Systematic reviews of diagnostic test accuracy
Leeflang, M.M.G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Impact of adjustment for quality on results of meta-analyses of diagnostic accuracy

Mariska M.G. Leeflang, Johannes B. Reitsma, Rob J.P.M. Scholten, Anne W.S. Rutjes, Marcello Di Nisio, Jon J. Deeks, Patrick M.M. Bossuyt

Clin Chem. 2007;53(2):164-72
Abstract

Background: We examined whether and to what extent different strategies of defining and incorporating quality of included studies affect the results of meta-analyses of diagnostic accuracy.

Methods: We evaluated the methodological quality of 487 diagnostic accuracy studies in 30 systematic reviews with the QUADAS (Quality Assessment of Diagnostic accuracy Studies) checklist. We applied three strategies that varied both in the definition of quality and in the statistical approach to incorporate the quality-assessment results into meta-analyses. We compared magnitudes of diagnostic odds ratios, widths of their confidence intervals, and changes in a hypothetical clinical decision between strategies.

Results: Following two definitions of quality, we concluded that only 70 or 72 of 487 studies were of “high quality.” This small number was partly due to poor reporting of quality items. None of the strategies for accounting for differences in quality led systematically to accuracy estimates that were less optimistic than ignoring quality in meta-analyses. Limiting the review to high-quality studies considerably reduced the number of studies in all reviews, with wider confidence intervals as a result. In 18 reviews, the quality adjustment would have resulted in a different decision about the usefulness of the test.

Conclusions: Although reporting the results of quality assessment of individual studies is necessary in systematic reviews, reader wariness is warranted regarding claims that differences in methodological quality have been accounted for. Obstacles for adjusting for quality in meta-analyses are poor reporting of design features and patient characteristics and the relatively low number of studies in most diagnostic reviews.
3.1 Introduction

Health care professionals seeking the best information about diagnostic tests increasingly turn to systematic reviews of test-accuracy studies, yet a review’s summary estimate can be biased if the studies in the review are flawed. An evaluation of the quality of the original studies, therefore, is an essential issue of any systematic review.

The methodological quality of studies can be defined in terms of their susceptibility to bias. Studies with methodological shortcomings, such as inclusion of healthy control individuals or selective use of multiple reference standards to verify index test results, have produced different measures of test accuracy. In most cases, such deficiencies have been associated with inflated estimates of diagnostic accuracy. The inclusion of lower-quality studies in a meta-analysis may therefore produce unrealistically high accuracy estimates. Accounting for quality differences can be expected to produce less optimistic summary estimates of diagnostic accuracy.

Design feature variability and the presence of studies with suboptimal designs in a systematic review may also increase heterogeneity in results among studies. Given these considerations, one can expect strategies that account for quality in meta-analyses of diagnostic accuracy to lead to more homogeneous results and therefore to more precise estimates, with narrower confidence intervals around the accuracy measures of interest, than estimates without quality adjustment.

Quality assessment of individual studies in a review may identify both design deficiencies that can lead to bias and sources of variation that can lead to heterogeneity. Several quality-assessment tools, most of which use a “checklist” approach, have been developed for diagnostic accuracy studies. A recently developed generic quality-assessment tool based on a modified Delphi procedure has been recommended by the Cochrane Collaboration as a starting point for quality assessment in diagnostic reviews.

Although quality appraisal has been recognized as an essential step of systematic reviews, how study quality should be addressed in meta-analyses of diagnostic accuracy studies is less clear. Strategies to incorporate study quality into meta-analyses can be broadly divided into 3 categories: including all studies, irrespective of quality; analyzing subgroups that differ in quality; and multivariable regression analysis. The slightly different recommendations given in the guiding reports are all based on sparse evidence.

To test the hypothesis that adjustment for quality produces less optimistic estimates of diagnostic accuracy and narrower confidence intervals, we compared 3 different strategies for incorporating quality in analyzing a number of previously published systematic reviews of diagnostic accuracy studies.
3.2 Methods

We studied 3 alternative strategies for quality incorporation in meta-analysis and comparing the results of analyzing all available studies irrespective of their quality, in a series of systematic reviews of diagnostic accuracy studies. Within each systematic review, we compared the summary diagnostic odds ratios (DORs) and the widths of the confidence intervals across these strategies.

