How does study quality affect the results of a diagnostic meta-analysis?

Marie E Westwood*, Penny F Whiting and Jos Kleijnen

Address: Centre for Reviews and Dissemination, University of York, UK

Email: Marie E Westwood* - mew3@york.ac.uk; Penny F Whiting - penny.whiting@bristol.ac.uk; Jos Kleijnen - jk13@york.ac.uk

* Corresponding author

Abstract

**Background:** The use of systematic literature review to inform evidence based practice in diagnostics is rapidly expanding. Although the primary diagnostic literature is extensive, studies are often of low methodological quality or poorly reported. There has been no rigorously evaluated, evidence based tool to assess the methodological quality of diagnostic studies.

The primary objective of this study was to determine the extent to which variations in the quality of primary studies impact the results of a diagnostic meta-analysis and whether this differs with diagnostic test type. A secondary objective was to contribute to the evaluation of QUADAS, an evidence-based tool for the assessment of quality in diagnostic accuracy studies.

**Methods:** This study was conducted as part of large systematic review of tests used in the diagnosis and further investigation of urinary tract infection (UTI) in children. All studies included in this review were assessed using QUADAS, an evidence-based tool for the assessment of quality in systematic reviews of diagnostic accuracy studies. The impact of individual components of QUADAS on a summary measure of diagnostic accuracy was investigated using regression analysis. The review divided the diagnosis and further investigation of UTI into the following three clinical stages: diagnosis of UTI, localisation of infection, and further investigation of the UTI. Each stage used different types of diagnostic test, which were considered to involve different quality concerns.

**Results:** Many of the studies included in our review were poorly reported. The proportion of QUADAS items fulfilled was similar for studies in different sections of the review. However, as might be expected, the individual items fulfilled differed between the three clinical stages. Regression analysis found that different items showed a strong association with test performance for the different tests evaluated. These differences were observed both within and between the three clinical stages assessed by the review. The results of regression analyses were also affected by whether or not a weighting (by sample size) was applied. Our analysis was severely limited by the completeness of reporting and the differences between the index tests evaluated and the reference standards used to confirm diagnoses in the primary studies. Few tests were evaluated by sufficient studies to allow meaningful use of meta-analytic pooling and investigation of heterogeneity. This meant that further analysis to investigate heterogeneity could only be undertaken using a subset of studies, and that the findings are open to various interpretations.

**Conclusion:** Further work is needed to investigate the influence of methodological quality on the results of diagnostic meta-analyses. Large data sets of well-reported primary studies are needed to address this question. Without significant improvements in the completeness of reporting of primary studies, progress in this area will be limited.
Background

The use of systematic literature review to inform evidence-based practice in diagnostics is rapidly expanding. Although the primary diagnostic literature is extensive, there remain a number of problems for systematic reviews of diagnostic tests. Appropriate methods for rigorous evaluation of diagnostic technologies have been well established [1-5]. However, available studies have generally been poorly designed and reported [6-8]. Similarly, although a number of quality checklists for diagnostic accuracy studies have been proposed [9] and there is growing evidence on the effects of bias in such studies [10], there has been no rigorously evaluated, evidence-based quality assessment tool for diagnostic studies.

The objective of this study was to investigate the impact of quality on the results of a diagnostic meta-analysis, using regression analysis. A large diagnostic systematic review was required to enable the use of regression analysis to investigate the impact of components of quality upon results.

We have recently completed a systematic review, which aimed to determine the most appropriate pathway for the diagnosis and further investigation of UTI in children [11]. It included an assessment of the accuracy of tests for three different clinical stages of UTI: the diagnosis UTI, localisation of infection, and further investigation of patients with confirmed UTI. The nature of the tests included in these three clinical sections of this review differed. Tests used to diagnose UTI were generally labora-
tory-based or near-patient methods, with relatively objective interpretation of results, e.g. dipstick tests and microscopy. By contrast, tests used to investigate confirmed UTI mainly utilised imaging technologies which are largely subjective in their interpretation, and where diagnostic thresholds are difficult to define. Tests used to localise infection spanned both categories. We hypothe-
sised that the components of methodological quality affecting results were likely to differ between the three sections of the review. Such potential differences may indi-
cate a need for topic-specific checklists for the assessment of quality in diagnostic studies.

