Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests

Jeroen G. Lijmer, MD
Ben Willem Mol, MD, PhD
Siem Heisterkamp, PhD
Gouke J. Bonsel, MD, PhD
Martin H. Prins, MD, PhD
Jan H. P. van der Meulen, MD, PhD
Patrick M. M. Bossuyt, PhD

During recent decades, the number of available diagnostic tests has been rapidly increasing. As for all new medical technologies, new diagnostic tests should be thoroughly evaluated prior to their introduction into daily practice. The number of test evaluations in the literature is increasing but the methodological quality of these studies is on average poor. A survey of the diagnostic literature (1990-1993) showed that only 18% of the studies satisfied 5 of the 7 methodological standards examined. Different guidelines have been written to help physicians with the critical appraisal of the diagnostic literature consisting of lists of criteria for the assessment of study quality. Criteria enable readers to check whether studies fulfill methodological criteria on study design, data collection, and methods of reporting the results.

As few diagnostic studies meet all of the methodological criteria, physicians and reviewers are faced with a difficult choice. Strict application of the methodological criteria would imply that only a small minority of the available data can be used in clinical practice. Alternatively, inclusion of a wider range of imperfect studies would require weighting of the evidence according to the relative importance of the criteria that such studies failed to satisfy. One article has reported such weights for methodological criteria. Unfortunately, these weights were established through a consensus procedure in a general internal medicine division at an academic medical center, rather than on empirical data. A data-driven approach had been previously used by Schulz et al when evaluating the influence of study design features on estimates of treatment effects in randomized controlled trials.

The purpose of our study was to assess empirically the impact of shortcomings in design, data collection, and reporting on the estimates of diagnostic accuracy. We compared estimates of diagnostic accuracy for a given test reported in studies with lower quality with estimates for the same test from other high-quality studies.
studies without these shortcomings. We hypothesized that estimates of diagnostic accuracy would be exaggerated in studies that failed to meet methodological standards.

METHODS

Data Sources and Data Extraction

An electronic search of the literature was performed to identify meta-analyses summarizing the accuracy of diagnostic tests. We focused on meta-analyses because they enabled us to identify a large number of studies on a single diagnostic problem. We concentrated on recent meta-analyses as we expected these meta-analyses to include both older studies, using suboptimal designs, as well as recent studies, applying a more up-to-date approach that lives up to current methodological standards. To be included, a meta-analysis had to be based on a systematic search of the literature, had to include at least 5 studies, and had to report sensitivities and specificities of included studies. The latter criterion was introduced to assure that from each reviewed study sensitivity and specificity were available and to allow for easy replication of our work.

The MEDLINE and EMBASE databases were searched (January 1996 to December 1997) using combinations of the words meta-analysis; diagnostic imaging; diagnostic tests; routine; sensitivity and specificity; and review, publication type. In addition, the Cochrane Library and the DARE database of the NHS Centre for Reviews and Dissemination were examined for relevant abstracts.

We retrieved 26 articles that included 5 or more studies. Fifteen articles had to be excluded: 7 were not based on a systematic literature search and 8 reported no list of sensitivities and specificities. A list of excluded articles is available from the authors.

For the 11 remaining articles, all original papers included in the analyses were retrieved (TABLE 1). The characteristics of these studies were extracted on a standard form by 1 of the authors (J.G.L.). The set of characteristics on the standard form was based on a synthesis from different lists of criteria for study quality. All studies were independently scored a second time by a second reviewer (B.W.M., P.M.M.B.). Disagreement was resolved by consensus, if necessary the judgment of a third reviewer was decisive.

Assessment of Study Quality

The optimal design for assessing the accuracy of a diagnostic test is considered to be a prospective blind comparison of the test and the reference test in a consecutive series of patients from a relevant clinical population. A relevant clinical population is a group of patients covering the spectrum of disease that is likely to be encountered in the current or future use of the test. There are several threats to the validity of a diagnostic study. Diagnostic accuracy can be overestimated if the test is evaluated in a group of patients already known to have the disease and a separate group of normal patients, rather than in a relevant clinical population. This will be referred to as a case-control study.

Selection bias can be present when not all patients presenting with the relevant condition are included in order of entry (consecutive) into the study, and when this selection is not random. If it was not clear from the text that a consecutive series of patients was included or a random subset, the corresponding study was scored as non-consecutive.