3.2.1 Study set

To include a broad sample of diagnostic studies that examined a variety of tests over time, we conducted a systematic electronic search for systematic reviews of diagnostic accuracy studies published between January 1999 and April 2002. This search produced a set of 28 reports of systematic reviews. Details of the search strategy are available from the authors. Inclusion criteria were (1) a systematic review of diagnostic test-accuracy studies, (2) inclusion of at least 10 original studies, (3) no exclusion of primary studies based on design features, and (4) the ability to reproduce the 2 by 2 tables from the original studies. The 28 reports yielded 30 systematic reviews. Details of the inclusion process are reported elsewhere.

A variety of conditions and index tests were studied in these 30 reviews (Table 3.1). The median number of studies in a review was 14 (interquartile range, 10 to 20). The median sample size of the individual studies was 100 (interquartile range, 43 to 288).

3.2.2 Assessment of methodological quality

We assessed the methodological quality of all 487 studies included in the 30 reviews with items from the QUADAS instrument (Table 3.2). We limited ourselves to the 7 QUADAS items most closely related to methodological quality and did not use the items that referred to quality of reporting. We dichotomized each item by scoring as deficient any study feature that was not reported.

QUADAS item 1 (Table 3.2) refers to both the generalizability of results and the possibility that the study may produce biased results. We assessed 3 patient-spectrum components that refer to the distorted selection of participants, because previous studies have linked these components to biased accuracy estimates. These components were consecutive enrollment of patients, case-control or 2-gate design vs. cohort design, and avoidance of limited challenge. Limited challenge was defined as the exclusion of patients with disease characteristics that may produce false-positive or false-negative results (e.g., exclusion of patients with existing lung disorders in an accuracy study of spiral computed tomography for the diagnosis of pulmonary embolism). A 2-gate study was defined as a case-control study in which cases and controls are sampled from 2 distinct source populations by means of different selection criteria.
Two independent assessors conducted quality assessments, and consensus meetings resolved disagreements. If necessary, a third person made the final decision.

