Evidence of bias and variation in diagnostic accuracy studies

Anne W.S. Rutjes, Johannes B. Reitsma, Marcello Di Nisio, Nynke Smidt, Jeroen C. van Rijn, Patrick M.M. Bossuyt

An abridged version of this article appeared in the Feb. 14, 2006, issue of *CMAJ*.

Abstract

Background: Studies with methodologic shortcomings can overestimate the accuracy of a medical test. We sought to determine and compare the direction and magnitude of the effects of a number of potential sources of bias and variation in studies on estimates of diagnostic accuracy.

Methods: We identified meta-analyses of the diagnostic accuracy of tests through an electronic search of the databases MEDLINE, EMBASE, DARE and MEDION (1999–2002). We included meta-analyses with at least 10 primary studies without preselection based on design features. Pairs of reviewers independently extracted study characteristics and original data from the primary studies. We used a multivariable meta-epidemiologic regression model to investigate the direction and strength of the association between 15 study features on estimates of diagnostic accuracy.

Results: We selected 31 meta-analyses with 487 primary studies of test evaluations. Only 1 study had no design deficiencies. The quality of reporting was poor in most of the studies. We found significantly higher estimates of diagnostic accuracy in studies with nonconsecutive inclusion of patients (relative diagnostic odds ratio [RDOR] 1.5, 95% confidence interval [CI] 1.0–2.1) and retrospective data collection (RDOR 1.6, 95% CI 1.1–2.2). The estimates were highest in studies that had severe cases and healthy controls (RDOR 4.9, 95% CI 0.6–37.3). Studies that selected patients based on whether they had been referred for the index test, rather than on clinical symptoms, produced significantly lower estimates of diagnostic accuracy (RDOR 0.5, 95% CI 0.3–0.9). The variance between meta-analyses of the effect of design features was large to moderate for type of design (cohort v. case–control), the use of composite reference standards and the use of differential verification; the variance was close to zero for the other design features.

Interpretation: Shortcomings in study design can affect estimates of diagnostic accuracy, but the magnitude of the effect may vary from one situation to another. Design features and clinical characteristics of patient groups should be carefully considered by researchers when designing new studies and by readers when appraising the results of such studies. Unfortunately, incomplete reporting hampers the evaluation of potential sources of bias in diagnostic accuracy studies.

Cite this article as *CMAJ* 2006;174(4). DOI:10.1503/cmaj.050090

DOI:10.1503/cmaj.050090

Although the number of test evaluations in the literature is increasing, much remains to be desired in terms of methodology. A series of surveys have shown that only a small number of studies of diagnostic accuracy fulfil essential methodologic standards.[1–3]

Shortcomings in the design of clinical trials are known to affect results. The biasing effects of inadequate randomization procedures and differential dropout have been discussed and demonstrated in several publications.[4–6] A growing understanding of the potential sources of bias and variation has led to the development of guidelines to help researchers and readers in the reporting and appraisal of results from randomized trials.[7,8] More recently, similar guidelines have been published to assess the quality of reporting and design of studies evaluating the diagnostic accuracy of tests. For many of the items in these guidelines, there is no or limited empirical evidence available on their potential for bias.[9]

In principle, such evidence can be collected by comparing studies that have design deficiencies with studies of the same test that have no such imperfections. Several large meta-analyses have used a meta-regression approach to account for differences in study design.[10–12] Lijmer and colleagues examined a number of published meta-analyses and showed that studies that involved nonrepresentative patients or that used different reference standards tended to overestimate the diagnostic performance of a test.[13] They looked at the influence of 6 methodologic criteria and 3 reporting features on the estimates of diagnostic accuracy in a limited number of clinical problems.

We conducted this study of a larger and broader set of meta-analyses of diagnostic accuracy to determine the relative importance of 15 design features on estimates of diagnostic accuracy.

Methods

Data sources: systematic reviews

An electronic search strategy was developed to identify all systematic reviews of studies evaluating the diagnostic accuracy of tests that were published between January 1999 and April 2002 in MEDLINE (OVID and PubMed), EMBASE (OVID), the Database of Abstracts of Reviews of Effect (DARE) of the Centre for Reviews and Dissemination (www.york.ac.uk/inst/crd

/darehp.htm) and the MEDION database of the University of Maastricht (www.mediondatabase.nl/) (Appendix 1). The focus was on recent reviews, since we expected a larger number of studies in these and more variety in terms of studies with and without design deficiencies.

Systematic reviews were eligible if they included at least 10 primary studies of the accuracy of the same test, if study selection had not been based on one or more of the design features that we intended to evaluate, and if sensitivity and specificity were provided for at least 90% of the studies in the review (Fig. 1). Languages were restricted to English, German, French and Dutch. If 2 or more reviews addressed the same combination of index test and target condition, we included only the largest one to avoid duplicate inclusion of primary studies.

One of us (A.R.) completed the search and performed the initial selection of systematic reviews on the basis of abstracts and titles. Potentially eligible reviews were independently assessed by 2 researchers (A.R. and N.S., or A.R. and M.D.).

Standardized extraction forms and background documents were prepared for the evaluation of the eligibility of the systematic reviews and for the extraction of data and design features from the primary studies. All assessors attended a training session to become familiar with the use of these forms. No masking of authorship or journal name was applied during this or any of the following phases of the project. Inclusion criteria were tuned during the data extraction of the first few primary studies.

Data sources: primary studies

Paper copies of the reports of all of the primary studies were retrieved once a systematic review was included. We excluded primary studies if we were unable to reproduce the 2×2 tables.

A series of items was extracted from each report that addressed study design, patient group, verification procedure, test execution and interpretation, data collection, statistical analysis and quality of reporting. From this series, we assembled a list of 15 items as potential sources of bias or variation (Appendix 2). These items were selected on the basis of recent systematic reviews of the available literature.[9,14,15] Table 1 displays 9 additional items that were selected to evaluate the quality of reporting.

One epidemiologist (A.R.) assessed all of the articles. A second independent assessment was performed by one member of a team of 5 clinicians and trained epidemiologists (N.S., M.D., J.R., J.vR., P.B.). Disagreements were discussed. If necessary, the ruling of a third assessor (J.R. or P.B.) was decisive.