Veriﬁcation bias looms if the decision to perform the reference test is based on the result of the test under examination. In many diagnostic studies with an invasive reference test, most of the positive test results and only a small part of the negative test results are veriﬁed. Alternatively, negative test results are veriﬁed by a different, often less thorough, standard, for example follow-up. We will refer to these 2 forms of veriﬁcation bias as partial veriﬁcation bias and differential reference standard bias, respectively. In cases in which more than 10% of the study group was not subjected to the reference test, the study was scored as applying partial veriﬁcation; in cases in which different reference tests were used, the study was scored as differential reference standard. All other cases were scored as complete veriﬁcation.

Interpreting the reference test with knowledge of the results of the test under study can lead to an overestimation of a test’s accuracy, especially if the reference test is open to subjective interpretation. If the sequence of testing is reversed, it is important that the results of the test under study are interpreted without knowledge of the reference test. If it was not clear from the text that the interpretation of both tests was done while investigators were blinded the study was scored as not blinded.

In addition to characteristics of the study design, we also looked at methods of data collection and reporting. The data collection was categorized as either prospective or retrospective. In case of doubt, the method of data collection was scored as unknown. The reference test, the test under study, and the study population should be described with sufficient detail to allow for replication, validation, and generalization of the study. Descriptions of the tests were scored as sufﬁcient if clear deﬁnitions of positive and negative test results were mentioned in the text. Description of the study population was sufﬁcient if 2 of the following characteristics were described: age of participants, female to male ratio, or distribution of symptoms.

Statistical Analysis

The results of an individual study on diagnostic accuracy can be summarized in a 2 × 2 table. From this table, frequently used measures such as sensitivity, speciﬁcity, and predictive values can easily be calculated. Another measure for the diagnostic accuracy of a test is the diagnostic odds ratio (DOR), the odds for a positive test result in diseased persons relative to the odds of a positive result in nondiseased persons. The DOR is a single statistic of the results in a 2 × 2 table,
incorporating sensitivity as well as specificity. Expressed in terms of sensitivity and specificity the formula is:

\[
DOR = \frac{\text{Sensitivity}}{(1 - \text{Sensitivity})} \div \frac{\text{Specificity}}{(1 - \text{Sensitivity})}
\]

The effect of study characteristics was examined with a regression model that is adapted from the summary receiver operating characteristic curve model, developed for meta-analyses of diagnostic tests. The basic model contains the logarithm of the diagnostic odds ratio computed for a single study as a dependent variable and 2 explaining parameters, 1 for the intercept and 1 for the slope of the curve, for each meta-analysis. The intercept can be interpreted as the common DOR of the corresponding test and the parameter for the slope expresses variation of the DOR across individual studies due to threshold differences.

We added covariates to this model to examine whether, on average, studies that failed to meet the methodological criteria yielded different DORs. The resulting parameter estimates of the covariates can be interpreted after antilogarithm transformation as relative DORs (RDORs). They indicate the diagnostic performance of a test in studies failing to satisfy the methodological criterion, relative to its performance in studies with the corresponding feature. If the RDOR is larger than 1, studies not satisfying the criterion yield larger estimates of the DOR than studies with this corresponding feature.

In summary, the dependent variable of the model was the logarithm DOR. Explaining variables were 2 parameters for each meta-analysis (the common DOR and the threshold parameter) and 9 covariates to examine the effect of the different study characteristics, 1 for each feature. All study characteristics were evaluated simultaneously in a multivariate model.

A weighted linear regression analysis was used, with weights proportional to the reciprocal of the variance of the log DOR. This weighted linear regression assumes fixed effects. In case of zero entries, the DOR is not defined. This problem was solved by adding 0.5 to all cells of the 2 × 2 table for all studies in a meta-analysis. The model was fitted using maximum likelihood estimation, and programmed using statistical software (S-plus 4.5, Mathsoft Inc, Cambridge, Mass).

