Clinical Chemistry 58:11
1534–1545 (2012)

Review

Systematic Reviews of Studies Quantifying the Accuracy of Diagnostic Tests and Markers

Johannes B. Reitsma,[1*] Karel G.M. Moons,[1] Patrick M.M. Bossuyt,[2] and Kristian Linnet[3]

Systematic reviews of diagnostic accuracy studies allow calculation of pooled estimates of accuracy with increased precision and examination of differences in accuracy between tests or subgroups of studies. Recently, several advances have been made in the methods used in performing systematic reviews of diagnostic test accuracy studies, most notably in how to assess the methodological quality of primary diagnostic test accuracy studies by use of QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) instrument and how to develop sound statistical models for metaanalysis of the paired measures of test accuracy (bivariate metaregression model of sensitivity and specificity). This article provides an overview of the different steps within a diagnostic systematic review and highlights these advances, illustrated with empirical data. The potential benefits of some recent developments in the areas of network metaanalysis and individual patient data metaanalysis for diagnostic tests are also discussed.
© 2012 American Association for Clinical Chemistry

There is growing awareness that proper evaluation of diagnostic tests, including biochemical tests, is a requirement for making informed decisions regarding the approval of these tests and recommendations for their use *(1)*. In response, more and more primary evaluation studies of diagnostic tests are being performed and reported in the literature. In view of the increasing number of such studies, healthcare professionals more frequently turn to systematic reviews when seeking the best evidence about diagnostic tests.

In this series of 4 reports, various approaches to the evaluation of diagnostic tests and markers are covered, ranging from single-test accuracy studies [report 1 *(2)*], to multiple-test studies [report 2 *(3)*], to the evaluation of tests by their impact on patient outcomes and cost-effectiveness [report 4 *(4)*]. Each category of study has unique aspects that must be considered in the performance of systematic reviews of these kinds of studies. The focus of this third report in the series is on the methodology used for performing systematic reviews of diagnostic accuracy studies. This report will help readers of systematic reviews of diagnostic studies to judge key aspects of such reviews that may affect the validity (risk of bias) or applicability (generalizability) of the results of a review.

The reasons for performing a review of accuracy studies of diagnostic tests, (bio)markers, or even multivariable diagnostic models incorporating several diagnostic tests or markers, are similar to those for performing a review of (randomized) studies on therapeutic interventions:

- Provide a transparent overview of all relevant studies highlighting differences in design and conduct.
- Perform statistical pooling of estimates of the diagnostic accuracy of a test from individual studies to increase precision.
- Use the increased statistical power gained by pooled estimates to generate or confirm hypotheses about differences in accuracy of a test between clinical subgroups, to examine the impact of study features on the accuracy of a test, and to compare accuracy between different tests.

However, compared with systematic reviews of therapeutic interventions, reviews of diagnostic test accuracy carry additional complexities, particularly as related to the joint interest in 2 measures of accuracy per study (sensitivity and specificity) and the existence of specific forms of biases in diagnostic research *(5, 6)*. Several advances in review methodology have been made in recent years to tackle these complexities *(7, 8, 9)*. An important milestone confirming the relevance and maturity in methods of diagnostic reviews is the uptake of reviews about the accuracy of diagnostic tests in the Cochrane Collaboration database *(9, 10)*.

Here we describe the key steps in performing a systematic review and metaanalysis of diagnostic test accuracy studies, thereby highlighting recent advances in methodology. We illustrate this with a recently pub-

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands; [2] Department Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, the Netherlands; [3] Section of Forensic Chemistry, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark.
* Address correspondence to this author at: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, the Netherlands. Fax +31-887555485; e-mail j.b.reitsma-2@umcutrecht.nl.
Received January 17, 2012; accepted June 11, 2012.
Previously published online at DOI: 10.1373/clinchem.2012.182568

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

Table 1. Main steps in a systematic review of diagnostic test accuracy studies.

Steps	Key issues
I. Framing the review question	Target condition
	Intended role of the index test(s)
	Population of interest
II. Searching and locating studies	Multiple databases
	Multiple search terms related to target condition and index test
	No search filters
III. Assessment of methodological quality	To assess risk of bias and sources of variation
	QUADAS-2 checklist
IV. Metaanalyzing diagnostic accuracy data	Descriptive figures: forest plots, and ROC plot
	Use of hierarchical random effects models:
	+ hierarchical summary ROC approach
	+ bivariate metaregression of sensitivity and specificity
	Study-level covariates to examine differences in accuracy between index tests or subgroups
V. Interpreting results and drawing conclusions	Precision and variability across studies of relevant accuracy measures
	Absolute numbers of true-positive, false-positive, true-negative, false-negative findings derived from summary accuracy measures

lished systematic review on diagnostic decision rules alone or in combination with a D-dimer assay for the diagnosis of pulmonary embolism; see Appendix 1 *(11)*. In the final section some recent developments are discussed.

Key Steps in a Diagnostic Accuracy Review

Table 1 lists the different steps within a diagnostic review and highlights the key issues within each step. These steps are similar to those of a review of randomized therapeutic studies, but it is probably fair to say that each step in a diagnostic review is more complicated.