Table 3.1. Characteristics of the systematic reviews in our study set.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Target condition</th>
<th>Index test(s)</th>
<th>No. of included studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balk et al., 2001¹</td>
<td>Acute myocardial infarction</td>
<td>Laboratory test</td>
<td>9</td>
</tr>
<tr>
<td>Berger et al., 2000¹⁶</td>
<td>Gallstones</td>
<td>Physical examination</td>
<td>12</td>
</tr>
<tr>
<td>Devillé et al., 2000¹⁷</td>
<td>Herniated discs</td>
<td>Physical examination</td>
<td>11</td>
</tr>
<tr>
<td>Fiellin et al., 2000¹⁴</td>
<td>Alcohol abuse</td>
<td>Questionnaires</td>
<td>14</td>
</tr>
<tr>
<td>Gould et al., 2001¹⁹</td>
<td>Pulmonary nodules</td>
<td>FDG-PET¹</td>
<td>29</td>
</tr>
<tr>
<td>Hobby et al., 2000²⁰</td>
<td>Tears in wrist cartilage</td>
<td>MRI</td>
<td>11</td>
</tr>
<tr>
<td>Hoffman et al., 2000²¹</td>
<td>Prostate cancer</td>
<td>Laboratory test</td>
<td>21</td>
</tr>
<tr>
<td>Hoogendam et al., 1999²²</td>
<td>Prostate cancer</td>
<td>Physical examination</td>
<td>13</td>
</tr>
<tr>
<td>Huicho et al., 2002²³</td>
<td>Urinary tract infection</td>
<td>Laboratory test</td>
<td>18</td>
</tr>
<tr>
<td>Hurley, 2000²⁴</td>
<td>Gram-negative infections</td>
<td>Laboratory test</td>
<td>27</td>
</tr>
<tr>
<td>Kelly et al., 2001²⁵</td>
<td>Gastroesophageal carcinoma</td>
<td>Ultrasound</td>
<td>13</td>
</tr>
<tr>
<td>Kim et al., 2001²⁶</td>
<td>Coronary artery disease</td>
<td>Echocardiography</td>
<td>40</td>
</tr>
<tr>
<td>Koelemay et al., 2001²⁷</td>
<td>Peripheral arterial disease</td>
<td>MRA</td>
<td>9</td>
</tr>
<tr>
<td>Kwok et al., 1999²⁸</td>
<td>Coronary artery disease</td>
<td>Echocardiography</td>
<td>19</td>
</tr>
<tr>
<td>Lau et al., 2001²⁹</td>
<td>Acute myocardial infarction</td>
<td>Laboratory test</td>
<td>10</td>
</tr>
<tr>
<td>Lederle et al., 1999²⁰</td>
<td>Abdominal aortic aneurysm</td>
<td>Physical examination</td>
<td>10</td>
</tr>
<tr>
<td>Li, 2001³¹</td>
<td>Endotracheal tube placement</td>
<td>Capnography</td>
<td>10</td>
</tr>
<tr>
<td>Mitchell et al., 1999²²</td>
<td>Cervix lesions</td>
<td>Cytology</td>
<td>17</td>
</tr>
<tr>
<td>Mol et al., 1999³³</td>
<td>Down syndrome</td>
<td>Ultrasound</td>
<td>23</td>
</tr>
<tr>
<td>Nelemans et al., 2000³⁴</td>
<td>Peripheral arterial disease</td>
<td>MRA</td>
<td>13</td>
</tr>
<tr>
<td>Safriel et al., 2002³⁵</td>
<td>Pulmonary emboli</td>
<td>CT</td>
<td>10</td>
</tr>
<tr>
<td>Sloan et al., 2000³⁶</td>
<td>Gonorrhea and chlamydial infection</td>
<td>Physical examination</td>
<td>14</td>
</tr>
<tr>
<td>Smith-Bindman et al., 2001³⁷</td>
<td>Down syndrome</td>
<td>Ultrasound</td>
<td>28</td>
</tr>
<tr>
<td>Sonnad et al., 2001³⁸</td>
<td>Prostate cancer</td>
<td>MRI</td>
<td>21</td>
</tr>
<tr>
<td>Vasquez et al., 2000³⁹</td>
<td>Acute cholecystitis</td>
<td>Scintigraphy</td>
<td>15</td>
</tr>
<tr>
<td>Visser et al., 2000⁴⁰</td>
<td>Peripheral arterial stenosis</td>
<td>Ultrasound</td>
<td>17</td>
</tr>
<tr>
<td>Westwood et al., 2002⁴¹</td>
<td>Carotid stenosis</td>
<td>MRA</td>
<td>24</td>
</tr>
<tr>
<td>Wiese et al., 2000⁴²</td>
<td>Vaginal trichomoniasis</td>
<td>Cytology</td>
<td>29</td>
</tr>
</tbody>
</table>

¹ FDG-PET, [¹⁸F]fluorodeoxyglucose positron emission tomography; MRI, magnetic resonance imaging; MRA, magnetic resonance angiography; CT, computed tomography.
Chapter 3

3.2.3 Meta-analysis

We used the summary ROC model of Moses and Littenberg for our meta-analysis. Their model uses linear regression analysis to examine how D, the natural logarithm of the DOR, changes as a function of S, which is the sum of logit(sensitivity) and logit(1 – specificity). S is related to the threshold for classifying a test as positive.

We modelled the intercept and slope of the model as fixed effects but included a random effect to allow for variation beyond chance among studies. We weighted studies by the inverse of the variance of the log DOR to allow for the precision with which each study measured the log DOR. This procedure gave more weight to larger studies.