RESULTS

The subjects of the 11 included articles and the reviewed tests are sum-

<table>
<thead>
<tr>
<th>Source, y</th>
<th>Diagnostic Problem</th>
<th>Type of Test</th>
<th>Reference Test</th>
<th>No. of Studies</th>
<th>Literature Search</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Enzyme-linked immunosorbent assay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bonis et al,22 1997</td>
<td>Measuring improvement in hepatitis C</td>
<td>Alanine aminotransferase</td>
<td>Histology</td>
<td>12</td>
<td>1990-1995</td>
</tr>
<tr>
<td>Hallan and Asberg,23 1997</td>
<td>Suspicion of acute appendicitis</td>
<td>C-reactive protein</td>
<td>Histology and follow-up</td>
<td>20</td>
<td>1983-1997</td>
</tr>
<tr>
<td>Mol et al,26 1997</td>
<td>Diagnosis of tubal pathology in subfertility</td>
<td>Migration inhibitory factor</td>
<td>Laparoscopy and/or hysterosalpingography</td>
<td>14</td>
<td>1981-1994</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Computed tomographic scan</td>
<td>Histology or cytology</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Magnetic resonance imaging</td>
<td>Histology or cytology</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Siegman-Igra et al,28 1997</td>
<td>Diagnosis of catheter-related bloodstream infection</td>
<td>Semiquantitative catheter culture</td>
<td>Blood culture and clinical information</td>
<td>13</td>
<td>1977-1992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quantitative catheter culture</td>
<td>Blood culture and clinical information</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Qualitative blood culture</td>
<td>Blood culture and clinical information</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Double test</td>
<td>Cytogenetic diagnosis and follow-up</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>De Vries et al,21 1996</td>
<td>Evaluation of peripheral arterial disease</td>
<td>Duplex ultrasonography</td>
<td>Angiography</td>
<td>8</td>
<td>1985-1994</td>
</tr>
</tbody>
</table>
marized in Table 1. Two articles reviewed 3 tests, 3 articles reviewed 2 tests, and 6 articles reviewed 1 test. This resulted in a total of 18 separate meta-analyses for this analysis. These 18 meta-analyses summarized the results of 193 published studies. Nine studies could not be used in the final analysis because only abstracts were available for 4 and 2 × 2 table calculations were not possible for 5. Of the 184 studies remaining, some evaluated multiple diagnostic tests. A total of 218 diagnostic test evaluations were available for analysis.

The overall results of the quality assessment of the included studies are listed in Table 2. Most studies used a clinical cohort and described the cut-off that the test evaluated (98% and 89%, respectively). Only 15 (6.8%) of the 218 studies satisfied all 8 criteria used. Sixty-four (30%) of the 218 studies satisfied 6 or more criteria.

Table 2. Results of the Scoring of Study Quality (N = 218)

<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Score</th>
<th>No. of Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clinical population</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>Case-control</td>
<td>5</td>
</tr>
<tr>
<td>Verification</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complete</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>Different reference tests</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Partial</td>
<td>54</td>
</tr>
<tr>
<td>Interpretation of test results</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blinded</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Not blinded</td>
<td>148</td>
</tr>
<tr>
<td>Patient selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Consecutive</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>Nonconsecutive</td>
<td>121</td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prospective</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>Retrospective</td>
<td>28</td>
</tr>
<tr>
<td>Details test</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sufficient</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>Insufficient</td>
<td>23</td>
</tr>
<tr>
<td>Details reference test</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sufficient</td>
<td>138</td>
</tr>
<tr>
<td></td>
<td>Insufficient</td>
<td>80</td>
</tr>
<tr>
<td>Details population</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sufficient</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>Insufficient</td>
<td>86</td>
</tr>
</tbody>
</table>

The results from the regression analysis are presented in the FIGURE. Studies using a case-control design tended to overestimate the DOR 3-fold compared with studies with a clinical cohort (RDOR, 3.0). Studies using different reference tests for positive and negative test results had an RDOR of 2.2, showing approximately a 2-fold overestimation of the DOR compared with studies that used 1 reference test. Studies verifying only part of the population had on average the same DOR as studies that subjected all patients to the reference test (RDOR, 1.0). There were no studies verifying only part of the population and in conjunction using different reference standards. Interpretation of the reference test with knowledge of the outcomes of the test under study resulted in a RDOR of 1.3, causing an overestimation of the DOR by approximately 30% compared with studies with adequate blinding. Selective inclusion of patients into the study did not change the estimation of diagnostic accuracy significantly.

Retrospective data collection was not associated with an overestimation or underestimation of diagnostic accuracy in comparison with studies with prospective or unknown data collection. In a univariate analysis (data not shown) the RDOR of studies with unknown data collection was nearest to that of prospective studies, we therefore collapsed these 2 categories. The DORs in articles without a sufficient description of the test under study or the study population were, respectively, about 70% and 40% higher than estimates in articles reporting sufficient details. Studies reporting no details of the cut-off of the reference test had DORs that were approximately 30% smaller than studies reporting the details of the reference tests.

**COMMENT**

This study describes the quantitative effects of characteristics of study design on estimates of diagnostic accuracy. By collecting data from studies in published diagnostic meta-analyses, we were able to examine the effect of study characteristics on diagnostic accur
racy. Our analysis shows that studies of lower methodological quality, particularly those including nonrepresentative patients or applying different reference standards, tend to overestimate the diagnostic performance of a test.