A secondary aim of this study was to contribute to the evaluation of QUADAS, an evidence-based tool for the assessment of the quality of diagnostic accuracy studies that was specifically developed for use in systematic reviews of diagnostic tests [12], by investigating the importance of specific QUADAS items.

Methods

We used QUADAS [12] (Table 1) to assess the quality of primary studies included in the review. Items were rated as 'yes', 'no', or 'unclear'. We examined differences in the individual QUADAS items fulfilled, as well as their impact on test performance. The review divided the diagnosis and further investigation of UTI into the following three clinical stages: diagnosis of UTI, localisation of infection, and further investigation of the UTI. Each stage used different types of diagnostic test, which were considered to involve different quality concerns.

Table 1: QUADAS

<table>
<thead>
<tr>
<th>Item #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Was the spectrum of patients representative of the patients who will receive the test in practice?</td>
</tr>
<tr>
<td>2.</td>
<td>Were selection criteria clearly described?</td>
</tr>
<tr>
<td>3.</td>
<td>Is the reference standard likely to correctly classify the target condition?</td>
</tr>
<tr>
<td>4.</td>
<td>Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests? (disease progression bias)</td>
</tr>
<tr>
<td>5.</td>
<td>Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis? (partial verification bias)</td>
</tr>
<tr>
<td>6.</td>
<td>Did patients receive the same reference standard regardless of the index test result? (differential verification diagnosis)</td>
</tr>
<tr>
<td>7.</td>
<td>Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? (incorporation bias)</td>
</tr>
<tr>
<td>8.</td>
<td>Was the execution of the index test described in sufficient detail to permit replication of the test?</td>
</tr>
<tr>
<td>9.</td>
<td>Was the execution of the reference standard described in sufficient detail to permit its replication?</td>
</tr>
<tr>
<td>10.</td>
<td>Were the index test results interpreted without knowledge of the results of the reference standard? (test review bias)</td>
</tr>
<tr>
<td>11.</td>
<td>Were the reference standard results interpreted without knowledge of the results of the index test? (diagnostic review bias)</td>
</tr>
<tr>
<td>12.</td>
<td>Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? (clinical review bias)</td>
</tr>
<tr>
<td>13.</td>
<td>Were uninterpretable/ intermediate test results reported?</td>
</tr>
<tr>
<td>14.</td>
<td>Were withdrawals from the study explained?</td>
</tr>
</tbody>
</table>
We analysed results grouped by clinical stage. Within these groups, we pooled studies of similar tests or test combinations where sufficient data were available and where pooling was clinically meaningful. (Table 2) The minimum number of studies that we required for regression analysis was ten. This choice was made based on published guidance [13,14].

We estimated summary receiver operator characteristic (SROC) curves using the following equation [15]:

\[
Sensitivity = \frac{1}{1 + \left(\frac{a}{1 - b}\right) \left(1 - \frac{Specificity}{1 - Specificity}\right)^{1+b}}
\]

a and b were estimated by regressing D against S for each study:

\[
D = a + bS
\]

\[
D = \{\logit (\text{sensitivity}) - \logit (1-\text{specificity})\} = \log\text{diagnostic odds ratio (DOR)}
\]

\[
S = \{\logit (\text{sensitivity}) + \logit (1-\text{specificity})\}
\]

We used both weighted and unweighted models. For the weighted model we weighted on sample size. We chose to weight on sample size rather than inverse variance, a method sometimes used in this type of analysis, as we believe that weighting on the inverse variance can produce biased results. The reason for this bias is that the DOR is associated with its variance and so large DORs will inevitably have large variances, which will be reflected in the weightings.