I. FRAMING THE REVIEW QUESTION

How the question to be answered is framed is a critical step in any review, but the framing process is more complex in the context of diagnostic test accuracy (DTA)[4] reviews. The reason is that tests can have different roles and be applied at different points in the diagnostic work-up of patients *(12)*. As described in the second report of this series, the diagnostic process in clinical practice typically consists of multiple tests being applied in a specific order. Such ordered lists of steps are also known as diagnostic pathways or work-ups. On the basis of test results obtained at each step, reassurance can be provided to patients owing to their low test-based likelihood of a serious condition, treatment can be offered if the likelihood of disease is high enough, or further testing may be required when there is remaining uncertainty. Therefore, the composition of the patient population changes along the diagnostic pathway. Such population changes that occur on the basis of earlier test results are likely to change the accuracy of a subsequent test, as explained in the second report of this series. One example would be the ability of positron-emission tomography–computed tomography (PET-CT) scanning to detect distant metastasis in patients with esophageal cancer who have been scheduled for a major operation with curative intent. This major operation is not worthwhile when distant metastases are present. PET-CT has been evaluated in different types of studies reflecting different clinical scenarios in which this test can be applied. For example, one group of studies has examined whether PET-CT can detect additional metastasis in patients in whom ultrasound and MRI did not find any metastasis. This is an example of an add-on question: will the addition of a new test correct previous errors by finding additional cases? Any remaining metastases are likely to be small or otherwise difficult to detect, and PET-CT scanning for the detection of these metastases later in the process is more difficult than when PET-CT is evaluated earlier in the diagnostic pathway as a possible alternative to MRI.

Because the accuracy of a test is likely to differ depending on its place in the diagnostic pathway, specifying the intended role and placement of the index test in the diagnostic pathway or clinical context is the first and critical step for any DTA review. A clear statement in the review of the potential role of the index test (e.g., the intended change in the diagnostic pathway) will facilitate the interpretation of the results of the review *(12)*. Highlighting the intended change in the diagnostic pathway is helpful in determining the appropriate patient population for which the right comparator test (if applicable) is being selected and in choosing and interpreting accuracy measures. A modification of the PICO (population, intervention, comparison, out-

[4] Nonstandard abbreviations: DTA, diagnostic test accuracy; PET, positron emission tomography; IPD, individual patient data.

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

come) system used in therapeutic studies can also be helpful in framing the question in diagnostic reviews with the following elements:

P: Population, the patients in whom the test will be applied in practice. Important elements: setting, presenting symptoms, prior testing;
I: Index test(s), the test under evaluation;
C: Comparator test(s), relevant if there is an interest in comparing the accuracy of different index tests;
O: Outcome, the target condition and how the final diagnosis will be made (i.e., reference standard).

In our example of clinical decision rules for pulmonary embolism (see Appendix 1) the key point of interest is whether these rules (alone or in combination with a D-dimer test) can be used to select patients who do not require further invasive or costly testing. This type of intended role of a test has been referred to as triage. Therefore the number of patients with negative test results but with a final diagnosis of pulmonary embolism is of key interest (i.e., missed cases).

II. SEARCHING FOR AND LOCATING STUDIES

Identifying all relevant studies is a key objective in any systematic review. Searching for diagnostic accuracy studies proves to be more difficult than for intervention studies because there are no specific search terms for diagnostic test accuracy studies, such as "randomized clinical trial" for therapeutic intervention studies *(13)*. Search strategies in diagnostic reviews are generally based on combining different sets of terms related to: (*a*) the test(s) under evaluation and (*b*) the clinical condition of interest. Both MESH terms and free text words describing the index test and condition of interest should be used in the search. The articles of these 2 sets can then be combined using the Boolean "AND" operator. The use of filters to limit a set of articles to diagnostic accuracy studies is not recommended because the use of these filters can cause a meaningful number of relevant studies to be missed (i.e., in situations in which the filter is not sensitive enough) or, in case of highly sensitive filters, can lead to a situation in which hardly any articles are eliminated from those that need to be screened *(14)*.

Limiting the search to a single database, for example MEDLINE, is generally not considered adequate for systematic reviews *(15)*. Relying solely on MEDLINE may result in the retrieval of a set of reports unrepresentative of all reports that would have been identified through a comprehensive search of several sources. EMBASE is a logical additional source to be searched because it also covers all areas of healthcare *(16–18)*. Many more specific databases exist that can be useful additional sources depending on the topic of the review. The IFCC database may be a useful extra source because it contains diagnostic reviews of tests or markers used in the domain of clinical chemistry (website: www.ifcc.org).

In our pulmonary embolism example, both MEDLINE and EMBASE have been searched using multiple alternative search terms for "pulmonary embolism" combined with multiple search words that can indicate studies reporting results from diagnostic studies. The full search strategy is available as an appendix on the website of the *Annals of Internal Medicine (11)*.

III. QUALITY ASSESSMENT

Quality assessment in diagnostic accuracy reviews focuses on 2 different, but related, concepts. Assessing both these concepts is important because they may explain why findings differ between studies. The first dimension to consider is whether the results of a study may be biased. Similarly to therapeutic intervention studies, diagnostic accuracy studies have key features in the design and execution that can produce incorrect results within a study. This is often described as "internal validity." Examples of threats to internal validity are the use of an inappropriate reference standard, studies in which all patients are verified by the reference standard to determine the presence or absence of the target condition (partial verification), or knowledge of the outcome of the reference standard when the results of the index test are interpreted.

Even if there are no apparent flaws directly leading to bias, a diagnostic accuracy study may generate results that are not applicable for the particular question that the review tries to answer. The patients in the study may not be similar to those in whom the test is used, the test may be used at a different point in the care pathway, or the test may be used in a different way than in practice. This refers to the issue of external validity or generalizability of results.

In 2003, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) checklist was published as a generic tool for the quality assessment of primary studies within a diagnostic review *(19)*. The QUADAS tool consists of 14 items covering risk of bias, sources of variation, and reporting quality. Each item is rated "yes," "no," or "unclear," where yes indicates the absence of problems. Recently, the QUADAS tool has been revised on the basis of comments from users and experts in the field *(8)*. The items of the revised QUADAS-2 checklist are shown in Table 2. Reviewers are encouraged to add additional items that are relevant for the particular review that they are carrying out.