The largest effect on the estimation of diagnostic accuracy was generated by studies using cases and controls, also labeled as spectrum bias.28 Often, mild cases that are difficult to diagnose are omitted from case-control studies, causing an overestimation of sensitivity as well as specificity. Another large effect was seen in studies that used different reference tests for the verification of positive and negative test results. The effect of this differential verification depends on the quality of the different reference tests used. Using a “gold” reference test for the positive test results and a poor reference test for the negative results can lead to an overestimation of both sensitivity and specificity of a test.28 For example, some studies evaluating the diagnostic performance of C-reactive protein (CRP) for the diagnosis of acute appendicitis used surgery and pathology as a reference test for patients with a high CRP. Patients with a low CRP were not operated on and clinical follow-up determined whether they were classified as having acute appendicitis. As low-grade infections with low CRPs can resolve spontaneously, this verification strategy fails to identify all false-negative test results. This way the diagnostic performance of CRP will be overestimated. If the poor reference standard fails to identify true-negative test results, the use of different reference standards can lead to an underestimation of the diagnostic performance of a test.

The terms verification bias or workup bias are sometimes used when not all patients are subjected to the reference test.29,30 We prefer the term partial verification bias to differentiate this situation from the situation in which different reference standards were used for verification. In theory, verifying more positive test results than negative test results will lead to an overestimation of sensitivity and an underestimation of specificity, resulting in a either an increase or a decrease of the DOR.20 In the analysis reported here, partial verification resulted in DORs comparable with those from studies with complete verification. The absence of an association with estimation of diagnostic accuracy could be caused by the definition we used. Partial verification will only lead to bias if systematically more abnormal than normal (or more normal than abnormal) test results are subjected to the reference standard. In many studies it was not clear why some patients were not subjected to the reference standard and if this was related to the test under study. Therefore, we scored studies as partial verification when 10% of the patients were not verified, including studies in which patients were not verified due to a random error. This could weaken the possible effect of partial verification in our analysis. Some other studies also have shown no effect of partial verification on the overall diagnostic accuracy of a test. In these studies only a shift of threshold values along the receiver operating characteristic curve was observed.31,32

The average effect of inappropriate blinding was small. The studies included in this analysis used many different reference standards ranging from diagnostic imaging to histology. In case the reference standard is objective, no effect is to be expected. Thus, in clinical situations with a subjective reference standard, the effect of not blinding could be larger.

Given the distinction made between case-control and cohort studies, we did not observe an influence of nonconsecutive sampling of patients. Retrospective collection did not generate different results than studies with prospective data collection (when corrected for all other methodological flaws).

When looking at the criteria for the methods of reporting, we found a sufficient description of the test and a sufficient description of the population associated with an overestimation of diagnostic accuracy in cases of their absence. As these criteria are not directly related to the study design, it is unclear how they lead to an overestimation of diagnostic accuracy. Somehow they seem to be predictors of methodological flaws in studies. In contrast with these findings, studies with an insufficient description of the reference standard generated less optimistic results compared with studies with an adequate description. An explanation could be that these studies possibly had a large variation in the interpretation of the reference standard. Large interobserver variation is associated with a poor diagnostic accuracy, leading to an underclassification of diseased persons.39 However, the same argument could be used for lack of description of the test under study. Many times we had difficulties in deciding whether the reference test was described with enough detail. How much detail is needed if the reference standard is histology? The definition was largely dependent on the clinical situation under study. The scores of 20 studies changed after the consensus reading in comparison with the first reading.

Four meta-analyses have examined the quantitative effect of study characteristics on diagnostic performance. These analyses were published before 1996 and, hence, not included in our sample. All were limited to a single test.34,37 Some of these studies found that partial verification and absence of blinding affected the estimates of diagnostic accuracy,35,34,36 while others found no effect of these characteristics.35 As in the first 3 meta-analyses, we found an overestimation of diagnostic accuracy in studies without appropriate blinding, as in the latter we found no effect of partial verification. One of these meta-analyses also looked at the reporting method of the test and the reference test in combination with other criteria and also found that insufficient description was associated with overestimation of the diagnostic accuracy.37

The decision to limit our analysis to data from recent meta-analyses could have affected our results. Extending our sample to older meta-analyses would most likely change the relative frequency of the study characteristics, but not necessarily the relative size of their effects on diagnostic accuracy.
Publication bias also has to be taken into account since only diagnostic studies published in scientific journals were included in the analysis. One can speculate that studies have a higher likelihood of being published when they are either of good quality or when they show encouraging results. Such a selective publication policy could lead to an inflation of the associations we found. It is difficult to examine the effect of publication bias since there is no registration of unpublished diagnostic studies. For future research in this field and for reviewers of diagnostic tests, such a central registration of diagnostic research protocols would be useful.

