performed these differences may be chance findings.

Table 5: Comparison of meta-analyses of test accuracy conducted in the 5 medical fields where most reviews had been published

Characteristics		Cardiovascular Disease (n=156)	Obstetrics & Gynecology (n=77)	Gastrointestinal Disease (n=65)	Infectious Disease (n=77)	Oncology (n=120)	P-value
Number of studies, median [25th–75th percentile]		20 [12–40]	16 [10–24]	17 [9–29]	19 [11–32]	18 [11–30]	0.099
Number of index tests, median [25th–75th percentile]		2 [1–3]	1 [1–2]	2 [1–3]	1 [1–2]	1 [1–2]	0.047
Number of reference standards, median [25th–75th percentile]		1 [1–1]	1 [1–2]	1 [1–1]	1 [1–1]	1 [1–1]	0.006
Searching and study selection	Exact search provided, n (%)	46 (29)	17 (22)	17 (26)	19 (25)	25 (21)	0.540
	Years searched, n (%)	147 (94)	73 (95)	60 (92)	70 (91)	112 (93)	0.843
	Exclusion based on quality criteria, n (%)	39 (25)	9 (12)	18 (28)	10 (13)	20 (17)	0.022
	Exclusion based on minimum sample size, n (%)	40 (26)	7 (9)	11 (17)	18 (23)	37 (31)	0.003
	MEDLINE plus at least one additional database, n (%)	66 (42)	42 (55)	36 (55)	45 (58)	55 (46)	0.091
Quality assessment	Settings, n (%)	63 (40)	21 (27)	23 (35)	34 (44)	43 (36)	0.220
	Consecutive patients, n (%)	63 (40)	41 (53)	24 (37)	39 (51)	44 (37)	0.077
	Prospective patient sampling, n (%)	80 (51)	54 (70)	39 (60)	45 (58)	70 (58)	0.103
	Any blinding, n (%)	106 (68)	43 (56)	39 (60)	60 (78)	77 (64)	0.040
	Verification bias, n (%)	70 (45)	38 (49)	29 (45)	37 (48)	60 (50)	0.902
	QUADAS, n (%)	30 (19)	8 (10)	15 (23)	21 (27)	21 (18)	0.086
	STARD, n (%)	13 (8)	7 (9)	6 (9)	7 (9)	11 (9)	0.996
Provides data for re-analysis, n (%)		92 (59)	50 (65)	37 (57)	44 (57)	79 (66)	0.576
Heterogeneity testing, n (%)		84 (54)	51 (66)	39 (60)	42 (55)	61 (51)	0.253
Random effects methods, n (%)		96 (62)	44 (57)	39 (60)	39 (51)	61 (52)	0.385
Advanced methods, n (%)		17 (11)	7 (9)	7 (11)	5 (6)	14 (12)	0.809

QUADAS = Quality Assessment of Diagnostic Accuracy Studies; STARD = Standards for Reporting of Diagnostic Accuracy.

P-values are from Fisher exact tests for nominal variables and Kruskal-Wallis tests for continuous or count variables. Studies investigating tests belonging to different categories have been excluded. Meta-analyses that included index tests belonging to more than one category have been excluded from this analysis.

Comparison of reviews of the 5 most commonly assessed test categories

The five most commonly assessed test categories were imaging tests (44 percent), biomarkers (28 percent), aspects of the clinical examination (15 percent), histological tests (14 percent, including cytological and culture-based tests), and physiologic/challenge (5 percent) tests. For the comparison between test types we excluded 75 studies (11 percent of those relevant to the top 5 test categories) that considered tests belonging to more that one test type (for example, reviews of multiple index tests belonging to different categories). There were several significant differences in the reporting and methods characteristics of meta-analyses assessing different test types (**Table 6**). The most striking of these differences pertained to the methods used by the reviews to appraise the quality of primary studies. Generally reviews of histological, cytological or culture based tests were less likely to assess quality items such as assessor blinding, verification bias, and prospective or consecutive patient recruitment. The use of the QUADAS instrument to guide quality assessment was also less common in reviews of histological, cytological or culture based tests. Similar patterns were observed for physiologic tests (although the number of available reviews was substantially smaller).