We assessed between study heterogeneity through visual examination of forest plots and statistically using the Q statistic [16]. Where sufficient data were available, we used regression analysis to investigate whether individual QUADAS items and additional variables thought likely to be associated with diagnostic accuracy were associated with the DOR and hence whether differences in these items between the studies accounted for some of the observed heterogeneity. Where data were available, the following additional variables were investigated:

- Patient age (<2 years, <5 years, <12 years and <18 years) was included to examine possible variation with age within the paediatric population.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Tests included in the regression analysis (number of studies)</th>
<th>Tests for which there were insufficient studies to permit regression analysis (number of studies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dipstick: nitrite (23), LE (14), nitrite or LE positive (15)</td>
<td>Clinical history (6)</td>
<td>Microscopy: pyuria (28), bacteriuria (22)</td>
</tr>
<tr>
<td>Microscopy: pyuria (28), bacteriuria (22)</td>
<td></td>
<td>Dipstick: nitrite and LE positive (9), glucose (4), protein (2), blood (1), protein and LE positive (1), combinations of 3 dipstick tests (5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Microscopy: pyuria or bacteriuria (8), pyuria and bacteriuria (8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Culture: standard (1), dipslide (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Combinations of different tests (10)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Localisation</th>
<th>Ultrasound (20)</th>
<th>Clinical history (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Imaging techniques: MCUG (7), MRI (1), CT (1), IVP (4), cystography (2), scintigraphy (3)</td>
<td>Laboratory based tests (16)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Further investigation</th>
<th>Detection of reflux: Ultrasound (28): standard (11), contrast enhanced (17)</th>
<th>Detection of reflux: IVP (4), voiding radionuclide cystography (3), NAG/creatinine ratio (1), scintigraphy (3), risk scoring system (1)</th>
</tr>
</thead>
</table>

Table 2: Tests included/excluded in the regression analysis

We used both weighted and unweighted models. For the weighted model we weighted on sample size. We chose to weight on sample size rather than inverse variance, a method sometimes used in this type of analysis, as we believe that weighting on the inverse variance can produce biased results. The reason for this bias is that the DOR is associated with its variance and so large DORs will inevitably have large variances, which will be reflected in the weightings.
• The geographic region where studies were conducted was included to account for possible regional differences in test technology and infective agent.

• Specific variations in index test technique were also included. For microscopy for pyuria and bacteriuria a variable on whether the sample was centrifuged was included, and for microscopy for bacteriuria a variable for Gram stain was included. For ultrasound for the detection of reflux a variable for whether or not the ultrasound involved a contrast agent was included.

• The SROC model [15], was extended to include each of the 14 QUADAS items and each of the variables above as individual covariates [17]. As each QUADAS item can be scored as "yes", "no" or "unclear", we included QUADAS items as categorical variables with 3 possible outcomes, thus including the comparisons of "yes vs no", and "yes vs unclear". This allowed us to make some distinction between associations of aspects of methodological quality with test performance and associations of completeness of reporting with test performance. A number of QUADAS items only received two of the three possible scores (i.e. were scored either "yes" or "no", or "yes" or "unclear", or "no" or "unclear"). These items were therefore included as dichotomous variables.

A multivariate linear regression analysis was conducted. Initially, we performed univariate analysis with all items included separately in the model. Items that showed moderate evidence of an association with D, defined as p < 0.10, were investigated further using step-down regression analysis. All items found to show moderate evidence of an association in the univariate models were entered into the multivariate model, then dropped in a step-wise fashion with the item with the weakest evidence of an association (largest p-value) dropped first. For covariates with more than one level, evidence of an association of one indicator variable with test performance was considered sufficient for inclusion in the model. The final model was achieved when all items remaining showed strong evidence of an association with D, defined as p < 0.05. Interaction terms were not included. Associations of covariates with D were expressed as relative diagnostic odds ratios (RDOR). The DOR is used as an overall measure of diagnostic accuracy. It is calculated as the odds of positivity among diseased persons, divided by the odds of positivity among non-diseased. When a test provides no diagnostic evidence then the DOR is 1.0. The RDOR is calculated as the DOR when the covariate is present divided by the DOR when the covariate is absent. It therefore provides an indicator of the overall impact on diagnostic accuracy of the presence of a given covariate.