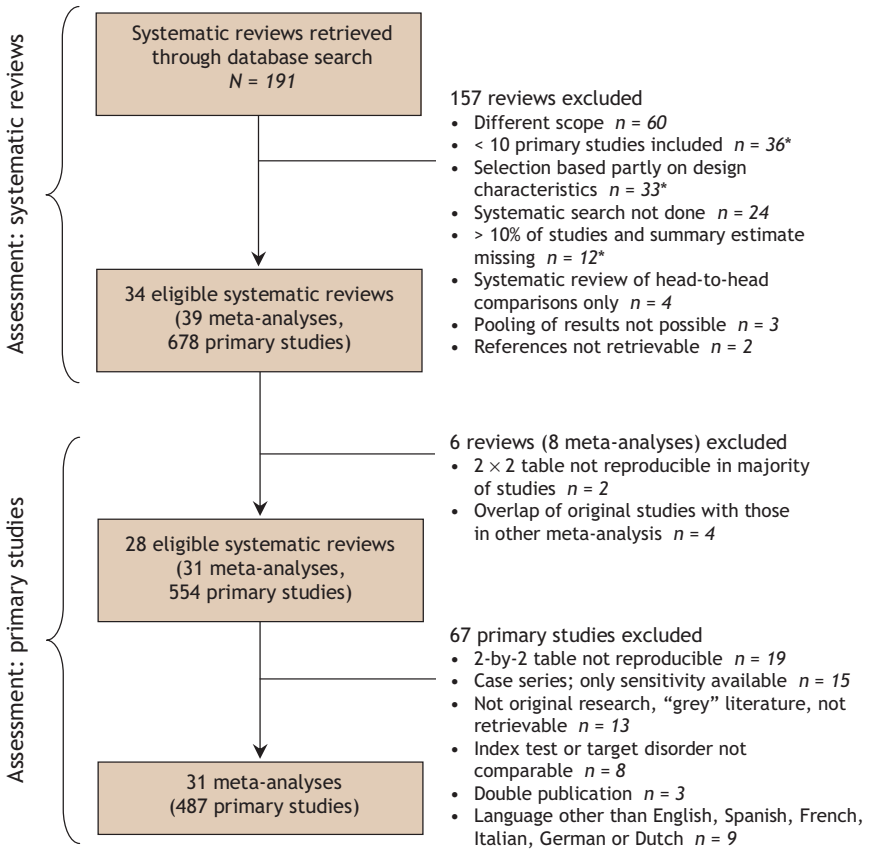


Fig. 1: Process of selecting and assessing systematic reviews and primary studies of the accuracy of diagnostic tests. *Exclusion criteria can overlap.

Data analysis

We used a meta-epidemiologic regression approach to evaluate the effect of design deficiencies on estimates of diagnostic accuracy across the systematic reviews.[16–18] Covariates indicating design features were used to examine whether, on average, studies that failed to meet certain methodologic criteria yielded different estimates of accuracy. The diagnostic odds ratio (DOR) was used as the summary measure of diagnostic accuracy.

Our model can be regarded as a random-effects regression extension of the summary receiver-operating-characteristic (ROC) model used in many systematic reviews of diagnostic accuracy.[19]

We modelled the DOR in a particular study of a test as a function of the summary DOR for that test, the threshold for positivity in that study, the effect of a series of design features, and residual error. We wanted to determine the average effect of the respective design features, expecting that the effect would differ between meta-analyses and that it can be more prominent for one test and less prominent for another. Using a regression approach, we adjusted the effect of one design feature for the potentially confounding effect of other design features. We allowed the DOR to be related to the positivity threshold in each meta-analysis, allowing for an ROC-like relation between sensitivity and specificity across studies in each meta-analysis.

More formally, our model, a single model including all studies from each meta-analysis, expresses the observed (log) DOR d_{ij} in study j in meta-analysis i using the following equation 1:

$$d_{ij} = \alpha_i + \beta_i S_{ij} + \sum_m (\gamma_m + \upsilon_{im}) X_{ijm} + e_{ij}$$

where S_{ij} is the positivity threshold in each study defined as the sum of logit(sensitivity) and logit(1 – specificity); α_i is the overall accuracy of the test studied in meta-analysis i; β_i is the coefficient indicating whether the DOR varies with S in each meta-analysis; X_{ijm} is the value of the design feature covariate m in study j included in meta-analysis i; γ_m is the average effect of feature m across all meta-analyses; and υ_{im} expresses the deviation from that average effect in meta-analysis i, calculated as follows (equation 2):

$$\upsilon_{im} \sim N(0, \sigma^2_{\upsilon m})$$

If the variance of an effect between meta-analyses (υ_{im}) is close or equal to zero, the average effect of a design feature is about the same in each meta-analysis. Larger values of υ_{im} indicate that the magnitude, or even the direction, of that design feature differs substantially from one meta-analysis to another. The error term e_{ij} is also normally distributed as follows (equation 3):

$$e_{ij} \sim N(0, \tau^2_{ij} + \sigma^2)$$

and it combines 2 sources of error: sampling error, which is specific for each study j, and a single residual error term, which is assumed to be constant across meta-analyses. The sampling error or imprecision e of the (log) DOR in each study j, is defined as follows (equation 4):

$$\tau^2_{ij} = \frac{1}{a_{ij}} + \frac{1}{b_{ij}} + \frac{1}{c_{ij}} + \frac{1}{d_{ij}}$$

where a_{ij}, b_{ij}, c_{ij}, d_{ij} are the 4 cells of the 2×2 table of study j in meta-analysis i.

The coefficient γ_m of a particular design feature estimates the change in the log-transformed DOR between studies with and without that feature. It can be interpreted, after antilogarithm transformation, as a relative diagnostic odds ratio (RDOR). It shows the mean DOR of studies with a specific design deficiency relative to the mean DOR of studies without this deficiency. If the relative DOR is larger than 1, it implies that studies with that design deficiency yield larger estimates of the DOR than studies without it.

We used the PROC MIXED procedure of SAS to estimate the parameters of this model (SAS version 9.1, SAS Institute Inc, Cary, NC). This procedure allows for the specification of random effects and the specification of the known variances of the (log) DOR, which can be kept constant (inverse variance method). Further details on how to fit these models can be found in articles by van Houwelingen and colleagues.[16,17]

We used the following multivariable modelling strategy.