In the multivariable quality-adjustment strategies, covariates representing quality items were added to the model; this step allowed the intercept and slope in the regression analysis to differ between subgroups of studies defined by the correspond-

<table>
<thead>
<tr>
<th>Table 3.2. QUADAS items included in the 2 definitions of “high quality.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence-based definition</td>
</tr>
<tr>
<td>1. Was the spectrum of patients representative of the patients who will receive the test in practice?</td>
</tr>
<tr>
<td>2. Were selection criteria clearly described?</td>
</tr>
<tr>
<td>3. Is the reference standard likely to correctly classify the target condition?</td>
</tr>
<tr>
<td>4. Is the time period between reference standard and index test short enough?</td>
</tr>
<tr>
<td>5. Did the whole sample receive verification using a reference standard for diagnosis?</td>
</tr>
<tr>
<td>6. Did patients receive the same reference standard regardless of the index test results?</td>
</tr>
<tr>
<td>7. Was the reference standard independent from the index test?</td>
</tr>
<tr>
<td>8. Was the execution of the index test described in sufficient detail to permit replication of the test?</td>
</tr>
<tr>
<td>9. Was the execution of the reference standard described in sufficient detail to permit replication of the test?</td>
</tr>
<tr>
<td>10. Were the index test results interpreted without knowledge of the results of the reference standard?</td>
</tr>
<tr>
<td>11. Were the reference standard results interpreted without knowledge of the results of the index test?</td>
</tr>
<tr>
<td>12. Were the same clinical data available when test results were interpreted as would be available in practice?</td>
</tr>
<tr>
<td>13. Were uninterpretable / intermediate results reported?</td>
</tr>
<tr>
<td>14. Were withdrawals from the study explained?</td>
</tr>
</tbody>
</table>

3.2.3 Meta-analysis

We used the summary ROC model of Moses and Littenberg for our meta-analysis. Their model uses linear regression analysis to examine how D, the natural logarithm of the DOR, changes as a function of S, which is the sum of logit(sensitivity) and logit(1 – specificity). S is related to the threshold for classifying a test as positive.

We modelled the intercept and slope of the model as fixed effects but included a random effect to allow for variation beyond chance among studies. We weighted studies by the inverse of the variance of the log DOR to allow for the precision with which each study measured the log DOR. This procedure gave more weight to larger studies.

In the multivariable quality-adjustment strategies, covariates representing quality items were added to the model; this step allowed the intercept and slope in the regression analysis to differ between subgroups of studies defined by the correspond-
Quality adjustment in meta-analyses of diagnostic accuracy

ing covariate. In all strategies, we estimated the summary DOR over all studies the meta-analysis at the mean S value of these studies. Because the DOR cannot be calculated in 2 by 2 tables containing a zero, we added 0.5 to all 4 cells in these situations as a continuity correction\textsuperscript{44,48}.

3.2.4 Strategies for incorporating quality

We compared the following 3 statistical approaches to account for quality in meta-analyses: (1) The “restrict” strategy applied to meta-analysis of high-quality studies only. Studies were regarded as “high quality” when they fulfilled all quality criteria. (2) The “adjust all” strategy involved multivariable adjustment for all individual quality items by including all these items in a single multivariable model, irrespective of the strength of the association between these items and the DOR. (3) The “selective adjustment” strategy consisted of multivariable adjustment for only those quality items that were significantly associated with the DOR in a univariable analysis (P for entry <0.2)\textsuperscript{49,50}.

These strategies were compared with a reference strategy in which all studies within the original meta-analysis were included, irrespective of their quality characteristics.

Differences in results between strategies may depend both on the definition of quality and on the statistical approach used. We therefore considered 2 different sets of quality items to define higher-quality studies. The first set was chosen because there is empirical evidence that they can lead to biased results\textsuperscript{4,5}. This set, referred to as the “evidence-based” quality definition, includes QUADAS items 5, 6, 10, and 11 (Table 3.2). The second set of quality items (QUADAS items 1, 5, and 6) is referred to as the “common practice” quality definition and was selected because these 3 items are often applied in diagnostic reviews\textsuperscript{5,11}. The restrict strategy and the adjust-all strategy were applied twice, once with the evidence-based definition of quality and once with the common-practice definition.