![Graph](http://www.biomedcentral.com/1471-2288/5/20)

**Figure 1**
Numbers of quality items fulfilled by studies in the three sections of the review.
Results

Results of QUADAS assessment
The proportion of QUADAS items fulfilled by studies included in our systematic review was similar for each of the three clinical stages assessed in the review. Studies evaluating tests to diagnose UTI fulfilled a median of 8 (range 5–13) items, those evaluating tests used to localise infection also fulfilled a median of 8 (range 3–13) items, and those evaluating further investigations fulfilled a median of 7.5 (range 3–12) items. Figure 1 illustrates the number of QUADAS items fulfilled by studies in each category. The similarity in numbers of QUADAS items fulfilled masks apparent differences in the individual items fulfilled.

Figure 2 shows the proportion of studies that scored "yes", "no" and "unclear" for each of the QUADAS items, separately for the three sections of the review.

Tests for the diagnosis of UTI (n = 79 studies) [18-96]
The use of an inappropriate spectrum of patients and inadequate reporting of inclusion criteria were problematic for studies in this category. The majority of studies provided insufficient details on how the reference standard was performed. Studies failed to report sufficient details on clinical review bias, diagnostic review bias and test review bias to judge whether these were avoided. Study withdrawals and handling of uninterpretable results were also poorly reported.

Tests for the localisation of infection (n = 39 studies) [48,69,97-133]
The time delay between the index test and reference standard was more of a problem with these studies than with those on the diagnosis of UTI. The use of an appropriate reference standard was also an issue in some of these studies. Spectrum composition and reporting of details of how children were selected for inclusion in the study was better in these studies than in the studies of the diagnosis of UTI. Only around half of studies provided sufficient details of how the index test and reference standard were performed to allow replication of these tests. More studies in this category, almost 40%, provided information indicating that test and diagnostic review bias had been avoided, in the remainder of studies this information was not reported. As with studies of the diagnosis of UTI, reporting of clinical review bias, handling of uninterpretable results, and withdrawals from the study was poor.
Tests for the further investigation of confirmed UTI (n = 71 studies) [134-204]

As with studies of the diagnosis of UTI, spectrum composition and reporting of inclusion criteria were poor in this group. The time delay between the index test and reference standard was also an issue in many of these studies. Around half of studies reported that diagnostic and test review bias had been avoided, the remaining studies did not report whether the index test and reference standard were interpreted blind to the results of each other. This was similar to the situation seen for studies on the localisation of infection. Reporting of the reference standard was poor. As in all previous groups, studies also provided very little information on whether appropriate clinical information was available when test results were interpreted, how uninterpretable results were handled, and whether there were any withdrawals from the study and if so whether all withdrawals were accounted for.

Results of multivariate regression analysis

Tests for the diagnosis of UTI

Tests involving dipstick or microscopy techniques were the only categories where enough studies were available to enable regression analysis. Table 3 summarises the results of the regression analysis for studies assessing dipstick tests. For dipstick to detect urinary nitrite (23 studies) [20,26,28,34,36,40,41,43,52,54-57,60,63,66,72,74,84,88,93-95], the weighted analysis found that studies reporting that clinical review bias had been avoided had a DOR 4.7 (95% CI: 1.7, 12.7) times greater than those which did not report on whether clinical information was available to those interpreting the test results (p = 0.004). This is what would be expected, as the DOR is likely to be higher when those interpreting test results have access to appropriate clinical information similar to that, which would be available in practice. No studies reported the presence of clinical review bias. This was the only item investigated to show strong evidence of an association with test performance in the weighted multivariate analysis, although age and geographic region did show moderate evidence of an association in the univariate analysis. The unweighted analysis showed slightly different results. The same three items were found to show at least moderate evidence of an association in the univariate analysis. However, only country remained in the multivariate model, suggesting that studies conducted in North America showed higher accuracy than studies conducted in Europe or other areas (p < 0.05).