Table 1: Quality of reporting study characteristics in 487 studies of the diagnostic accuracy of tests

Characteristic	Reported; no. (%) of studies	
	Yes	No
Dates of inclusion period	238 (49)	249 (51)
Definition of positive and negative results of index test	426 (87)	61 (13)
Definition of positive and negative results of reference standard	362 (74)	125 (26)
Sex or age distribution of study population	406 (83)	81 (17)
No. of readers		
Of index test	198 (41)	289 (59)
Of reference standard	111 (23)	376 (77)
Description of educational background of readers		
Of index test	187 (38)	300 (62)
Of reference standard	131 (27)	356 (73)
Training of readers	26 (5)†	426 (88)
Description of reproducibility of index test or reference standard*	70 (14)	417 (86)
Confidence intervals or standard errors for accuracy measures	81 (17)	406 (83)

*Includes reference to article stating test reproducibility.
†An additional 35 studies (7%) reported that no training was given.

We excluded covariates from the multivariable model when 50% or more of the studies failed to provide information on that design covariate. If that proportion was 10% or less, the corresponding studies were assigned to the potentially flawed category. Otherwise, the nonreported category was kept as such in the analysis. The results of the univariable analysis were used to decide whether categories of a design feature with only a few studies could be grouped together. Categories were combined only if the underlying mechanism of bias was judged to be similar and if the univariable effect estimates were comparable.

Results

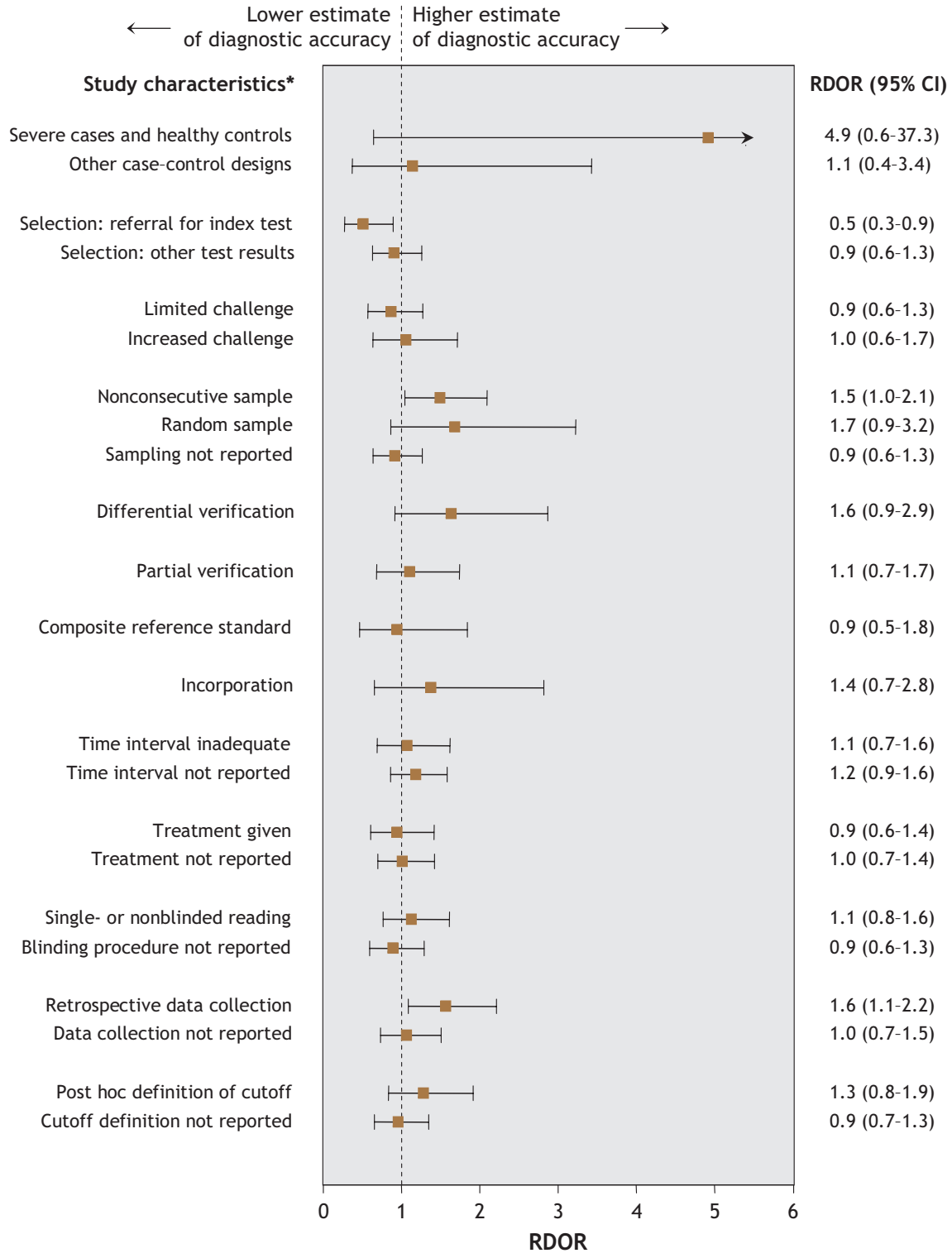
Our search identified 191 potentially eligible systematic reviews, from which we were able to include 31 meta-analyses[20–47] of 487 primary studies (Fig. 1). Two meta-analyses of the same clinical problem but with different restrictions of patient selection were analyzed as one meta-analysis.[20,34] Another meta-analysis had to be split into 4 separate meta-analyses because of differences in test techniques between the studies.[46] Because of the exclusion of some primary studies (Fig. 1) and the splitting of a meta-analysis, 6 meta-analyses had fewer than 10 studies.[20,32,46] The included meta-analyses addressed a wide range of diagnostic problems in different clinical settings (Appendix 3). Index tests varied, from signs and symptoms derived from history taking or physical examination to laboratory tests and imaging tests. This diversity in tests is also reflected in the pooled DORs, which ranged from 1.2 to 565 (median 30).