Table 6: Comparison of meta-analyses of test accuracy conducted in the 5 test categories assessed in most meta-analyses

Characteristics		Histological tests* (n=68)	Clinical examination (n=81)	Imaging tests (n=296)	Biomarkers (n=172)	Physiologic tests† (n=22)	P-value
Number of studies, median [25th–75th percentile]		19 [11–28]	14 [10–24]	19 [12–33]	21 [11–35]	11 [8–27]	0.010
Number of index tests, median [25th–75th percentile]		1 [1–2]	2 [1–4]	1 [1–2]	1 [1–2]	1 [1–1]	0.001
Number of reference standards, median [25th–75th percentile]		1 [1–1]	1 [1–1]	1 [1–1]	1 [1–1]	1 [1–1]	0.869
Searching and study selection	Exact search provided, n (%)	14 (21)	31 (38)	71 (24)	42 (24)	5 (23)	0.096
	Years searched, n (%)	58 (85)	77 (95)	274 (93)	160 (93)	19 (86)	0.161
	Exclusion based on quality criteria, n (%)	9 (13)	17 (21)	53 (18)	39 (23)	1 (5)	0.171
	Exclusion based on minimum sample size, n (%)	10 (15)	9 (11)	93 (31)	25 (15)	6 (27)	<0.001
	MEDLINE plus at least one additional database, n (%)	26 (38)	45 (56)	145 (49)	96 (56)	7 (32)	0.039
Quality assessment	Settings, n (%)	14 (21)	55 (68)	79 (27)	83 (48)	11 (50)	<0.001
	Consecutive patients, n (%)	24 (35)	36 (44)	121 (41)	85 (49)	4 (18)	0.030
	Prospective patient sampling, n (%)	32 (47)	44 (54)	193 (65)	104 (60)	8 (36)	0.007
	Any blinding, n (%)	29 (43)	58 (72)	202 (68)	117 (68)	14 (64)	0.001
	Verification bias, n (%)	22 (32)	40 (49)	152 (51)	93 (54)	5 (23)	0.003
	QUADAS, n (%)	4 (6)	14 (17)	66 (22)	45 (26)	1 (5)	0.001
	STARD, n (%)	2 (3)	3 (4)	27 (9)	23 (13)	1 (5)	0.034
Provides data for re-analysis, n (%)		47 (69)	42 (52)	177 (60)	101 (59)	13 (59)	0.325
Heterogeneity testing		31 (46)	44 (54)	184 (62)	110 (64)	10 (458)	0.034
Random effects methods, n (%)		33 (49)	49 (60)	168 (57)	109 (63)	9 (41)	0.112
Advanced methods, n (%)		3 (4)	8 (10)	38 (13)	15 (9)	2 (9)	0.280

QUADAS = Quality Assessment of Diagnostic Accuracy Studies; STARD = Standards for Reporting of Diagnostic Accuracy.
P-values are from Fisher exact tests for nominal variables and Kruskal-Wallis tests for continuous or count variables. Studies investigating tests belonging to different categories have been excluded. Meta-analyses covered more than one of the relevant clinical topics have been excluded from this analysis.

* Including cytological and culture-based tests.
† Including challenge tests.

Discussion

We performed a comprehensive review of 760 medical test accuracy meta-analyses published over the last 25 years. This work provides a "snapshot" of the available literature and an overview of longitudinal trends in methods and reporting, with the aim of identifying where future reviews could be improved. Meta-analyses of test accuracy are increasingly being pursued: in recent years approximately 100 such reports are published annually. Overall, the available literature appears to have several limitations: most reviews do not appraise important quality items, statistical analyses use methods that may be suboptimal for test accuracy and direct comparisons of index tests are scarce. Our findings regarding the limitations of existing systematic reviews of test accuracy generally agree with previously published, smaller-scale surveys of reviews of test accuracy. We have summarized some of these previous empirical investigations in **Table 7**. Generally, previous assessments of systematic reviews of diagnostic tests have assessed much smaller numbers of studies or have been limited to a single clinical topic (e.g., oncology[28]). Furthermore, with the exception of a report focusing on the statistical methods used for meta-analysis,[10] no previous overview has included an adequate number of studies spread over several years that would allow the exploration of trends over time.

We also found that many aspects of the methods and reporting of systematic reviews of medical test accuracy have improved over time. Searching of multiple electronic databases without language restrictions has become more common; quality items such as verification bias, spectrum bias and blinding have been increasingly been considered in quality appraisal; and advanced statistical methods that simultaneously model sensitivity and specificity are beginning to be adopted.

Empirical studies comparing the reporting of methodological quality items have documented an increase in the clarity of reporting of quality items in systematic reviews of therapeutic interventions[29] after the International Committee of Medical Journal Editors endorsed the Quality of Reporting of Meta-analyses (QUOROM) checklist, compared to before. Similar data exist on the impact of the Consolidated Standards of Reporting Trials (CONSORT) statement[30,31] for reporting of randomized trials.[32,33] We observed that the QUADAS[18,20] and STARD[21,22] checklists were used more often by recent systematic reviews and we hypothesize that they may have had a similar influence on the reporting of meta-analyses of test accuracy.