For dipsticks measuring urinary leukocyte esterase (14 studies) [20,28,34,36,43,56,57,60,63,66,72,84,94,95] and for dipsticks for the presence of either nitrite or leukocyte esterase (15 studies) [19-21,28,34,56,60,63,66,84-86,92,94-96], no items showed strong evidence of an association with the DOR in the weighted analysis. However, for urinary leukocyte esterase, the unweighted analysis found strong evidence of an association between patient age and the DOR. There was strong evidence (p = 0.015) that the dipstick was more accurate in children aged <12 years than in those aged <2 years (RDOR = 28.1, 95% CI: 2.3, 343.3). There was no evidence of any difference in accuracy between children aged <18 years and those aged <2 years (p = 0.703), and very little evidence of any difference between children aged <5 years and those aged <2 years (p = 0.158).

Table 4 summarises the results of the regression analysis for studies that assessed the accuracy of microscopy. In studies evaluating microscopy to detect pyuria three items showed a strong association with test performance in the weighted analysis (28 studies) [19-23,28,34,35,41,43,46,47,49,50,58,59,63,67,70,75,77,80,81,83,85,92-94]. The DOR was 1.3 (95% CI: 1.1, 1.6; p = 0.007) times higher in studies that adequately reported details of the reference standard execution. The DOR was lower, RDOR = 0.2 (95% CI: 0.1, 0.4; p < 0.001) in studies that did not report on reasons for withdrawals compared to studies in which it was unclear whether there were any withdrawals, and 1.8 times higher (95% CI: 1.0, 3.4; p = 0.056) in studies in which withdrawals were accounted for compared to those in which this was unclear. The DOR was lower, RDOR = 0.2 (95% CI: 0.1, 0.3; p < 0.001), in studies where samples were centrifuged compared to studies in which samples were not centrifuged. In the unweighted analysis, only centrifugation showed any evidence of an association with test performance (p = 0.08). All of these items, with the exception of centrifugation, relate to the completeness of reporting. The association for centrifugation is counter-intuitive, as we would expect centrifugation of the sample to lead to improved test accuracy.

Two items showed a strong evidence of an association with the DOR in the weighted analysis of studies evaluating microscopy to detect bacteriuria (22 studies) [20,21,23,28,34,35,41,47,50,61-64,67,70,76,77,80,85,90,91,94]. The DOR was 3.0 (95% CI: 1.6, 5.5; p = 001) times greater in studies in which incorporation bias was present compared to those in which it was avoided, and 5.3 (95% CI: 2.3, 12.0; p = 0.001) times greater if samples were Gram stained. We would expect both Gram staining and the presence of incorporation bias to increase test performance as found in the analysis. The unweighted analysis found very similar results.

Tests for the localisation of infection

Only the evaluation of ultrasound for the localisation of infection provided sufficient data to enable the conduct of regression analysis (20 studies) [48,69,97,99-102,109-
None of the QUADAS items, or other items investigated, showed moderate evidence of an association with the DOR in this analysis, using either the weighted or unweighted model.

Tests for the further investigation of confirmed UTI
Table 5 summarises the results of the regression analysis for studies assessing this clinical stage. The use of ultrasound to detect reflux was the only test in this category with sufficient data to support regression analysis (28 studies) [69,135,137,140,141,150,152,153,155,164,169,170,172,176-178,181,185,187,189,190,195-198,202-204]. Three items showed strong evidence of an association with the DOR in the weighted analysis. The DOR was 8.0 (95% CI: 2.9, 22.0; p < 0.001) times greater in studies that used contrast enhanced ultrasound compared to those that used standard ultrasound. As this was also thought to be a clinically important variable it was included in all further analyses. The DOR was 1.4 (95% CI: 1.0, 1.9; p = 0.033)