The characteristics of the included studies are listed in

Table 2: Effect of study characteristics on estimates of diagnostic accuracy from multivariable analysis

Item no.*	Label†	No. of studies / no. of meta-analyses	RDOR (95% CI)	Variance in effect between meta-analyses
1	Cohort‡	445/31	1.0	0.7
	Severe cases and healthy controls	5/2	4.9 (0.6–37.3)	
	Other case-control design	37/7	1.1 (0.4–3.4)	
2	Selection: symptoms/signs‡	160/26	1.0	0.0
	Selection: referral for index test	36/9	0.5 (0.3–0.9)	
	Selection: other test results	291/24	0.9 (0.6–1.3)	
3	No limited challenge‡	359/31	1.0	0.1
	Limited challenge	85/23	0.9 (0.6–1.3)	
	Increased challenge	43/14	1.0 (0.6–1.7)	
4	Consecutive sample‡	130/30	1.0	0.1
	Nonconsecutive sample	173/29	1.5 (1.0–2.1)	
	Random sample	17/6	1.7 (0.9–3.2)	
	Sampling method not described	167/28	0.9 (0.6–1.3)	
5	Same reference standard‡	388/29	1.0	0.2
	Differential verification	99/14	1.6 (0.9–2.9)	
6	Complete verification‡	453/31	1.0	0.0
	Partial verification	34/15	1.1 (0.7–1.7)	
7	Single reference standard‡	395/28	1.0	0.4
	Composite reference standard	92/14	0.9 (0.5–1.8)	
8	No incorporation‡	463/31	1.0	0.0
	Incorporation	24/8	1.4 (0.7–2.8)	
9	Time interval adequate‡	236/28	1.0	0.0
	Time interval inadequate	45/15	1.1 (0.7–1.6)	
	Time interval not reported	206/28	1.2 (0.9–1.6)	
10	Treatment withheld‡	250/28	1.0	0.0
	Treatment given	54/11	0.9 (0.6–1.4)	
	Treatment not reported	183/25	1.0 (0.7–1.4)	
11	Double-blinded reading‡	84/21	1.0	0.0
	Single- or nonblinded reading	187/17	1.1 (0.8–1.6)	
	Blinding procedure not reported	216/17	0.9 (0.6–1.3)	
12	Prospective data collection‡	301/31	1.0	0.1
	Retrospective data collection	106/21	1.6 (1.1–2.2)	
	Data collection not reported	80/22	1.0 (0.7–1.5)	
13	Predefined or standard cutoff‡	338/31	1.0	0.0
	Post hoc definition of cutoff	59/15	1.3 (0.8–1.9)	
	Cutoff definition not reported	90/18	0.9 (0.7–1.3)	

Note: RDOR = relative diagnostic odds ratio estimated in a multivariable random-effects meta-epidemiological regression model.
*Items 14 (noninterpretable results) and 15 (dropouts) were not included in the multivariable analysis because of incomplete reporting (reported in less than 50% of the studies).
†See Appendix 2 for descriptions of labels.
‡Reference category.



*See Appendix 2 for descriptions of the study characteristics.

Fig. 2: Effects of study design characteristics on estimates of diagnostic accuracy. RDOR = relative diagnostic odds ratio (adjusted RDORs were estimated in a multivariable random-effects meta-epidemiologic regression model).

Table 2. Most of the 487 studies used a clinical cohort (445 [91%]), verified all index test results with a reference standard (453 [93%]) and interpreted the reference standard without integrating index test results (463 [95%]). Only 1 study fulfilled all 13 desired design features.

The quality of reporting per item varied, from reasonably good (age and sex distribution, definition of positive and negative index test results, and reference standard results) to poor (Table 1).

The results of the univariable analysis are presented in Appendix 4. Incomplete reporting precluded the investigation of 2 potential sources of bias. Information about noninterpretable test results and information about dropouts were reported in less than 50% of the studies and were therefore not analyzed any further. Of the remaining 13 design features, 6 were not reported in more than 10% of the studies (Table 2).

The relative effects of all of the characteristics in the multivariable model are shown in Table 2 and depicted in Fig. 2. The reference groups listed in Table 2 have, by definition, an RDOR of 1 and are therefore not presented in Fig. 2.

The largest overestimation of accuracy was found in studies that included severe cases and healthy controls (RDOR 4.9, 95% confidence interval 0.6–37). Only 5 studies in 2 meta-analyses used such a design, which explains the broad confidence interval. In addition, the heterogeneity in effect between meta-analyses was large (0.7), because there was severe overestimation in one of the meta-analyses (detection of gram-negative infection with Gelation Limulus amebocyte lysate) and a much smaller effect in the other meta-analysis (detection of lifetime alcohol abuse or dependence with the CAGE questionnaire). The design features associated with a significant overestimation of diagnostic accuracy were nonconsecutive inclusion of patients and retrospective data collection. Random inclusion of eligible patients and differential verification also resulted in higher estimates of diagnostic accuracy, but these effects were not significant. The selection of patients on the basis of whether they had been referred for the index test, rather than on clinical symptoms, was significantly associated with lower estimates of accuracy.

The RDORs presented in Table 2 and Fig. 2 are average effects across different meta-analyses, and effects varied between meta-analyses. The amount of variance between meta-analyses provides an indication of the heterogeneity of an effect (Table 2). Moderate to large differences were found for study design (cohort v. case–control design), the use of composite reference standards and differential verification. For the other design features, the variance between meta-analyses was close to zero.

Interpretation

Our analysis has shown that differences in study design and patient selection are associated with variations in estimates of diagnostic accuracy. Accuracy was lower in studies that selected patients on the basis of whether they had been referred for the index test rather than on clinical symptoms, whereas it was significantly higher in studies with nonconsecutive inclusion of patients and in those with retrospective data collection. Comparable or even higher estimates of diagnostic accuracy occurred in studies that included severe cases and healthy controls and in those in which 2 or more reference standards were used to verify index test results, but the corresponding confidence intervals were wider in these studies.

We found that studies that used retrospective data collection or that routinely collected clinical data were associated with an overestimation of the DOR by 60%. In studies in which data collection is planned after all index tests have been performed, researchers may find it difficult to use unambiguous inclusion criteria and to identify patients who received the index test but whose test results were not subsequently verified.[48,49]

Studies that used nonconsecutive inclusion of patients were associated with an overestimation of the DOR by 50% compared with those that used a consecutive series of patients. Studies conducted early in the evaluation of a test may have preferentially excluded more complex cases, which may have led to higher estimates of diagnostic accuracy. Yet if clear-cut cases are excluded, because the reference standard is costly or invasive, diagnostic accuracy will be underestimated. These 2 mechanisms, with opposing effects, may explain why other studies have reported different results, either lower estimates of accuracy in studies with nonconsecutive inclusion[50] or, on average, no effect on accuracy estimates.[13]

We found that studies that selected patients on the basis of whether they had been referred for the index test or on the basis of previous test results tended to lower diagnostic accuracy compared with studies that set out to include all patients with prespecified symptoms. The interpretation of this finding is not straightforward. We speculate that, with this form of patient selection, patients strongly suspected of having the target condition may bypass further testing, whereas those with a low likelihood of having the condition may never be tested at all. These mechanisms tend to lower the proportion of true-positive and true-negative test results.[51]