3.2.5 Comparisons and analysis

We compared the summary DOR and its 95% confidence interval for the reference strategy, which included all studies, with the 3 quality-adjusting strategies in all 30 systematic reviews. Differences in results between strategies were analyzed within each systematic review with the Wilcoxon signed rank test to determine whether a strategy consistently led to higher or lower estimates of diagnostic accuracy. To investigate whether the strategies that adjusted for quality also resulted in more precise summary DOR estimates, we again used the Wilcoxon signed rank test statistic to compare the different approaches with respect to the absolute widths of the natural logarithm of the 95% confidence interval around the mean DOR.

To determine whether the change in summary DOR would affect clinical decisions, we used 4 arbitrary categories, which were defined by the absolute size of the summary DOR. If a meta-analysis resulted in a point estimate of the DOR <16, the test
was regarded as not useful. We regarded a test with a DOR of 16 to 81 as moderately useful, a test with a DOR of 81 to 361 as useful, and a test with a DOR >361 as very useful. The DOR values of 16, 81, and 361 correspond to sensitivity-specificity pairs of 80%-80%, 90%-90%, and 95%-95%, respectively.

We used SAS for Windows, version 9.1.3 (SAS Institute) for all analyses and the PROC MIXED procedure in SAS to fit all models.

### 3.3 Results

Figure 3.1 summarizes how often the 7 QUADAS items were fulfilled in the 487 studies. Nonreporting of items was common, particularly for blinding of the index test (49%) and the reference test (72%), adequate time interval between the index and reference standard (42%), and whether patients were consecutively included (34%).

Nine of the 30 reviews included studies of the case-control or 2-gate type. Whether all patients had received the reference standard and whether the reference standard was the same for each patient were well reported (99% of the studies). In 3 reviews, the primary studies used different reference standards to verify index test results.

Applying the evidence-based definition of quality (items 5, 6, 10, and 11 of the QUADAS checklist) identified 72 (15%) of the 487 primary studies as high quality. With this definition, 12 of the 30 systematic reviews had no high-quality studies, and 9 reviews included at least 3 high-quality studies.

Applying the common-practice definition identified 70 high-quality studies (14%). With this definition, 9 systematic reviews contained no high-quality studies, and 11 reviews had at least 3 high-quality studies. Use of both definitions yielded only 3 reviews that contained 3 high-quality studies.

![Figure 3.1. Overall results of quality assessment of the various QUADAS items in the 487 primary studies. Items 1a, 1b, and 1c refer to the different components of patient spectrum as we extracted them.](image-url)
Figure 3.2. Point estimates of the DOR and confidence intervals of all analyses. The abscissa represents the DOR, and the ordinate lists each meta-analysis by the first author, with the number of included studies in parentheses. Dotted lines reflect a DOR of 16 (i.e., a test with 80% sensitivity and 80% specificity), a DOR of 81 (90% sensitivity and 90% specificity), and DOR of 391 (95% sensitivity and 95% specificity). Analyses are indicated as follows: not incorporating quality (■), evidence-based restricted (▲), common-practice restricted (□), evidence-based multivariable (▲), common-practice multivariable (Δ), and selective adjustment (x).
Chapter 3

3.3.1 Comparing the pooled estimates of the various strategies

The summary DORs and the corresponding 95% confidence intervals were obtained for all 30 systematic reviews with the reference and 3 quality-adjustment strategies (Figure 3.2).

The evidence-based restrict strategy, which analyzed only high-quality studies according to the evidence-based definition, could be applied in 9 reviews containing 3 high-quality studies. In 3 cases, the DOR for the high-quality studies was higher than the DOR obtained by ignoring quality and including all studies, whereas the opposite occurred in 5 cases \( (P = 0.64) \). In 1 review, the DOR did not change, because all studies were high-quality studies according to the evidence-based definition. We found only 2 or fewer high-quality studies in the other reviews, and we did not calculate a summary estimate based on these small numbers.