A recent focused empirical assessment[10] of meta-analyses of medical tests concluded that the increased use of bivariate random effects statistical models for sensitivity and specificity coincided with the development of easy-to-use routines for performing such analyses (e.g., the `metandi` and `midas` commands in Stata).[34,35] We observed the same pattern. Further, we observed that, at the same time, simpler, but less appropriate methods such as the fixed effects SROC model of Moses and Littenberg[12,24] (which accounts for only part of the uncertainty in the bivariate probability model) begun to be utilized less.

We found substantial differences in methods and reporting of test accuracy studies across different types of medical tests. These differences may reflect either heterogeneous diffusion of methodological advances between research groups focusing on specific test types, or the reviewers' assessment that specific methodological approaches are not applicable to specific test types. In contrast, we found few differences in comparisons across different medical fields.

Our work has several limitations that need to be considered when interpreting our results. First, we relied on searches using methodological filters for identifying reports of meta-analyses medical test accuracy studies.[36] Although such searches may miss some relevant studies, it

appears unlikely that this would substantially affect our results. More comprehensive searches would require the examination of a much larger number of abstracts with little expected incremental yield in terms of eligible studies. Second, we only used MEDLINE supplemented by screening of the reference lists of eligible studies and those of relevant review articles. Given the large number of meta-analyses included in our overview, the overall results would be unlikely to change with the addition of studies indexed by non-MEDLINE databases. Third, we did not perform double extraction for all eligible studies. However, we implemented several procedures for standardizing the definition of the extracted variables during data extraction and performed extensive quality control of the final dataset. Further, a substantial proportion of eligible articles were extracted in duplicate.

This comprehensive overview of meta-analyses of test accuracy highlights the current status and the temporal evolution of a complex research field. Available meta-analyses of medical tests have several limitations in regards to methodological approaches and reporting characteristics; however, over time reviews have increasingly performed more comprehensive assessments of study quality and have used more appropriate statistical methods addressing the particular challenges relevant to reviews of test accuracy. The large and rapidly expanding number of available meta-analyses identified by this overview reflects the growing interest in "evidence-based diagnosis".[37,38] Increasing use of quality checklists is expected to facilitate further improvements in the quality assessment of primary studies included in meta-analyses of medical tests. Similarly, increasing diffusion of methodological advances, availability of software to perform advanced statistical analyses and clear guidelines for the conduct and reporting of meta-analyses of test accuracy will hopefully lead to further improvements in the practice of systematic reviews of medical tests.

Table 7: Summary of selected previously published overviews of systematic review of test accuracy

Characteristics	Irwig, 1994[39]	Whiting, 2005[40]	Dinnes, 2005[41]	Mallet, 2006[28]	Moher, 2007[42]	Willis, 2011[10]	Current project
Number of included SRs	11	114	189 (133 used statistical synthesis methods)	89 (25 assessed in detail)	23 (diagnostic/prognostic SRs among 300 SRs identified)	236	760
Selection criteria	All inclusive; meta-analysis of test accuracy as primary focus.	All inclusive	All inclusive	Cancer diagnosis; included SRs regardless of the use of quantitative synthesis methods. Screening tests + tests for risk factors were excluded; computer decision tools were also excluded.	All inclusive	All inclusive; reviews had to have searched ≥2 databases, stated search terms and inclusion criteria, and used a statistical method to summarize test accuracy.	All inclusive; reviews had to have used a statistical method to summarize test accuracy. Excluded HTAs, Cochrane reviews and AHRQ EPC reports.
Years covered	Jan 1990 – Dec 1991	1995 – 2001	Up to 2002	1990 – 2003	November 2004	Up to 2008	1966 – 2009
Databases searched	MEDLINE, experts, bibliographies of retrieved papers	DARE	DARE (update of the search used in Whiting et al.[40])	MEDLINE, Embase, MEDION, Cancerlit, HTA, DARE, Cochrane Database of Systematic Reviews	MEDLINE	MEDLINE, Embase, CINAHL, Cochrane Library, PsychInfo, Global health, HMIC, AMED	MEDLINE, bibliographies of papers and relevant reviews
Items extracted	Literature review methods; data extraction and presentation; statistical analysis methods.	Quality assessment methods.	Systematic review methods; statistical analysis and reporting; trends over time in statistical method use.	Objectives and setting of the SRs; participant characteristics. In the 25 studies assessed in detail: quality assessment methods; whether meta-analysis was performed; reporting of results; availability of data for reanalysis.	Basic reporting characteristics of reviews (bibliometric features, outcomes considered, whether any quantitative method was used).	Statistical and quality assessment methods; settings of test use; trends over time in statistical method use.	Literature search and study selection methods; quality assessment methods; statistical analyses and reporting methods; trends over time in multiple aspects of the review process, statistical analysis and reporting; availability of data for re-analysis.
Main findings	2 of 11 studies reported the complete search strategy; all studies analyzed sensitivity and specificity, 2 studies used the sROC method and 2 did not provide a summary estimate; 6 of 11 studies discussed variability in reference standards; 7 studies reported comparisons between 2 or more index tests.	49% of SRs had not conducted quality assessment; in most cases information on quality was incorporated in narrative synthesis; 13% of reviews used quality as an inclusion criterion.	70% of SRs used quantitative methods; 52% used MEDLINE as the only source; 69% performed quality assessment; median number of studies=18; 68% of SRs do not report tests for heterogeneity (58% of those using statistical analyses); naïve pooling has decreased over time.	75% of SRs stated inclusion criteria, 40% reported details of study design, 17% reported on the clinical setting, 17% reported on disease severity, 49% reported on tumor stage. Of the 25 reviews assessed in detail, 56% reported sensitivity, specificity, and sample sizes for individual studies. Of the 89 reviews, 61% attempted to formally synthesize results of the studies and 32% reported formal assessments of study quality.	No SRs were updates of previous reviews; harms were considered in 54% and costs in 35% of the reviews were this information was considered relevant; median number of included studies=39; quantitative synthesis was performed in 48% of SRs.	27% of SRs used advanced statistical methods (BREM or hsROC); between 2006 and 2008 QUADAS was used in 40% of the studies; imaging tests are the most commonly assessed test category; 80% of tests are normally used in specialist settings.	As detailed in this report.