An extreme form of selective patient inclusion occurred in the studies that included severe cases and healthy controls. These case–control studies had much higher estimates of diagnostic accuracy (RDOR 4.9), although the low number of such studies led to wide confidence intervals. Severe cases are easier to detect with the use of the index test, which would lead to higher estimates of sensitivity in studies with more severe cases.[52] The inclusion of healthy controls is likely to lower the occurrence of false-positive results, thereby increasing specificity.[52] Other studies have also reported overestimation of diagnostic accuracy in this type of case–control studies.[13,50]

Verification is a key issue in any diagnostic accuracy study. Studies that relied on 2 or more reference standards to verify the results of the index test reported odds ratios that were on average 60% higher than the odds ratios in studies that used a single reference standard. The origin of this difference probably resides in differences between reference standards in how

they define the target conditions or in their quality.[53] If misclassifications by the second reference standard are correlated with index test errors, agreement will artificially increase, which would lead to higher estimates of diagnostic accuracy. Our result is in line with that of the study by Lijmer and colleagues,[13] who reported a 2-fold increase with a confidence interval overlapping ours.

As in the study by Lijmer and colleagues, we were unable to demonstrate a consistent effect of partial verification. This may be because the direction and magnitude of the effect of partial verification is difficult to predict. If a proportion of negative test results is not verified, this tends to increase sensitivity and lower specificity, which may leave the odds ratio unchanged.[54]

We were unable to demonstrate significant associations between estimates of DOR and a number of design features. The absence of an association in our model does not imply that the design features should be ignored in any given accuracy study, since the effect of design differences may vary between meta-analyses, or even within a single meta-analysis.

The results of our study need to be interpreted with the following limitations and strengths in mind. We were hampered by the low quality of reporting in the studies. Several design-related characteristics could not be adequately examined because of incomplete reporting (e.g., frequency of indeterminate test results and of dropouts, patient selection criteria, clinical spectrum, and the degree of blinding). We used the odds ratio as our main accuracy measure, which is a convenient summary statistic,[55,56] but it may be insensitive to phenomena that produce opposing changes in sensitivity and specificity. Further studies should explore the effects of these design features on other accuracy measures, such as sensitivity, specificity and likelihood ratios.

Our study can be seen as a validation and extension of the study of Lijmer and colleagues.[13] To ensure independent validation, we did not include any of their meta-analyses in our study. Furthermore, we replaced the fixed-effects approach used by them with a more appropriate random-effects approach, which allowed the design covariates to vary between meta-analyses. This explains the wider confidence intervals in our study, despite the fact that we included 269 studies more than Lijmer and colleagues did.

In general, the results of our study provide further empirical evidence of the importance of design features in studies of diagnostic accuracy. Studies of the same test can produce different estimates of diagnostic accuracy depending on choices in design. We feel that our results should be taken into account by researchers when designing new primary studies as well as by reviewers and readers who appraise these studies. Initiatives such as STARD (Standards for Reporting of Diagnostic Accuracy [www.consort-statement.org/stardstatement.htm]) should be endorsed to improve the awareness of design features, the quality of reporting and, ultimately, the quality of study designs. Well-reported studies with appropriate designs will provide more reliable information to guide decisions on the use and interpretation of test results in the management of patients.

This article has been peer reviewed.

From the Deptartment of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands (Rutjes, Reitsma, van Rijn, Bossuyt); the Department of Medicine and Aging, School of Medicine, and Aging Research Center, Gabriele D'Annunzio University Foundation, Chieti-Pescara, Italy (Di Nisio); and the Institute for Research in Extramural Medicine, VU University Medical Center, Amsterdam, the Netherlands (Smidt)

Competing interests: None declared.

Contributors: Johannes Reitsma and Patrick Bossuyt initiated and supervised the study. Anne Rutjes wrote the first draft of the study protocol, designed and established the database and wrote the first draft of the article. All of the authors collected the data. Anne Rutjes and Johannes Reitsma analyzed the data and, along with Patrick Bossuyt, provided the first interpretation of the implications of the study results. All of the authors contributed to the final manuscript and gave final approval of the version to be published. Patrick Bossuyt is the guarantor.

Acknowledgements: We thank Jeroen G. Lijmer for his useful comments on earlier drafts of the study protocol and for securing project funding. We also thank Aeilko H. Zwinderman and Augustinus A. Hart for their statistical input.

The study was funded by a research grant from the Netherlands organization for scientific research (NWO; registration no. 945-10-012). The funding source had no involvement in the development of the study design, the collection, analysis and interpretation of the data, the writing of the report or the decision to submit the paper for publication.

REFERENCES

1. 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.
2. Harper R, Reeves B. Compliance with methodological standards when evaluating ophthalmic diagnostic tests. *Invest Ophthalmol Vis Sci* 1999;40:1650-7.
3. Estrada CA, Bloch RM, Antonacci D, et al. Reporting and concordance of methodologic criteria between abstracts and articles in diagnostic test studies. *J Gen Intern Med* 2000;15:183-7.
4. Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. *JAMA* 1995;273:408-12.
5. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609-13.
6. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42-6.
7. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Control Clin Trials* 1996;17:1-12.
8. Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol* 1998;51:1235-41.
9. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189-202.
10. Romagnuolo J, Bardou M, Rahme E, et al. Magnetic resonance cholangiopancreatography: a meta-analysis of test performance in suspected biliary disease. *Ann Intern Med* 2003;139:547-57.
11. Nederkoorn PJ, van der Graaf Y, Hunink MG. Duplex ultrasound and magnetic resonance angiography compared with digital subtraction angiography in carotid artery stenosis: a systematic review. *Stroke* 2003;34:1324-32.
12. Whiting P, Rutjes AW, Dinnes J, et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol* 2005;58:1-12.
13. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
14. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;49:7-18.
15. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem* 2003;49:1-6.
16. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002;21:589-624.
17. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Stat Med* 1993;12:2273-84.
18. Sterne JA, Juni P, Schulz KF, et al. Statistical methods for assessing the influence of study characteristics on treatment effects in "meta-epidemiological" research. *Stat Med* 2002;21:1513-24.
19. 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.