Quality assessment of included studies takes time and effort because reporting is often incomplete or unclear. Furthermore, several quality items require a subjective judgment of the assessor, for example when

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

Table 2. Revised QUADAS-2 checklist

Table 1. Risk of bias and applicability judgments in QUADAS-2[a]

Domain	Patient selection	Index test	Reference standard	Flow and timing
Description	Describe methods of patient selection	Describe the index test and how it was conducted and interpreted	Describe the reference standard and how it was conducted and interpreted	Describe any patients who did not receive the index tests or reference standard or who were excluded from the 2 × 2 table (refer to flow diagram)
	Describe included patients (previous testing, presentation, intended use of index test, and setting)			Describe the interval and any interventions between index tests and the reference standard
Signaling questions (yes, no, or unclear)	Was a consecutive or random sample of patients enrolled?	Were the index test results interpreted without knowledge of the results of the reference standard?	Is the reference standard likely to correctly classify the target condition?	Was there an appropriate interval between index tests and reference standard?
	Was a case control design avoided?	If a threshold was used, was it prespecified?	Were the reference standard results interpreted without knowledge of the results of the index test?	Did all patients receive a reference standard?
	Did the study avoid inappropriate exclusions?			Did all patients receive the same reference standard?
				Were all patients included in the analysis?
Risk of bias (high, low, or unclear)	Could the selection of patients have introduced bias?	Could the conduct or interpretation of the index test have introduced bias?	Could the reference standard, its conduct, or its interpretation have introduced bias?	Could the patient flow have introduced bias?
Concerns about applicability (high, low, or unclear)	Are there concerns that the included patients do not match the review question?	Are there concerns that the index test, its conduct, or its interpretation differ from the review question?	Are there concerns that the target condition as defined by the reference standard does not match the review question?	

[a] Reproduced with permission from Whiting et al. *(8)*.

judging whether the spectrum of patients in a study matches that of the intended population defined in the review question. Given these difficulties, the strong advice is that at least 2 persons should independently perform the quality assessment. These persons should have relevant knowledge of both the methodological issues in diagnostic accuracy studies and the clinical topic area.

Results of quality assessment can be presented in tables and graphs. Tables can be used to document all features of the included studies including the QUADAS-2 items. Such a table takes up a lot of space and does not provide a useful succinct summary for the reader. These tables are often reported as supplemental material on the websites of journals.

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

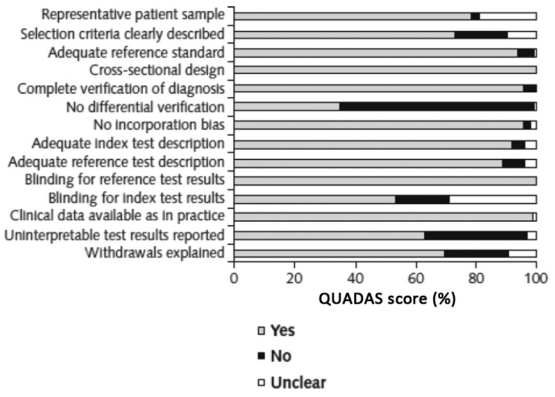

Fig. 1. Quality assessment of included studies performed with the original QUADAS checklist in a recently published systematic review on diagnostic decision rules used alone or in combination with a D-dimer assay for the diagnosis of pulmonary embolism.

Reproduced with permission from Lucassen et al. *(11)*.

Two graphical summaries are recommended for presenting the results of the quality assessment. The methodological quality graph presents, for each quality assessment item, the percentage of included studies in which the item was rated "yes," "no," and "unclear" in a stacked bar chart. This type of graph provides the reader with a quick overview of the study quality within the whole review. The methodological quality graph of our pulmonary embolism example is given in Fig. 1, which shows that potential areas of concern are differential verification (e.g., the use of different reference standard for different groups of patients in a study) and the large proportion of studies providing no data that uninterpretable test results were present.

A systematic review provides an opportunity to investigate how features of study design, execution, and reporting may have an impact on study findings. One way is to give a narrative summary of the quality assessment and discuss how susceptible the results are to particular biases. Another approach is to do a sensitivity analysis in which studies that fail to meet some standard of quality are excluded. Metaregression allows direct examination of the impact of specific individual quality items on diagnostic accuracy (see next section).

IV. METAANALYSIS AND PRESENTATION OF POOLED DIAGNOSTIC TEST ACCURACY RESULTS

Metaanalysis is the use of statistical techniques to combine the results from a set of individual studies. We can use metaanalysis to obtain summaries of the results of relevant included studies, such as an estimate of the mean diagnostic accuracy of a test or marker, the statistical uncertainty around this mean expressed with 95% CIs, and the variability of individual study findings around mean estimates. Metaanalytical regression models can statistically compare the accuracy of 2 or more different tests and examine how test accuracy varies with specific study characteristics.

In the metaanalysis of diagnostic accuracy studies the focus is on 2 statistical measures of diagnostic accuracy: the sensitivity of the test (the proportion of patients with the target disease who have an abnormal test result) and the specificity of the test (the proportion of patients without the target disease who have a normal test result). Statistical methods for diagnostic test accuracy have to deal with 2 outcomes simultaneously (i.e., sensitivity and specificity) rather than a single outcome measure (e.g., a relative risk or odds ratio) as is the case for reviews of therapeutic interventions *(5)*. The diagnostic metaanalytical models have to allow for the trade-off between sensitivity and specificity that can arise because studies may vary in the threshold value used to define test positives and test negatives [also see report 1 in our series *(20)*]. Another feature of diagnostic reviews is the many potential sources for variation in test accuracy results between studies. Examining factors that can (partly) explain variation in these results and the use of a random effects model are key features of a DTA review.

DESCRIPTIVE STATISTICS

The first step in the analysis is to visualize the results from the individual studies within a review. There are 2 types of figures that can be used: forest plots of sensitivity and specificity and plots of these measures in ROC space.