The restrict strategy with the common-practice definition could be used in 11 reviews. This restrict strategy produced a higher DOR in 4 meta-analyses and a lower estimate in 7 others. The mean odds ratio was not significantly higher or lower when quality was not incorporated, compared with the different restrictive strategies (Table 3.3).

When we included all the items of the evidence-based quality definition as covariates in the multivariable model, model building failed in 9 reviews. In these reviews, at least 1 of the quality criteria was not fulfilled by any of the included studies. In 9 of the other 21 reviews, the adjust-all strategy resulted in a DOR estimate that was higher than when quality was not incorporated; 11 times the estimate was lower.

Table 3.3. Comparison of DORs and 95% confidence interval widths of different quality-incorporating strategies.

<table>
<thead>
<tr>
<th>Alternative strategy (no. of analyses)</th>
<th>Higher</th>
<th>Lower</th>
<th>Equal</th>
<th>P</th>
<th>Broader</th>
<th>Smaller</th>
<th>Equal</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence-based restricted (9)</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0.64</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0.078</td>
</tr>
<tr>
<td>Common-practice restricted (11)</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0.52</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
</tr>
<tr>
<td>Evidence-based multivariable (21)</td>
<td>9</td>
<td>11</td>
<td>1</td>
<td>0.31</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Common-practice multivariable (23)</td>
<td>10</td>
<td>13</td>
<td>0</td>
<td>0.68</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Selective adjustment (30)</td>
<td>5</td>
<td>7</td>
<td>18</td>
<td>0.85</td>
<td>12</td>
<td>0</td>
<td>18</td>
<td>0.001</td>
</tr>
</tbody>
</table>

*All strategies were compared with the overall meta-analysis, in which all studies within the original meta-analysis were included irrespective of their quality characteristics.

H₀: DOR<sub>overall</sub> = DOR<sub>alternative</sub>

H₀: CI<sub>overall</sub> = CI<sub>alternative</sub>

b H₀, null hypothesis; CI, confidence interval.
In 1 review, all of the original studies could be regarded as of high quality, so there was no change in the summary DOR.

With the common-practice definition, we were able to make a multivariable adjust-all model in 23 reviews. The estimated DOR was higher in 10 reviews and lower in 13. The differences between analyzing studies irrespective of their quality and analyses with the 2 multivariable strategies were not significant (Table 3.3).

The selective-adjustment strategy included only items that were significantly associated with accuracy in a univariable analysis \((P < 0.2)\). In 18 reviews, none of the QUADAS items was significantly associated with accuracy, and the use of all original studies in a meta-analysis yielded the same summary DOR as when quality was disregarded. In 5 reviews, only one single QUADAS item had a significant effect, and in a further 5, 1, and 1 meta-analyses respectively 2, 3, and 4 items were significant. The selective-adjustment strategy led to a higher estimate in 5 cases and to a lower estimate in 7 cases, compared with the meta-analysis in which quality was not incorporated.

Figure 3.3 shows the relative DORs (compared with not including quality in the analysis) for the various quality-adjustment strategies. The symmetrical distribution around unity illustrates that there is no systematic trend in underestimating or overestimating the DOR of a test. However, in 5 cases, the alternative strategy resulted in a DOR >5 times higher than when quality was disregarded; in 3 cases the relative DOR was <0.2.
None of the quality-adjustment strategies produced systematically narrower confidence intervals for the summary DOR than analyzing studies irrespective of their quality (Table 3.3). The confidence intervals were significantly wider with the restrict and adjust-all strategies ($P < 0.01$) but did not significantly differ with the selective-adjustment method ($P = 0.08$).

Because differences between strategies can be due to both differences in quality definitions and differences in statistical methods, we compared the results between statistical methods within 1 definition. We also compared the results with 2 quality definitions within 1 strategy. We observed no systematic differences between the 2 approaches, either for the summary estimates or for their 95% confidence intervals.