20. Balk EM, Ioannidis JPA, Salem D, et al. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. *Ann Emerg Med* 2001;37:478-94.
21. Berger MY, van der Velden JJ, Lijmer JG, et al. Abdominal symptoms: Do they predict gallstones? A systematic review. *Scand J Gastroenterol* 2000;35:70-6.
22. Deville WL, van der Windt DA, Dzaferagic A, et al. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. *Spine* 2000;25:1140-7.
23. Fiellin DA, Reid MC, O'Connor PG. Screening for alcohol problems in primary care: a systematic review. *Arch Intern Med* 2000;160:1977-89.
24. Gould MK, Maclean CC, Kuschner WG, et al. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *JAMA* 2001;285:914-24.
25. Hobby JL, Tom BD, Bearcroft PW, et al. Magnetic resonance imaging of the wrist: diagnostic performance statistics. *Clin Radiol* 2001;56:50-7.
26. Hoffman RM, Clanon DL, Littenberg B, et al. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. *J Gen Intern Med* 2000;15:739-48.
27. Hoogendam A, Buntinx F, de Vet HC. The diagnostic value of digital rectal examination in primary care screening for prostate cancer: a meta-analysis. *Fam Pract* 1999;16:621-6.
28. Huicho L, Campos-Sanchez M, Alamo C. Metaanalysis of urine screening tests for determining the risk of urinary tract infection in children. *Pediatr Infect Dis J* 2002; 21:1-11.
29. Hurley JC. Concordance of endotoxemia with gram-negative bacteremia. A meta-analysis using receiver operating characteristic curves. *Arch Pathol Lab Med* 2000; 124:1157-64.
30. Kelly S, Harris KM, Berry E, et al. A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma. *Gut* 2001;49:534-9.
31. Kim C, Kwok YS, Heagerty P, et al. Pharmacologic stress testing for coronary disease diagnosis: a meta-analysis. *Am Heart J* 2001;142:934-44.
32. Koelemay MJ, Lijmer JG, Stoker J, et al. Magnetic resonance angiography for the evaluation of lower extremity arterial disease: a meta-analysis. *JAMA* 2001;285: 1338-45.
33. Kwok Y, Kim C, Grady D, et al. Meta-analysis of exercise testing to detect coronary artery disease in women. *Am J Cardiol* 1999;83:660-6.
34. Lau J, Ioannidis JP, Balk EM, et al. Diagnosing acute cardiac ischemia in the emergency department: a systematic review of the accuracy and clinical effect of current technologies. *Ann Emerg Med* 2001;37:453-60.
35. Lederle FA, Simel DL. Does this patient have abdominal aortic aneurysm? *JAMA* 1999;281:77-82.
36. Li J. Capnography alone is imperfect for endotracheal tube placement confirmation during emergency intubation. *J Emerg Med* 2001;20:223-9.
37. Mitchell MF, Cantor SB, Brookner C, et al. Screening for squamous intraepithelial lesions with fluorescence spectroscopy. *Obstet Gynecol* 1999;94(Suppl 1):889-96.
38. Mol BW, Lijmer JG, van der Meulen J, et al. Effect of study design on the association between nuchal translucency measurement and Down syndrome. *Obstet Gynecol* 1999;94:864-9.
39. Nelemans PJ, Leiner T, de Vet HC, et al. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. *Radiology* 2000;217:105-14.
40. Safriel Y, Zinn H. CT pulmonary angiography in the detection of pulmonary emboli: a meta-analysis of sensitivities and specificities. *Clin Imaging* 2002;26:101-5.
41. Sloan NL, Winikoff B, Haberland N, et al. Screening and syndromic approaches to identify gonorrhea and chlamydial infection among women. *Stud Fam Plann* 2000; 31:55-68.
42. Smith Bindman R, Hosmer W, Feldstein VA, et al. Second-trimester ultrasound to detect fetuses with Down syndrome: a meta-analysis. *JAMA* 2001;285:1044-55.
43. Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. *Acad Radiol* 2001;8:149-57.
44. Vasquez TE, Rimkus DS, Hass MG, et al. Efficacy of morphine sulfate-augmented hepatobiliary imaging in acute cholecystitis. *J Nucl Med Technol* 2000;28:153-5.
45. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US — a meta-analysis. *Radiology* 2000;216: 67-77.
46. Westwood ME, Kelly S, Bery E, et al. Use of magnetic resonance angiography to select candidates with recently symptomatic carotid stenosis for surgery: systematic review. *BMJ* 2002;324:198-201.
47. Wiese W, Patel SR, Patel SC, et al. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. *Am J Med* 2000;108:301-8.
48. Oostenbrink R, Moons KG, Bleeker SE, et al. Diagnostic research on routine care data: prospects and problems. *J Clin Epidemiol* 2003;56:501-6.
49. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J Clin Epidemiol* 2003;56:1118-28.
50. Pai M, Flores LL, Pai N, et al. Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis. *Lancet Infect Dis* 2003;3:633-43.
51. Sackett DL, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA, editor. *The evidence base of clinical diagnosis.* London (UK): BMJ Publishing Group; 2002. p. 19-38.
52. Rutjes AW, Reitsma JB, Vandenbroucke JP, et al. Case–control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005;51:1335-41.
53. Whiting P, Rutjes AW, Reitsma JB, et al. 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.
54. Pepe MS. Incomplete data and imperfect reference tests. In*: The statistical evaluation of medical tests for classification and prediction.* Oxford (UK): Oxford University Press; 2004. p. 168-213.
55. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129-35.
56. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res* 2002;2:4.

Correspondence to: *Dr. Anne W.S. Rutjes, Department of Clinical Pharmacology and Epidemiology, Consorzio Mario Negri Sud, Via Nazionale 8, 66030 Santa Maria Imbaro, Chieti, Italy; fax +39 087 2570206*

Editor's take

- Clinicians need to know the diagnostic accuracy of the medical tests they use. Yet, determinations of test characteristics (sensitivity, specificity and likelihood ratios) derived from comparisons with a "gold standard" vary markedly between studies.

- In this study, the authors examined the sources of variation across 15 design features of 487 published studies of diagnostic accuracy. Only 1 study had no design deficiencies. Estimates of accuracy were highest in studies that selected nonconsecutive patients, that used severe cases and healthy controls and that analyzed retrospective data.

Implications for practice: The marked variation in estimates should make clinicians cautious when reading studies reporting on the diagnostic accuracy of tests. It is important that such studies be properly designed and reported.