Forest plots display the estimates of sensitivity and specificity of each study, the corresponding CIs, and the underlying raw numbers in a paired way (Fig. 2). These plots give a visual impression of the variation in results between studies, an indication of the precision by which sensitivity and specificity have been measured in each study, the presence of outliers, and a sense for the mean values of sensitivity and specificity.

Plotting the pairs of sensitivity and specificity estimates from separate studies in the ROC space provides additional insight regarding the variation of results between studies, in particular whether sensitivity and specificity are negatively correlated (Fig. 3). The x axis of the ROC plot displays the (1 − specificity) obtained in the studies in the review and the y axis shows the corresponding sensitivity. The rising diagonal line indicates values of sensitivity and specificity belonging to a test that is not informative, i.e., the chances for a positive test result are identical for patients with and without the target disease. Better (e.g., more informative) tests will have higher values of both sensitivity and

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

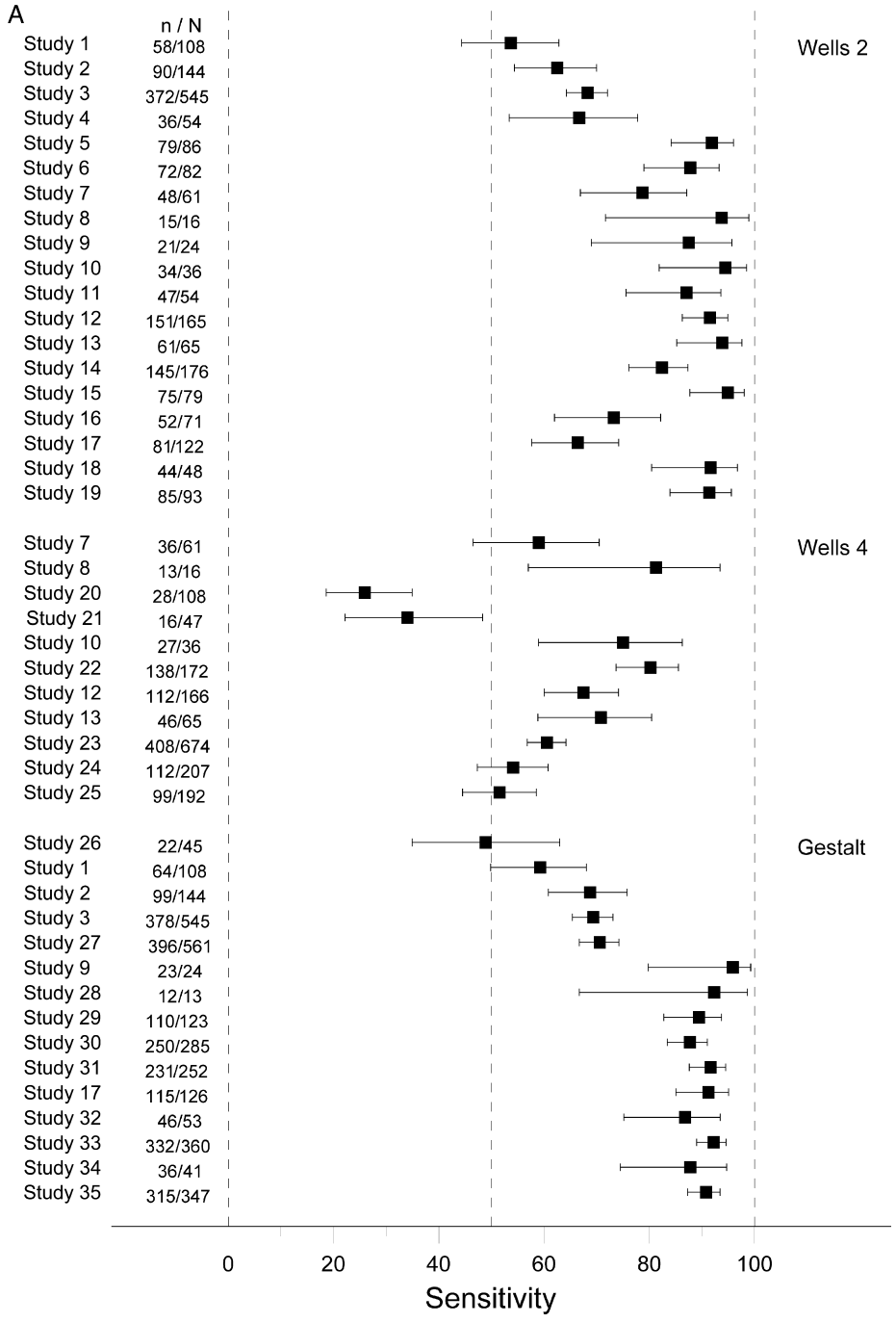

Fig. 2. Paired forest plot of sensitivity (A) and specificity (B) and the corresponding 95% CIs from studies examining the diagnostic accuracy of the Wells rule with a cutoff value of 2, Wells rule with cutoff value of 4, and Gestalt for the diagnosis of pulmonary embolism.

Studies within a rule are sorted by prevalence. Adapted from Lucassen et al. *(11)*.