The judgment about the usefulness of a test based on the magnitude of the summary DOR was not affected in 12 of the 30 reviews with any of the quality-adjustment strategies (Figure 3.2). In 18 reviews, the quality-adjusted DOR obtained with 1 or more of the quality-adjustment strategies ended in a different category than the DOR obtained with all studies included. The DOR was higher in 14 cases and lower in 17 others (Figure 3.2).

### 3.4 Discussion

In this re-analysis of 30 previously published systematic reviews, we found no evidence for our hypothesis that adjustment for differences in methodological quality in meta-analysis leads to less optimistic summary diagnostic accuracy estimates with less variability in results among better-quality studies. We saw no such overall effects for strategies that relied on restriction to high-quality subsets, on multivariable adjustment for a set of quality items, or on selective multivariable adjustment for significant quality items.

A main problem that authors of systematic reviews encounter is poor reporting of study characteristics, and our study was no exception. We scored any study feature that was not reported as deficient. Dichotomizing QUADAS items into a simple “yes” or “no” can lead to loss of information, especially when many study characteristics are unreported. Some QUADAS items, such as the use of an adequate reference standard and the generalizability of the patient spectrum, could not be assessed at all in our data set. Both of these items can have a large effect on the performance of a test under study, and a proper incorporation of these characteristics could have resulted in a larger effect of the quality-adjustment strategies.

Because our analysis unit was the single meta-analysis, our sample size was only 30. Therefore, the power for detecting significant trends between strategies was limited, despite the inclusion of 487 individual studies. The 30 systematic reviews
covered a wide range of clinical topics and diagnostic tests, with a wide variability in the magnitude of the DOR. Our primary outcome variable was the DOR, which is a single accuracy indicator that incorporates both the sensitivity and specificity of a test. Such a single indicator is convenient in the analysis, but it also means that any given summary DOR can be produced by innumerable sensitivity-specificity combinations. In practice, the value of 1 accuracy measure, say sensitivity, may be more critical than another if the implications of false-positive and false-negative test results differ in severity.

In our analysis, we refrained from calculating summary quality scores for studies and labelling any study that exceeded a certain threshold score as high quality. Such summary quality scores have been extensively studied—and criticized—in systematic reviews of intervention studies. Different shortcomings in study design may cause different forms of bias, making it almost impossible to determine the weight that should be given to each quality item in calculating such quality scores. We also did not include a sequential analysis of the studies based on their quality ranking, which would have led to a quality-adjusted cumulative meta-analysis. This strategy also requires a hierarchical approach to study quality in that it assumes that some criteria are more important than others and that studies fulfilling more criteria are of higher quality.

Several previous studies have linked design features of diagnostic accuracy studies to changes in accuracy estimates. One systematic review documented the theoretical and empirical evidence for several sources of bias. Two publications, which examined these effects in a collection of systematic reviews, both reported significant effects for a number of features across meta-analyses. We can only speculate why we failed to find any systematic differences from incorporating these study features in the meta-analysis process. These earlier studies analyzed the impact of deficiencies in quality in a large number of diagnostic accuracy studies across a variety of systematic reviews, whereas our study assessed the impact of these quality items on estimates of diagnostic accuracy within systematic reviews. Furthermore, the number of studies with methodological deficiencies was small in a number of the systematic reviews included in our analysis, whereas other reviews contained only studies with deficiencies. Many of these studies with a deficient study design had a small sample size. Because the weight of an individual study depends on sample size, these studies had only a minor impact on the summary estimate of diagnostic accuracy. Furthermore, if 2 or more quality items influence accuracy but in opposing directions, the overall estimate obtained irrespective of quality may be similar to the estimate based on high-quality studies only. It is also possible that incomplete reporting has led to misclassification of design features in our project, which may have jeopardized our attempts to find differences in accuracy.

There are other potential explanations for our failed attempts at quality adjustment. The effects of several study-design features may not always be in the same predictable direction. Whether partial verification, for example, will lead to accuracy
Chapter 3

estimates that are unchanged, lower, or higher, depends on the pattern of verification and the reference standards being used. The ratio of patients with unverified positive index test results and patients of unverified negative test results matters, in particular when being verified or not is related to the presence or absence of the target condition.

