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

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/rd...
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 $2 \times 2$ 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. 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.rR., 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.**

<table>
<thead>
<tr>
<th>Systematic reviews retrieved through database search</th>
<th>N = 191</th>
</tr>
</thead>
<tbody>
<tr>
<td>34 eligible systematic reviews (39 meta-analyses, 678 primary studies)</td>
<td></td>
</tr>
<tr>
<td>28 eligible systematic reviews (31 meta-analyses, 554 primary studies)</td>
<td></td>
</tr>
<tr>
<td>31 meta-analyses (487 primary studies)</td>
<td></td>
</tr>
<tr>
<td>157 reviews excluded</td>
<td></td>
</tr>
<tr>
<td>Different scope $n = 60$</td>
<td></td>
</tr>
<tr>
<td>$&lt;10$ primary studies included $n = 36$</td>
<td></td>
</tr>
<tr>
<td>Selection based partly on design characteristics $n = 32$</td>
<td></td>
</tr>
<tr>
<td>Systematic search not done $n = 24$</td>
<td></td>
</tr>
<tr>
<td>$&gt;10%$ of studies and summary estimate missing $n = 12$</td>
<td></td>
</tr>
<tr>
<td>Systematic review of head-to-head comparisons only $n = 4$</td>
<td></td>
</tr>
<tr>
<td>Pooling of results not possible $n = 3$</td>
<td></td>
</tr>
<tr>
<td>References not retrievable $n = 2$</td>
<td></td>
</tr>
<tr>
<td>6 reviews (8 meta-analyses) excluded</td>
<td></td>
</tr>
<tr>
<td>$2 \times 2$ table not reproducible in majority of studies $n = 2$</td>
<td></td>
</tr>
<tr>
<td>Overlap of original studies with those in other meta-analysis $n = 4$</td>
<td></td>
</tr>
<tr>
<td>67 primary studies excluded</td>
<td></td>
</tr>
<tr>
<td>2-by-2 table not reproducible $n = 19$</td>
<td></td>
</tr>
<tr>
<td>Case series; only sensitivity available $n = 15$</td>
<td></td>
</tr>
<tr>
<td>Not original research, “grey” literature, not retrievable $n = 13$</td>
<td></td>
</tr>
<tr>
<td>Index test or target disorder not comparable $n = 8$</td>
<td></td>
</tr>
<tr>
<td>Double publication $n = 3$</td>
<td></td>
</tr>
<tr>
<td>Language other than English, Spanish, French, Italian, German or Dutch $n = 9$</td>
<td></td>
</tr>
</tbody>
</table>
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. 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.

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:

$$d_{ij} = \alpha_i + \beta_j S + \sum_m (\gamma_m + \nu_m) X_{jm} + e_{ij}$$

where $S_j$ is the positivity threshold in each study defined as the sum of logit(sensitivity) and logit(1 – specificity); $\alpha_i$ is the overall accuracy of the test studied in meta-analysis $i$; $\beta_j$ is the coefficient indicating whether the DOR varies with $S$ in each meta-analysis; $X_{jm}$ is the value of the design feature covariate $m$ in study $j$ included in meta-analysis $i$; $\gamma_m$ is the average effect of feature $m$ across all meta-analyses; and $\nu_m$ expresses the deviation from that average effect in meta-analysis $i$, calculated as follows (equation 2):

$$\nu_m \sim N(0, \sigma^2_m)$$

If the variance of an effect between meta-analyses ($\nu_m$) is close or equal to zero, the average effect of a design feature is about the same in each meta-analysis. Larger values of $\nu_m$ 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):

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

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):

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

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

The coefficient $\gamma_m$ 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.

We used the following multivariable modelling strategy.

<table>
<thead>
<tr>
<th>Table 1: Quality of reporting study characteristics in 487 studies of the diagnostic accuracy of tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
</tr>
<tr>
<td>Dates of inclusion period</td>
</tr>
<tr>
<td>Definition of positive and negative results of index test</td>
</tr>
<tr>
<td>Definition of positive and negative results of reference standard</td>
</tr>
<tr>
<td>Sex or age distribution of study population</td>
</tr>
<tr>
<td>No. of readers</td>
</tr>
<tr>
<td>Of index test</td>
</tr>
<tr>
<td>Of reference standard</td>
</tr>
<tr>
<td>Description of educational background of readers</td>
</tr>
<tr>
<td>Of index test</td>
</tr>
<tr>
<td>Of reference standard</td>
</tr>
<tr>
<td>Training of readers</td>
</tr>
<tr>
<td>Description of reproducibility of index test or reference standard*</td>
</tr>
<tr>
<td>Confidence intervals or standard errors for accuracy measures</td>
</tr>
</tbody>
</table>

*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 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. Another meta-analysis had to be split into 4 separate meta-analyses because of differences in test techniques between the studies. 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.