Appendix 1: Search terms used to retrieve systematic reviews of diagnostic accuracy studies

MEDLINE [OVID]

1. exp diagnostic imaging/
2. exp diagnostic tests, routine/
3. "sensitivity and specificity"/
4. review.pt.
5. meta analysis.pt.
6. meta-analysis/
7. 1 or 2 or 3
8. 4 or 5 or 6
9. 7 and 8
10. limit 9 to yr=1999
11. limit 9 to yr=2000
12. limit 9 to yr=2001
13. limit 9 to yr=2002
14. Editorial.pt.
15. Letter.pt.
16. Comment.pt.
17. 14 or 15 or 16
18. 10 not 17
19. 11 not 17
20. 12 not 17
21. 13 not 17

MEDLINE (PubMed)

("meta-analysis"[Publication Type] OR "meta-analysis"[MeSH Terms] OR "review"[Publication Type]) AND ("sensitivity and specificity"[MeSH Terms] OR "diagnostic imaging"[MeSH Terms] OR "diagnostic tests, routine"[MeSH Terms]) AND (("humans"[MeSH Terms]) AND ("1999"[PDAT] : "2002"[PDAT]))

EMBASE [OVID]

1. exp diagnostic imaging/
2. exp diagnostic tests, routine/
3. "sensitivity and specificity"/
4. meta-analysis/
5. review.pt.
6. 1 or 2 or 3
7. 4 or 5
8. 6 and 7
9. limit 8 to yr=2002
10. limit 8 to yr=2001
11. limit 8 to yr=2000
12. limit 8 to yr=1999
13. Conference Paper.pt.
14. Editorial.pt.
15. Letter.pt.
16. 13 or 14 or 15
17. 9 not 16
18. 10 not 16
19. 11 not 16
20. 12 not 16

MEDION

1. DR (diagnostic reviews)
2. limit 1 to yr = 1999
3. limit 1 to yr = 2000
4. limit 1 to yr = 2001
5. limit 1 to yr = 2002

DARE

A staff member of the Centre for Reviews and Dissemination (CRD) provided an endnote database containing all systematic reviews for 2001 and 2002 that were identified by CRD as systematic reviews of either therapeutic or diagnostic studies. Search strategies and selection procedures used by CRD to retrieve systematic reviews for the DARE database can be found online (http://agatha.york.ac.uk/faq2.ht)

Appendix 2: Sources of bias and variation: definitions of items and background information

Item	Label	Description
Patient group		*The accuracy of a test may vary between patient groups that differ in disease severity, comorbid conditions or alternative diagnoses*[9,14,52,53]
1	Cohort	Cohort design, where the index test is performed before the reference standard[49,52]
	Severe cases and healthy controls	Case–control design selecting severe cases and healthy controls[13,50,52]
	Other case–control design	Case–control design avoiding selection from extreme ends of the spectrum[52]
2	Selection: symptoms/signs	Patient selection based on symptoms or signs of target condition only
	Selection: referral for index test	Patient selection based on referral of patient for index test[49]
	Selection: other test results	Patient selection based on other test results or referral of patient for reference standard[49]
3	No limited challenge	No additional criteria to exclude patients with specific features that may lead to false-negative or false-positive index test results[9,14,52]
	Limited challenge	Additional criteria to exclude patients with specific features that may lead to false-negative or false-positive index test results[9,14,52]
	Increased challenge	Preferential inclusion of patients with specific features that may lead to false-negative or false-positive index test results
4	Consecutive sample	Consecutive inclusion of all patients fulfilling selection criteria[14,50,53]
	Nonconsecutive sample	Nonconsecutive inclusion of patients or cases (case–control design)[14,50,53]
	Random sample	Inclusion of random subsample of patients fulfilling selection criteria[14,53]
Verification procedure		*Ideally, all results of index test are verified with those of one, independent reference standard. Verification is instant, without intervening treatment*[9,54]
5	Same reference standard	All results of index test verified with the same reference standard
	Differential verification	Subset of index test results verified with an alternative reference standard[13,54]
6	Complete verification	All index test results verified with a reference standard
	Partial verification	Only subset of index test results verified with reference standard[13,54]
7	Single reference standard	Reference standard is single test or procedure
	Composite reference standard	Reference standard is combination of tests or procedures
8	No incorporation	Index test not incorporated as part of reference standard
	Incorporation	Index test incorporated as part of reference standard[9,53]
9	Time interval adequate	Acceptable time window between index test and reference standard
	Time interval inadequate	Unacceptable time window between index test and reference standard[9,14,53]
10	Treatment withheld	No treatment given to patients between index test and reference standard
	Treatment given	Treatment given between index test and reference standard[9,14,53]
Interpretation/reading		*Knowledge of the result of the reference standard while reading the result of the index test, or vice versa, may enhance agreement*
11	Double-blinded reading	Results of index test or reference standard interpreted without knowledge of the results of the other test
	Single- or nonblinded reading	Results of index test or reference standard, or both, interpreted without blinding[9,14,53]
Data collection		*Prospective data collection enables researchers to obtain high-quality data. Retrospective data collection is more vulnerable to missing data and incomplete patient flow*[14]
12	Prospective data collection	Data collection planned before performance of index test and reference standard
	Retrospective data collection	Data collection planned after performance of all index tests and reference standards[14]
Analysis		*Choices during data analysis may affect estimates of accuracy, including choice of cutoff value for positivity and exclusion of noninterpretable test results*[9,14,53]
13	Predefined or standard cutoff	Cutoff value for positivity of index test results defined before start of data collection[9]
	Post hoc definition of cutoff	Cutoff value for positivity defined post hoc after completion of data collection[9]
14	Noninterpretable results reported	Number of indeterminate and noninterpretable test results and outliers explicitly reported
	Noninterpretable results not reported	Number of indeterminate and noninterpretable test results and outliers not reported[9,14,53]
15	No dropouts	Data on more than 90% of the included patients were available for the analysis
	Dropouts	Data on less than 90% of the included patients were available for the analysis[9,14,53]

Appendix 3: Characteristics of selected meta-analyses of studies evaluating diagnostic accuracy of tests