Similar remarks have been made in the field of intervention studies, where more meta-epidemiologic studies like ours have been performed. The aim in meta-epidemiologic studies is to evaluate the importance of 1 or more design features across a substantial number of systematic reviews. These studies have shown that meta-epidemiologic studies require substantial numbers of systematic reviews with sufficient differences in methodological quality among the included studies. Furthermore, if the effects of design features vary in direction among reviews or even among studies within a single review, meta-epidemiologic studies may produce summary estimates that suggest no effect at all. Although we have found no systematic trend in results among strategies, reviews in which adjusting for quality has led to substantially different results clearly exist. Because we do not know the true magnitude of accuracy, it is impossible to tell whether the adjusted estimates were closer to the truth.

Not only did we fail to find support for our hypothesis that adjusting for quality will result in less optimistic estimates of test accuracy, we also found no evidence for the hypothesis that adjusting for quality leads to less heterogeneity in results and therefore to smaller confidence intervals. On the contrary, the alternative analyses generally produced broader confidence limits. The main reason for this result is that the alternative strategies were based on fewer studies.

Our study did not produce evidence for the superiority of one type of adjustment over another. Low-quality studies can produce accuracy statistics that do not differ from those obtained in high-quality studies. Although methodological quality may influence the results of meta-analyses, a direct association with results is not necessarily present.

In any review, poor quality will affect the trustworthiness of the conclusions of that review. Our results indicate that the strategy used to correct for quality may affect the estimated accuracy, but not in a predictable way. Our results also indicate that measuring and incorporating quality in a diagnostic review is not a simple task of routinely scoring a few standard quality items and then adjusting for these variables in a multivariable model.

There may be good reasons to identify some quality criteria as crucial for the credibility and applicability of any systematic review. An example could be the selection of the reference standard—QUADAS item 3. These criteria may then be used as inclusion criteria for the review, and authors of systematic reviews might want to report how many studies had to be excluded based on that criterion.
Quality adjustment in meta-analyses of diagnostic accuracy

Quality-assessment results of the studies included in a review remains a necessity because it notifies readers about the overall quality of the studies included in the review and may point out differences in design that can help to explain some of the heterogeneity in results. The QUADAS instrument can be used for that purpose. We propose to score “not reported” as a separate category where applicable, and we hope that a more widespread implementation of the STARD statement will lead to better reporting in future reports of diagnostic accuracy studies.\(^{61,62}\)

We feel it necessary that quality-assessment results in a systematic review be summarized in a table or a figure. A table can list the extent to which each of the studies fulfilled the quality criteria. A figure, such as the stacked bar chart in Figure 3.1, can then display the studies for which each of the respective criteria was fulfilled so that the reader can obtain an overview of the quality of the studies included in the review. Plotting results for all of the included studies in ROC space and coding individual studies by colour or with symbols can help readers recognize the characteristics of individual studies.

In our view, whether quality is also to be incorporated in a meta-analysis depends on several factors. In the first place, analyzing quality is not even an option if the number of included studies is too low. If the results are very heterogeneous, quality differences can be used to search for an explanation for the heterogeneity, and such a search can be accommodated by stratification or, if appropriate, regression analysis. Caution is needed because it is not unusual for the potential explanations for observed differences to outnumber the studies in a systematic review. It is important to recognize the major limitations of meta-epidemiologic approaches in meta-analysis.

Quality is a multidimensional concept, and the importance of individual quality items will vary from one research project to another. The goal of adjusting for quality differences in meta-analysis will remain attractive but elusive until we have large-scale systematic reviews and fully informative reporting in individual studies.

**Acknowledgments**

J.J.D. is supported in part by a Senior Scientist in Evidence Synthesis Award from the UK Department of Health.
Chapter 3

References


Quality adjustment in meta-analyses of diagnostic accuracy


Quality adjustment in meta-analyses of diagnostic accuracy