When reading the results of a single diagnostic study, it is difficult to weigh the methodological flaws against the available evidence. How large is the possible overestimation and will it have clinical consequences? In a study with different reference standards, without blinding and lacking a description of the test, the DOR would on average be overestimated by 5-fold based on the results of our analysis. This is equal to reporting a sensitivity and specificity of about 84% when in fact both should be 70%. Differences will be smaller if sensitivity and specificity are higher and if only a few minor criteria are not fulfilled.

Our results stress the importance of adequate methodology and the need for complete and reliable reporting of research. Assessment of quality is only feasible in the light of complete clarity on the methodology. Authors should therefore describe explicitly their methods of patient selection, methods of disease verification, and criteria for interpretation of the test and the reference test.

This study shows that shortcomings in design, data collection, and reporting affect estimates of diagnostic accuracy. Investigators should be aware of this when designing their studies and readers should be aware of this when interpreting the results. Our results can be of help in determining the merits of the available evidence when appraising literature. Greater editorial vigilance could help make researchers aware of current methodological standards and thereby decrease the potential for bias in future diagnostic studies.

Acknowledgment: The authors gratefully acknowledge lain Chalmers, MSc, for his review of and comments on the article.

REFERENCES

RESEARCH LETTER

NIH Research Grants: Funding and Re-funding

To the Editor: National Institutes of Health (NIH) grants play an important role in the careers of the research faculty in medical schools. Given the importance of NIH funding in an academic career, we sought to determine several aspects of NIH funding that have, to the best of our knowledge, not been examined in detail. These are (1) the median length of time of NIH funding during the academic life of an individual, (2) the predictive value, for long-term funding, of receiving NIH funding at any given time, and (3) the effect of length of the unfunded period on the likelihood that an investigator will be subsequently funded.

Methods. We obtained the funding histories of individuals who were awarded any NIH grants in the index years 1972 and 1982, and tracked individual funding histories for 25 years and 15 years, respectively. Data consisted of the number of grants each individual received annually following the index year. A total of 1707 individuals received awards in 1972 and 1639 in 1982. Because of the nature of record keeping at the NIH, the precise length of each award was not available. Based on the historical experience of NIH funding, we assumed that the average length of an NIH award is 4 years.

Results. The mean number of awards obtained by individuals throughout their careers was 2.5 for the 1972 and 3.3 for the 1982 cohort; the median for both groups was 2.0. About 40% of individuals who obtain an NIH grant never received another for the rest of their careers.

The best predictor of future funding appears to be the number of grants garnered during a 10-year period. Thus, for both the 1972 and 1982 cohorts, we computed the proportion of individuals obtaining 1, 2, or more than 2 awards over a 10-year period from 1972 to 1981 or 1982 to 1991, respectively. The distribution of the individuals into these 3 categories was similar in the 2 cohorts: 39.0% vs 35.0%, 18.3% vs 22.0%, and 42.4% vs 42.0%, respectively. For individuals who only obtained 1 grant for the 10-year period (1972-1981 or 1982-1991), 92.5% and 93%, respectively, did not obtain any funding for the subsequent period until 1997. Of those who received 1 additional grant (or a total of 2 grants for the 10-year period), 72% did not obtain any subsequent funding until 1997. These 2 cohorts were in contrast with those who obtained more than 2 grants for the 10-year interval. Only 27% of these individuals failed to obtain a grant for the subsequent period.

Comment. The cumulative period of NIH funding appears to be rather brief for most individuals. Neither the possession of an NIH grant nor the apparent ability to acquire one based on prior research training or history predicts whether an individual will receive future grants.

T. V. Rajan, MD, PhD
Jonathan Clive, PhD
University of Connecticut Health Center
Farmington

CORRECTIONS

Incorrect Equation: In the Original Contribution entitled “Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests” published in the September 15, 1999, issue of THE JOURNAL (1999;282:1061-1066), the equation was incorrect. On page 1063, the equation should have appeared:

\[
DOR = \frac{Sensitivity}{1 - Sensitivity} \cdot \frac{1 - Specificity}{Specificity}
\]

Incorrect Color Reproduction: In the Consensus Statement entitled “Anthrax as a Biological Weapon: Medical and Public Health Management” published in the May 12, 1999, issue of THE JOURNAL (1999;281:1735-1745), the color of the photomicrograph in Figure 1 on page 1727 was incorrectly reproduced. The correct image showing gram-positive anthrax bacilli in a peripheral blood smear from a rhesus monkey that died of inhalational anthrax is shown below, left. A second photomicrograph (lower magnification) of a gram stain of the peripheral blood of a rhesus monkey that died of inhalational anthrax is also shown below, right.

Figure. Gram Stains of Bacillus anthracis