Continued on page 1540

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

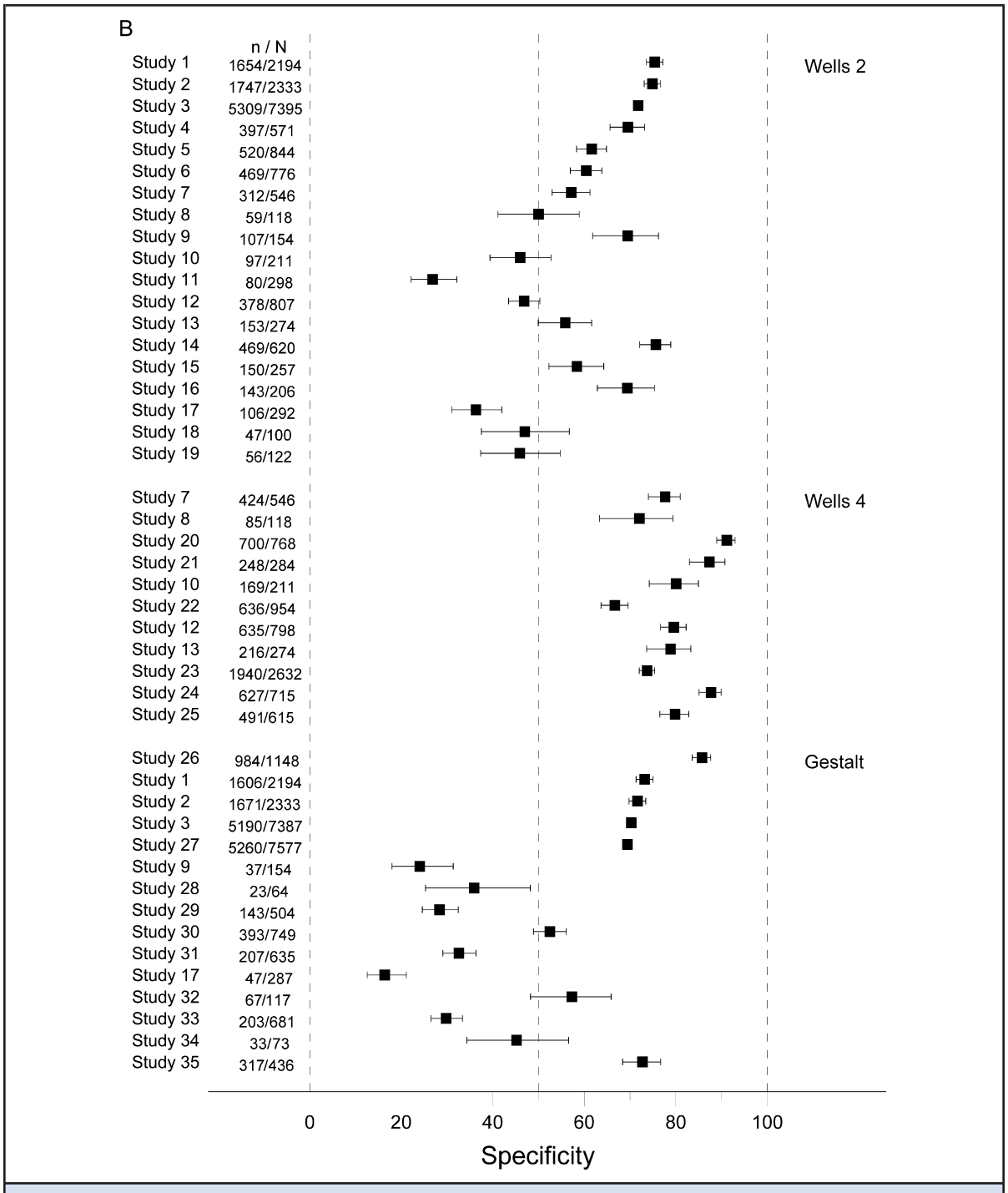

Fig. 2. Continued.

specificity and are therefore located more toward the top-left corner of the ROC-space. If there is a tradeoff (e.g., negative correlation) between sensitivity and specificity, a shoulderlike pattern in the ROC space will emerge. This pattern will be comparable to the pattern that arises in a single study of a test that produces a continuous result in which the threshold has been varied. Lowering the threshold will then increase the likelihood of a positive test result in patients with the target disease, thereby increasing sensitivity while at the same

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

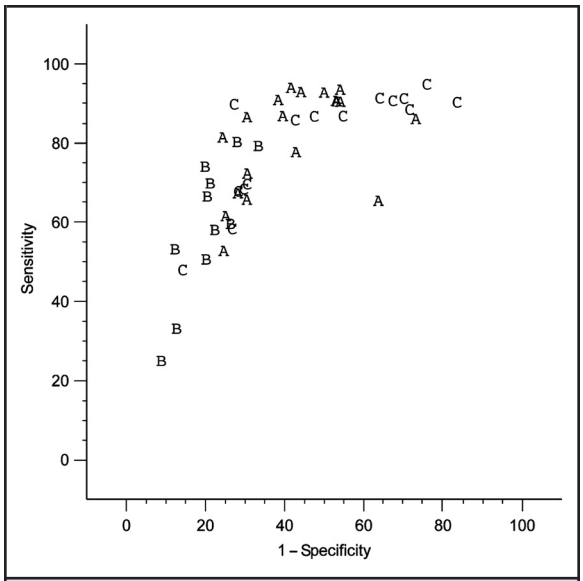

Fig. 3. Pairs of sensitivity and specificity values from studies examining 3 different rules for the diagnosis of pulmonary embolism (A, Wells rule with a cutoff value 2; B, Wells rule with a cutoff value of 4; C, Gestalt). Adapted from Lucassen et al. *(11)*.