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

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.
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).
The inclusion of healthy controls is likely to reduce the occurrence of false-positive results, thereby increasing specificity. Other studies have also reported overestimation of diagnostic accuracy in this type of case–control studies. 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 were selected and verified. 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 mechanisms, with opposing effects, may explain why other studies have reported different results, either lower estimates of accuracy in studies with nonconsecutive inclusion or, on average, no effect on accuracy estimates.

We found that 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.

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 or, on average, no effect on accuracy estimates.

We found that studies that selected 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,
they define the target conditions or in their quality.53 If mis-
classifications 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 Liemier and colleagues,78 who reported a 2-fold increase with a confidence interval overlapping ours.

As in the study by Liemier 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 sen-
sitivity 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 fea-
tures. 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 fol-
lowing 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 indetermi-
nate 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 conven-
tient summary statistic.55,56 but it may be insensitive to phe-
nomena that produce opposing changes in sensitivity and specificity. Further studies should explore the effects of these design features on other accuracy measures, such as sensitiv-
ity, specificity and likelihood ratios.

Our study can be seen as a validation and extension of the study of Liemier and colleagues.53 To ensure independent vali-
dation, 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 Liemier and colleagues did.

In general, the results of our study provide further empiri-
cal 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 ac-
count 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 Diag-
nostic Accuracy [www.consort-statement.org/stardstatement .html]) 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 appropri-
ate 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 Department 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 Uni-
versity Foundation, Chieti-Pescara, Italy (Di Nisio); and the Institute for Re-
search 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. Liemier 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 organiza-
tion for scientific research (NWO; registration no. 945-10-012). The funding source had no involvement in the development of the study design, the col-
lection, analysis and interpretation of the data, the writing of the report or the decision to submit the paper for publication.

REFERENCES

3. Estrada CA, Bloch RM, Antonacci D, et al. Reporting and concordance of method-
5. Moher D, Pain B, Jones A, et al. Does quality of reports of randomised trials affect es-
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 devel-
10. Romagnuolo J, Bardou M, Rahme E, et al. Magnetic resonance cholangiopancre-
11. Nederkoorn PJ, van der Graaf Y, Hunink MG. Duplex ultrasound and magnetic res-
onance angiography compared with digital subtraction angiography in carotid ar-
12. Whiting P, Rutjes AW, Dennes J, et al. A systematic review finds that diagnostic re-
14. Bossuyt PM, Reitsma JB, bruins DE, et al. The STARD statement for reporting stud-
16. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis:
17. van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-
19. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnos-
The marked variation in estimates of 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
17. 14 or 15 or 16
18. 10 not 17
19. 11 not 17
20. 12 not 17
21. 13 not 17