Studies are listed chronologically, based on the dates covered by their searches. AHRQ = Agency for Healthcare Research and Quality; AMED = Allied and Complementary Medicine database; BREM = bivariate random effects meta-analysis; CINAHL = Cumulative Index to Nursing and Allied Health Literature; DARE = Database of Reviews of Effects; EPC = Evidence-based Practice Center; HMIC = Health Management Information Consortium database; hsROC = hierarchical summary receiver operating characteristic meta-analysis method; HTAs = health technology assessments; QUADAS = Quality Assessment of Diagnostic Accuracy Studies; SR = systematic review; sROC = summary receiver operating characteristic method.

Abbreviations

ANOVA	analysis of variance
AHRQ	Agency for Healthcare Research and Quality
CI	confidence interval
EPC	Evidence-based Practice Center
OR	odds ratio
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
QUOROM	Quality of Reporting of Meta-analyses
ROC	Receiver operating characteristic
STARD	STAndards for the Reporting of Diagnostic Accuracy Studies

References

(1) Miettinen OS. The modern scientific physician: 3. Scientific diagnosis. CMAJ. 2001;165:781-82.

(2) Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making. 2009;29:E13-E21.

(3) Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144:850-855.

(4) Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol. 2006;187:271-81.

(5) Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58:982-90.

(6) van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Stat Med. 1993;12:2273-84.

(7) van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002;21:589-624.

(8) Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol. 1995;2 Suppl 1:S48-S56.

(9) Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865-84.

(10) Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. BMC Med Res Methodol. 2011;11:27.

(11) Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56:1129-35.

(12) Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993;12:1293-316.

(13) 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.

(14) Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926-30.

(15) 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;282:1061-66.

(16) Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174:469-76.

(17) Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? BMC Med Res Methodol. 2005;5:20.

(18) 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.

(19) 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:iii, 1-iii234.

(20) Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.

(21) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003;138:40-44.

(22) 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.

(23) Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol. 1995;48:119-30.

(24) Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making. 1993;13:313-21.

(25) Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol. 2006;59:1331-32.

(26) Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8:239-51.

(27) Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making. 2008;28:621-38.

(28) Mallett S, Deeks JJ, Halligan S, Hopewell S, Cornelius V, Altman DG. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. BMJ. 2006;333:413.

(29) Wen J, Ren Y, Wang L, Li Y, Liu Y, Zhou M et al. The reporting quality of meta-analyses improves: a random sampling study. J Clin Epidemiol. 2008;61:770-775.

(30) Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med. 2001;134:657-62.

(31) Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med. 2001;134:663-94.

(32) Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA. 2001;285:1992-95.

(33) Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust. 2006;185:263-67.

(34) Harbord RM. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. In: Sterne J, ed. Meta-Analysis: An Updated Collection from the Stata Journal. Stata Press; 2011.

(35) Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J Clin Epidemiol. 2004;57:925-32.

(36) Wilczynski NL, Haynes RB. Consistency and accuracy of indexing systematic review articles and meta-analyses in medline. Health Info Libr J. 2009;26:203-10.

(37) Fowler PB. Evidence-based diagnosis. J Eval Clin Pract. 1997;3:153-59.

(38) Whiting P, Harbord R, de S, I, Egger M, Sterne J. Evidence-based diagnosis. J Health Serv Res Policy. 2008;13 Suppl 3:57-63.

(39) 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.

(40) 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.

(41) 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;9:1-113, iii.

(42) Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med. 2007;4:e78.