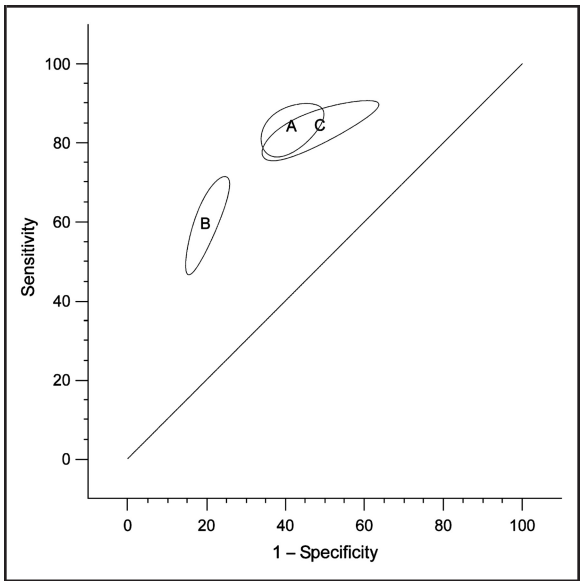

Fig. 4. ROC plot showing the summary estimate of sensitivity and specificity and the corresponding 95% confidence ellipse for 3 different clinical decision rules for the diagnosis of pulmonary embolism (A = Wells rule with cutoff value 2; B = Wells rule with cutoff value of 4; C = Gestalt). Adapted from Lucassen et al. *(11)*.

it increases the risk of a false-positive result in patients without the target disease, thereby lowering specificity. This trade-off or negative correlation will generate this shoulder-like pattern in the ROC space.

The ROC plot of our 3 clinical decision rules for pulmonary embolism clearly indicates the presence of negative correlations both within rules as well as across different rules (Fig. 3).

Metaanalysis of Diagnostic Accuracy Data

Metaanalyses of studies reporting sensitivities and specificities have often used the Moses–Littenberg linear regression approach *(21)* to obtain a summary ROC curve. It has become clear that this approach has statistical shortcomings *(5, 22)*, and therefore it is no longer recommended for evaluating differences between summary ROC curves between tests or examining the impact of covariates on accuracy.

To overcome the shortcomings of the Moses–Littenberg approach, 2 more rigorous statistical approaches have since been developed. These are the hierarchical summary ROC approach and the bivariate random effects model *(5, 22)*. Both models are hierarchical random effects models that take into account the between-study variation in sensitivities and specificities (e.g., random effects models) and their possible correlations as well as the precision of these estimates within a study (e.g., weighting of studies). Although the starting points of these 2 models are different, the 2 models are mathematically equivalent *(23)*. Both models can produce summary estimates of sensitivity and specificity and produce a statistically sound summary ROC line or provide 95% confidence ellipses around the mean values of sensitivity and specificity (Fig. 4).

EXAMINATION OF SOURCES OF VARIATION AND DIFFERENCES IN ACCURACY BETWEEN TESTS

Results from individual studies often vary within a review. There are several possible causes for variation, which can be categorized according to the groups shown in Table 3.

Both advanced models are regression models that allow flexibility in examining sources of heterogeneity by including study-level covariates. This feature provides the option of formally comparing the results of studies with a specific feature (e.g., partial verification) with the results of studies that have avoided partial verification. In the same way we can examine whether the accuracy results from studies examining test A are different from studies examining test B. Limiting the comparison of tests to studies with a cross-over design in which both index tests have been applied in the same patient may be a preferred approach. These so-called

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

Table 3. Causes for variation in sensitivity and specificity results between primary studies within a review.

Chance variation	The majority of diagnostic accuracy studies are moderate to small in sample size. Considerable variation by chance can then be expected, especially for sensitivity when the prevalence is low. The advanced models properly take into account the precision by which sensitivity and specificity have been measured in each study.
Differences in threshold	Explicit or implicit differences in thresholds for positivity between studies will lead to differences in sensitivity and specificity in opposite directions, creating negative correlations. The advanced models take the possible correlations into account.
Bias	Deficiencies in the design and conduct of diagnostic studies can lead to biased results, often producing more exaggerated results. Advanced models can examine the impact of deficiencies in design by including study-level covariates
Variation by clinical subgroups	Examine stratified results or summaries at a study level.
Unexplained variation	It is likely that remaining variation beyond chance will be present in DTA reviews. The advanced models use random effects to incorporate variation beyond chance.

paired comparisons provide more valid evidence than results generated from unpaired studies (separate studies) that may reflect other underlying differences in design and conduct between studies (i.e., confounding factors).

Diagnostic accuracy can vary between clinical subgroups. Examining such differences in a systematic review is problematic if primary studies do not report stratified results for these subgroups. In the absence of stratified results, researchers have to use study-level summaries of the covariate representing the clinical subgroup. Such summaries have limited power for detecting differences in accuracy between clinical subgroups. As an example, if reviewers are interested in whether accuracy of a test varies between men and women they could use the percentage of males in each study as a study-level covariate in their model. A study-level covariate reduces the power to find differences between males and females, which would clearly be the case if all included studies had similar percentages of males. Even if a clear difference in accuracy existed between males and females in each study, it would remain undetected in a regression model based on the percentage of males. Individual patient data (IPD) metaanalysis provides more power and flexibility to examine variation in accuracy between clinical subgroups.

Just as for any regression model used for examining covariates, clear boundaries exist that define what can be done or what is sensible given the sample size of the study. Insufficient statistical power or an increased risk of finding false-positive associations when many covariates are examined are concerns when diagnostic reviews are conducted. The number of different studies within a review is the key limitation for examining covariates.

The results from the bivariate model comparing the three different rules are summarized in Fig. 4 and Table 4.

These results show that (as expected) mean sensitivity is significantly lower for the Wells studies using a cutoff value of 2 compared with studies using a cutoff value of 4, but at the same time specificity is significantly lower. Such differences are expected when lowering the threshold for positivity. The results from the Gestalt studies are comparable with the Wells studies using a cutoff value of 2, although there appears to be more heterogeneity in the reported specificities of Gestalt studies (Fig. 2).

In the example review on pulmonary embolism the authors examined whether the prevalence in a study had an impact on the levels of sensitivity and specificity in a study by including it as a covariate in the bivariate metaregression model. There are several reasons why prevalence might be associated with sensitivity and specificity. An overview of these potential reasons is given in *(24)*.