#### 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
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

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

<table>
<thead>
<tr>
<th>Meta-analysis</th>
<th>Diagnostic problem</th>
<th>Type of index test</th>
<th>No. of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balk et al&lt;sup&gt;20&lt;/sup&gt;</td>
<td>Emergency department diagnosis of acute myocardial infarction</td>
<td>Biomarker: creatine kinase (CK)-MB</td>
<td>9</td>
</tr>
<tr>
<td>Berger et al&lt;sup&gt;21&lt;/sup&gt;</td>
<td>Diagnosis of gallstones</td>
<td>Symptom: upper abdominal pain</td>
<td>12</td>
</tr>
<tr>
<td>Devillé et al&lt;sup&gt;22&lt;/sup&gt;</td>
<td>Workup of herniated discs in patients selected for surgery</td>
<td>Test of Lasegue</td>
<td>11</td>
</tr>
<tr>
<td>Fiellin et al&lt;sup&gt;23&lt;/sup&gt;</td>
<td>Screening for lifetime alcohol abuse or dependence in primary care settings</td>
<td>CAGE questionnaire</td>
<td>14</td>
</tr>
<tr>
<td>Gould et al&lt;sup&gt;24&lt;/sup&gt;</td>
<td>Workup of pulmonary nodules</td>
<td>Positron emission tomography with the glucose analog 18-fluorodeoxyglucose (FDG-PET)</td>
<td>29</td>
</tr>
<tr>
<td>Hobby et al&lt;sup&gt;25&lt;/sup&gt;</td>
<td>Diagnosis of complete tears of the triangular fibrocartilage complex in the wrist</td>
<td>MRI</td>
<td>11</td>
</tr>
<tr>
<td>Hoffman et al&lt;sup&gt;26&lt;/sup&gt;</td>
<td>Workup of prostate cancer in men with nonspecific elevations of prostate specific antigen levels</td>
<td>Free:total prostate-specific antigen ratio</td>
<td>21</td>
</tr>
<tr>
<td>Hoogendam et al&lt;sup&gt;27&lt;/sup&gt;</td>
<td>Primary care screening for prostate cancer</td>
<td>Digital rectal examination</td>
<td>13</td>
</tr>
<tr>
<td>Huicho et al&lt;sup&gt;28&lt;/sup&gt;</td>
<td>Screening for urinary tract infection in children</td>
<td>Urine marker: dipstick nitrate</td>
<td>18</td>
</tr>
<tr>
<td>Hurley&lt;sup&gt;29&lt;/sup&gt;</td>
<td>Diagnosis of gram-negative infection</td>
<td>Gelation Limulus amebocyte lysate</td>
<td>27</td>
</tr>
<tr>
<td>Kelly et al&lt;sup&gt;30&lt;/sup&gt;</td>
<td>Workup of staging of gastroesophageal cancer</td>
<td>Endoscopic ultrasonography</td>
<td>13</td>
</tr>
<tr>
<td>Kim et al&lt;sup&gt;31&lt;/sup&gt;</td>
<td>Diagnosis of coronary artery disease</td>
<td>Dobutamine echocardiography</td>
<td>40</td>
</tr>
<tr>
<td>Koelemay et al&lt;sup&gt;32&lt;/sup&gt;</td>
<td>Evaluation of lower-extremity arterial disease in aortoiliac tract</td>
<td>3-dimensional magnetic resonance angiography (MRA)</td>
<td>9</td>
</tr>
<tr>
<td>Kwok et al&lt;sup&gt;33&lt;/sup&gt;</td>
<td>Detection of coronary artery disease in women</td>
<td>Exercise electrocardiography</td>
<td>19</td>
</tr>
<tr>
<td>Lau et al&lt;sup&gt;34&lt;/sup&gt;</td>
<td>Emergency department diagnosis of acute myocardial infarction</td>
<td>Biomarker: CK-MB</td>
<td>10</td>
</tr>
<tr>
<td>Lederle et al&lt;sup&gt;35&lt;/sup&gt;</td>
<td>Screening for abdominal aortic aneurysm</td>
<td>Abdominal palpation</td>
<td>10</td>
</tr>
<tr>
<td>Li&lt;sup&gt;36&lt;/sup&gt;</td>
<td>Confirmation of endotracheal tube placement</td>
<td>Capnography: end-tidal CO₂ devices</td>
<td>10</td>
</tr>
<tr>
<td>Mitchell et al&lt;sup&gt;37&lt;/sup&gt;</td>
<td>Screening for squamous intraepithelial lesions of the cervix</td>
<td>Papanicolaou smear screening</td>
<td>17</td>
</tr>
<tr>
<td>Mol et al&lt;sup&gt;38&lt;/sup&gt;</td>
<td>Screening for Down’s syndrome</td>
<td>Ultrasonographic marker: nuchal translucency measurement</td>
<td>23</td>
</tr>
<tr>
<td>Nelemans et al&lt;sup&gt;39&lt;/sup&gt;</td>
<td>Evaluation of peripheral arterial disease</td>
<td>2-dimensional time-of-flight MRA</td>
<td>13</td>
</tr>
<tr>
<td>Safriel et al&lt;sup&gt;40&lt;/sup&gt;</td>
<td>Diagnosis of pulmonary emboli</td>
<td>CT pulmonary angiography</td>
<td>10</td>
</tr>
<tr>
<td>Sloan et al&lt;sup&gt;41&lt;/sup&gt;</td>
<td>Diagnosis of gonorrhea and chlamydial infection</td>
<td>Sign: abdominal/lower-abdominal pain</td>
<td>14</td>
</tr>
<tr>
<td>Smith Bindman et al&lt;sup&gt;42&lt;/sup&gt;</td>
<td>Screening for Down’s syndrome</td>
<td>Ultrasonographic marker: femoral shortening</td>
<td>28</td>
</tr>
<tr>
<td>Sonnad et al&lt;sup&gt;43&lt;/sup&gt;</td>
<td>Workup of staging of prostate cancer</td>
<td>MRI</td>
<td>21</td>
</tr>
<tr>
<td>Vasquez et al&lt;sup&gt;44&lt;/sup&gt;</td>
<td>Workup of acute cholecystitis</td>
<td>Morphine sulfate-augmented hepatobiliary imaging</td>
<td>15</td>
</tr>
<tr>
<td>Visser et al&lt;sup&gt;45&lt;/sup&gt;</td>
<td>Workup of peripheral arterial stenosis</td>
<td>Colour-guided duplex ultrasonography</td>
<td>17</td>
</tr>
<tr>
<td>Westwood et al&lt;sup&gt;46&lt;/sup&gt;</td>
<td>Selecting candidates with recently symptomatic carotid artery stenosis for surgery</td>
<td>3-dimensional contrast-enhanced MRA</td>
<td>7</td>
</tr>
<tr>
<td>Wiese et al&lt;sup&gt;47&lt;/sup&gt;</td>
<td>Diagnosis of vaginal trichomoniasis</td>
<td>Wet-mount smear technique</td>
<td>29</td>
</tr>
</tbody>
</table>
## Appendix 4: Effect of study characteristics on estimates of diagnostic accuracy from univariable analysis

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

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).