Meta-analysis	Diagnostic problem	Type of index test	No. of studies
Balk et al[20]	Emergency department diagnosis of acute myocardial infarction	Biomarker: creatine kinase (CK)-MB	9
Berger et al[21]	Diagnosis of gallstones	Symptom: upper abdominal pain	12
Devillé et al[22]	Workup of herniated discs in patients selected for surgery	Test of Lasegue	11
Fiellin et al[23]	Screening for lifetime alcohol abuse or dependence in primary care settings	CAGE questionnaire	14
Gould et al[24]	Workup of pulmonary nodules	Positron emission tomography with the glucose analog 18-fluorodeoxyglucose (FDG-PET)	29
Hobby et al[25]	Diagnosis of complete tears of the triangular fibrocartilage complex in the wrist	MRI	11
Hoffman et al[26]	Workup of prostate cancer in men with nonspecific elevations of prostate specific antigen levels	Free:total prostate-specific antigen ratio	21
Hoogendam et al[27]	Primary care screening for prostate cancer	Digital rectal examination	13
Huicho et al[28]	Screening for urinary tract infection in children	Urine marker: dipstick nitrate	18
Hurley[29]	Diagnosis of gram-negative infection	Gelation Limulus amebocyte lysate	27
Kelly et al[30]	Workup of staging of gastroesophageal cancer	Endoscopic ultrasonography	13
Kim et al[31]	Diagnosis of coronary artery disease	Dobutamine echocardiography	40
Koelemay et al[32]	Evaluation of lower-extremity arterial disease in aortoiliac tract	3-dimensional magnetic resonance angiography (MRA)	9
Kwok et al[33]	Detection of coronary artery disease in women	Exercise electrocardiography	19
Lau et al[34]	Emergency department diagnosis of acute myocardial infarction	Biomarker: CK-MB	10
Lederle et al[35]	Screening for abdominal aortic aneurysm	Abdominal palpation	10
Li[36]	Confirmation of endotracheal tube placement	Capnography: end-tidal CO_2 devices	10
Mitchell et al[37]	Screening for squamous intraepithelial lesions of the cervix	Papanicolaou smear screening	17
Mol et al[38]	Screening for Down's syndrome	Ultrasonographic marker: nuchal translucency measurement	23
Nelemans et al[39]	Evaluation of peripheral arterial disease	2-dimensional time-of-flight MRA	13
Safriel et al[40]	Diagnosis of pulmonary emboli	CT pulmonary angiography	10
Sloan et al[41]	Diagnosis of gonorrhea and chlamydial infection	Sign: abdominal/lower-abdominal pain	14
Smith Bindman et al[42]	Screening for Down's syndrome	Ultrasonographic marker: femoral shortening	28
Sonnad et al[43]	Workup of staging of prostate cancer	MRI	21
Vasquez et al[44]	Workup of acute cholecystitis	Morphine sulfate-augmented hepatobiliary imaging	15
Visser et al[45]	Workup of peripheral arterial stenosis	Colour-guided duplex ultrasonography	17
Westwood et al[46]	Selecting candidates with recently symptomatic carotid artery stenosis for surgery	3-dimensional contrast-enhanced MRA	7
		2-dimensional contrast-enhanced MRA	7
		3-dimensional time-of-flight MRA	5
		2-dimensional time-of-flight MRA	5
Wiese et al[47]	Diagnosis of vaginal trichomoniasis	Wet-mount smear technique	29

Appendix 4: Effect of study characteristics on estimates of diagnostic accuracy from univariable analysis

Item no.	Label*	No. (%) of studies	RDOR (95% CI)
1	Cohort†	445 (91)	1.0
	Severe cases and healthy controls	5 (1)	4.3 (0.5–38.0)
	Other case-control design	37 (8)	1.0 (0.3–3.3)
2	Selection: symptoms/signs†	160 (33)	1.0
	Selection: referral for index test	36 (7)	0.6 (0.3–1.3)
	Selection: other test results‡	122 (25)	1.0 (0.6–1.6)
	Selection: referral for reference standard‡	150 (31)	1.0 (0.7–1.6)
	Selection procedure not reported‡	19 (4)	1.0 (0.5–2.3)
3	No limited challenge†	359 (74)	1.0
	Limited challenge	85 (17)	0.9 (0.6–1.3)
	Increased challenge	43 (9)	1.0 (0.6–1.7)
4	Consecutive sample†	130 (27)	1.0
	Nonconsecutive sample	173 (36)	1.5 (1.1–2.1)
	Random sample	17 (3)	1.7 (0.9–3.1)
	Sampling method not described	167 (34)	1.0 (0.7–1.4)
5	Same reference standard†	388 (80)	1.0
	Differential verification	99 (20)	1.5 (0.9–2.6)
6	Complete verification†	453 (93)	1.0
	Partial verification	34 (7)	1.0 (0.6–1.6)
7	Single reference standard†	395 (81)	1.0
	Composite reference standard‡	78 (16)	1.2 (0.6–2.2)
	Composition of reference standard not reported‡	14 (3)	1.2 (0.5–2.9)
8	No incorporation†	463 (95)	1.0
	Incorporation‡	17 (3)	1.2 (0.4–3.1)
	Not reported whether index test results were integrated‡	7 (1)	1.4 (0.6–3.3)
9	Time interval adequate†	236 (48)	1.0
	Time interval inadequate	45 (9)	1.2 (0.8–1.8)
	Time interval not reported	206 (42)	1.3 (0.9–1.7)
10	Treatment withheld†	250 (51)	1.0
	Treatment given	54 (11)	1.0 (0.6–1.4)
	Treatment not reported	183 (38)	1.1 (0.8–1.5)
11	Double-blinded reading†	84 (17)	1.0
	Nonblinded reading‡	24 (5)	1.0 (0.5–2.1)
	Single-blinded reading‡	163 (33)	1.1 (0.8–1.6)
	Blinding procedure not reported	216 (44)	0.9 (0.6–1.3)
12	Prospective data collection†	301 (62)	1.0
	Retrospective data collection	106 (22)	1.4 (1.0–1.9)
	Data collection not reported	80 (16)	1.0 (0.7–1.4)
13	Predefined or standard cutoff†	338 (69)	1.0
	Post hoc definition of cutoff	59 (12)	1.2 (0.8–1.8)
	Cutoff definition not reported	90 (19)	1.0 (0.7–1.4)
14	Noninterpretable results reported†	123 (25)	1.0
	Noninterpretable results not reported	364 (75)	0.7 (0.6–0.9)
15	No dropouts†	52 (11)	1.0
	Dropouts	29 (6)	0.6 (0.4–1.1)
	Dropouts not reported	406 (83)	1.2 (0.8–1.7)

RDOR: relative diagnostic odds ratio estimated in a univariable random-effects meta-epidemiological regression model.
*See Appendix 2 for descriptions of labels.
†Reference category.
‡Categories that were combined for multivariable analysis (see Methods).