In this case, the authors hypothesized that differences in prevalence could be seen as a proxy for differences in case mix between studies. In studies with lower prevalence, more patients may be in an early stage of the disease, which would hamper detection and lead to more false-negative results, and hence lower sensitivity. In this review, increased prevalence was associated with higher sensitivity and lower specificity (Table 4).

V. INTERPRETING RESULTS AND DRAWING CONCLUSIONS

This is the part of the review process in which all the results of the different steps within a systematic review have to be combined to answer the review question(s) at hand. Key ingredients include the methodological quality of the evidence, whether the included studies examined the same intended role of the test as ex-

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

Table 4. Mean (95% CI) values of sensitivity and specificity for 3 different clinical decision rules for pulmonary embolism, the impact of prevalence on sensitivity and specificity, and failure rate and efficiency of a strategy in which patients with a low probability of disease and a negative D-dimer receive no further testing.

Subgroup (no. of studies)	Sensitivity (95% CI)	Specificity (95% CI)
Type of rule:		
Wells cutoff value of 2 (n = 19)	84% (78%–89%)	58% (52%–65%)
Wells cutoff value of 4 (n = 11)	60% (49%–69%)	80% (75%–84%)
Gestalt (n = 15)	85% (78%–90%)	51% (39%–63%)
P value Wells 2 vs Wells 4	P <0.001	P <0.001
P value Wells 2 vs Gestalt	P = 0.96	P = 0.31
P value Wells 4 vs Gestalt	P <0.001	P <0.001
Impact prevalence within Wells 2 studies		
Prevalence 5%	67% (58%–75%)	72% (65%–79%)
Prevalence 15%	85% (80%–89%)	58% (52%–63%)
Prevalence 30%	91% (88%–94%)	47% (40%–55%)
P value for trend	P <0.001	P <0.001
Adding D-dimer testing to rule	Failure rate (95% CI)	Efficiency (95% CI)
Wells 4 with quantitative D-dimer (n = 4)	0.5% (0.2%–0.9%)	39% (30%–48%)
Wells 2 with qualitative D-dimer (n = 5)	0.9% (0.5%–1.7%)	40% (32%–49%)

pressed in the review question, and the precision and variability in accuracy results.

Reviews with a comparative question (e.g., is test A better than test B at a specific point in the diagnostic pathway?) can directly examine whether sensitivity or specificity or both are higher for one test than the other. A distinction should be made between primary studies directly comparing the 2 index tests in the same patient (direct evidence) and studies examining only one of these index tests (indirect evidence). Direct evidence is preferred because important factors that may have an impact on accuracy (i.e., potential confounding factors such as the population and choice of reference standard) will be constant when the index tests are compared. If sufficient studies with direct evidence are available, the main analysis or any sensitivity analyses should focus on these studies providing direct evidence.

If both sensitivity and specificity are higher or the entire summary ROC curve for one test is to the left and above that of the other test, the conclusion is straightforward. If sensitivity is higher for one test and specificity for the other or if the summary ROC curves of the 2 tests cross, it is important to examine and weigh the potential negative consequences associated with false-positive or false-negative test results. One way to provide this insight is to subject a hypothetical cohort of 1000 patients to both tests and calculate the number of patients with different correct and incorrect test results based on summary estimates of sensitivity and specificity and a reasonable estimate of the expected prevalence.

The intended role of a test is also helpful in structuring the interpretation of results. In triage questions, the number of missed cases (e.g., false-negative test results) is the key concern, so sensitivity or the negative predictive value are the key accuracy measures. The desired minimum level for these measures will still be a subjective choice and depend on the condition at hand. In our example, most experts will agree that the clinical decision rule should not miss more than 5% of the patients with pulmonary embolism, so therefore sensitivity should be at least 95%. From the results of the rules alone it is clear that a large part of the confidence ellipse and even the summary estimate of sensitivity do not meet this criterion (Table 4). This observation leads to a firm conclusion that clinical decisions alone are not suited for use in the triage of patients suspected to have pulmonary embolism. Therefore, D-dimer results have been added to the triage of patients suspected for pulmonary embolism. In this scenario patients will not undergo further testing if both the clinical decision rule AND the D-dimer are negative. The proportion of patients who had negative results for both tests but who had a final diagnosis of pulmonary embolism (failure rate) has been metaanalyzed. Adding a qualitative D-dimer to the clinical decision rule led to failure rates that were lower than 2% (Table 4). This frequency has been considered sufficiently low and therefore such strategies have been implemented in

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

practice. The efficiencies of such strategies are around 40%, meaning that in 40% of the patients no further testing is required.

Similar to any other review there is the threat of publication bias in DTA reviews *(18)*. Publication bias occurs when studies containing less favorable results are less likely to be published. Summary results based on published findings will then generate an overoptimistic picture of the accuracy of a test. Unfortunately, little information exists regarding the presence and magnitude of publication bias in diagnostic accuracy studies. Unlike randomized trials there are no registries for protocols of diagnostic accuracy studies.

Recent Developments

In this section we highlight some recent developments that are relevant for diagnostic accuracy reviews of biochemical tests and markers.

NETWORK METAANALYSIS

In many diagnostic scenarios there are several alternative tests available, which leads to the key question, which test is the best? Direct comparisons of tests (head-to-head comparison in the same patients by use of a cross-over design or a parallel randomized design) offer the most valid study design but are not always available in the literature. Systematic reviews focusing on more than one diagnostic test have to incorporate indirect comparisons (accuracy of different tests assessed in different populations). Network metaanalyses have been developed in the field of intervention to combine both direct and indirect comparisons within a single statistical model to allow for ranking of the available treatments *(25)*. In addition, these models provide estimates of heterogeneity and inconsistency of effects. Such network metaanalyses would be a welcome addition for ranking and selecting the best test among several alternatives.

IPD METAANALYSIS

IPD metaanalyses use individual patient data rather than published summary results of a study. In an IPD metaanalysis there is more flexibility and more statistical power to examine how patients' characteristics affect diagnostic test accuracy (subgroup analyses or effect modification). IPD metaanalysis also offers more flexibility in handling differences in thresholds for positivity for continuous index test results and for determining the optimal cutoff value *(26)*.

Concluding Remarks

Many improvements have been made in the methodology of performing systematic reviews of the accuracy of diagnostic tests and multivariable diagnostic models. Methods have been improved for locating diagnostic accuracy studies, for assessing the risk for bias and sources of variation, and for developing advanced and flexible models to metaanalyze 2 possible correlated outcomes. However, the biggest obstacle for generating high-quality, clinically useful diagnostic reviews is the poor methodological quality of the existing body of diagnostic accuracy studies reported in the literature. Fortunately, interest in the methods for the evaluation of diagnostic tests has grown considerably in the last decade. Higher-quality and more informative primary studies will in return generate more informative diagnostic reviews.

Appendix 1

Accuracy of diagnostic decision rules without and with D-dimer assay for the diagnosis of pulmonary embolism Pulmonary embolism (PE) is an important condition for physicians to consider because case fatality is high if left untreated. However, diagnosing PE in suspected patients is challenging because signs and symptoms are often nonspecific. Physicians constantly face the dilemma of not wanting to miss a PE while at the same time wanting to avoid performing too many unnecessary additional diagnostic procedures that can be expensive, burdensome, and possibly harmful. Diagnostic strategies in suspected patients therefore focus on identifying patients in whom PE can be safely ruled out on the basis of findings from the patient history and physical examination. Many different diagnostic decision rules for excluding PE on the basis of symptoms and signs, with or without D-dimer assay, have been developed and validated, but there remains uncertainty as to whether these different rules differ in their accuracy in a meaningful way. In this example we focus on 3 rules:

- Wells rule using a cutoff value of 2 for defining a positive (abnormal) test result;
- Wells rule using a cutoff value of 4;
- Gestalt rule.

In the Wells rules, points are scored when certain signs and symptoms (e.g. heart rate >100, previous deep venous thrombosis) are present, resulting in a total score. In the Gestalt rule, physicians provide an overall empirical assessment of the likelihood of pulmonary embolism being present after examination of a patient. To safely exclude pulmonary embolism a D-dimer test can be added to the clinical rule to refrain from further testing if both tests (rule + D-dimer assay) are negative. Further details and more rules can be found in the original review *(11)*.

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022

REVIEW AIMS:

To determine and compare the diagnostic accuracy of 3 different clinical decision rules: Wells-2 (n = 19 studies), Wells-4 (n = 11 studies), and Gestalt rule (n = 15 studies).

To examine whether a negative test from a rule in combination with a negative D-dimer test result is a safe and efficient strategy for excluding PE without referral for further burdening and invasive imaging.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: The Netherlands Organisation for Health Research and Development (ZonMW); K.G.M. Moons, the Netherlands Organisation for Scientific Research (projects 9120.8004 and 918.10.615).
Expert Testimony: None declared.

References

1. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29:E13–21.
2. Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. Clin Chem 2012;58:1292–301.
3. Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PM. Quantifying the added value of a diagnostic test or marker. Clin Chem 2012;58: 1408–17.
4. Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem [Epub ahead of print 2012 Jun 22].
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. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189–202.
7. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, on behalf of the Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149: 889–97.
8. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36.
9. Diagnostic test accuracy working group. http://srdta.cochrane.org/ (Accessed August 2012).
10. Leeflang MM, Debets-Ossenkopp YJ, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, et al. Galactomannan detection for invasive aspergillosis in immunocompromized patients. Cochrane Database Syst Rev 2008:CD007394.
11. Lucassen W, Geersing GJ, Erkens PM, Reitsma JB, Moons KG, Büller H, van Weert HC. Clinical decision rules for excluding pulmonary embolism: a meta-analysis. Ann Intern Med 2011;155: 448–60.
12. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332: 1089–92.
13. Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. J Clin Epidemiol 2005;58:444–9.
14. Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol 2006;59:234–40.
15. Savoie I, Helmer D, Green CJ, Kazanjian A. Beyond Medline: reducing bias through extended systematic review search. Int J Technol Assess Health Care 2003;19:168–78.
16. Fraser C, Mowatt G, Siddiqui R, Burr J. Searching for diagnostic test accuracy studies: an application to screening for open angle glaucoma (OAG) [Abstract]. Cochrane Colloquium Abstracts Journal; 2006. http://www.imbi.uni-freiburg.de/OJS/cca/index.php?journal=cca&page=article&op=view&path[]=1980 (Accessed October 2012).
17. Whiting P, Westwood M, Burke M, Sterne J, Glanville J. Systematic reviews of test accuracy should search a range of databases to identify primary studies. J Clin Epidemiol 2008;61: 357–64.
18. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technol Assess 2000;4:1–115.
19. 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.
20. Linnet K, Bossuyt PMM, Moons KGM, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. Clin Chem 2012;58:1292–301.
21. 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.
22. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865–84.
23. 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.
24. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5–12.
25. Li T, Puhan MA, Vedula SS, Singh S, Dickersin K; the Ad Hoc Network Meta-analysis Methods Meeting Working Group. Network meta-analysis-highly attractive but more methodological research is needed. BMC Med. 2011 27;9:79.
26. Khan KS, Bachmann LM, ter Riet G. Systematic reviews with individual patient data meta-analysis to evaluate diagnostic tests. Eur J Obstet Gynecol Reprod Biol 2003;108:121–5.

Downloaded from https://academic.oup.com/clinchem/article/58/11/1534/5620813 by guest on 30 January 2022