### Table 3: Results of the regression analysis for dipstick tests for the diagnosis of UTI

<table>
<thead>
<tr>
<th>Variable*</th>
<th>Univariate analysis</th>
<th>Multivariate model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RDOR (95% CI)</td>
<td>p-value</td>
</tr>
<tr>
<td><strong>Nitrite dipstick – weighted (n = 23 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical review bias avoided: yes vs unclear</td>
<td>4.7 (1.7, 12.7)</td>
<td>0.004</td>
</tr>
<tr>
<td>Age &lt;5 years vs &lt;2 years</td>
<td>5.8 (0.3, 101.8)</td>
<td>0.213</td>
</tr>
<tr>
<td>Age &lt;12 years vs &lt;2 years</td>
<td>2.2 (0.1, 35.8)</td>
<td>0.548</td>
</tr>
<tr>
<td>Age &lt;18 years vs &lt;2 years</td>
<td>3.6 (0.9, 14.9)</td>
<td>0.076</td>
</tr>
<tr>
<td>Europe vs North America</td>
<td>0.3 (0.1, 0.9)</td>
<td>0.041</td>
</tr>
<tr>
<td>Other areas vs North America</td>
<td>0.3 (0.1, 1.0)</td>
<td>0.050</td>
</tr>
<tr>
<td><strong>Nitrite dipstick – unweighted (n = 23 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical review bias avoided: yes vs unclear</td>
<td>3.4 (0.9, 13.7)</td>
<td>0.078</td>
</tr>
<tr>
<td>Age &lt;5 years vs &lt;2 years</td>
<td>7.3 (0.9, 61.8)</td>
<td>0.067</td>
</tr>
<tr>
<td>Age &lt;12 years vs &lt;2 years</td>
<td>2.9 (0.5, 18.3)</td>
<td>0.243</td>
</tr>
<tr>
<td>Age &lt;18 years vs &lt;2 years</td>
<td>3.8 (1.1, 13.4)</td>
<td>0.039</td>
</tr>
<tr>
<td>Europe vs North America</td>
<td>0.3 (0.1, 1.0)</td>
<td>0.044</td>
</tr>
<tr>
<td>Other areas vs North America</td>
<td>0.3 (0.1, 1.2)</td>
<td>0.089</td>
</tr>
<tr>
<td><strong>LE dipstick – weighted (n = 14 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LE dipstick – unweighted (n = 14 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test details reported: yes vs no</td>
<td>19.0 (1.9, 192.2)</td>
<td>0.017</td>
</tr>
<tr>
<td>Age &lt;5 years vs &lt;2 years</td>
<td>5.4 (0.5, 64.2)</td>
<td>0.158</td>
</tr>
<tr>
<td>Age &lt;12 years vs &lt;2 years</td>
<td>28.1 (2.3, 343.3)</td>
<td>0.015</td>
</tr>
<tr>
<td>Age &lt;18 years vs &lt;2 years</td>
<td>1.3 (0.3, 4.7)</td>
<td>0.703</td>
</tr>
<tr>
<td><strong>Nitrite or leukocyte esterase dipstick – weighted (n = 15 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference standard details reported</td>
<td>4.5 (0.9, 22.5)</td>
<td>0.064</td>
</tr>
<tr>
<td>North America vs Europe</td>
<td>5.0 (0.8, 10.5)</td>
<td>0.076</td>
</tr>
<tr>
<td>North America vs Other areas</td>
<td>1.1 (0.28, 5.0)</td>
<td>0.854</td>
</tr>
<tr>
<td><strong>Nitrite or leukocyte esterase dipstick – unweighted (n = 15 studies)</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Only items that showed moderate evidence (p < 0.10) for an association with the DOR in the univariate analysis are included.

*Items were dropped where there were too few studies in one category to allow a coefficient to be calculated.
times higher in studies that reported that disease progression bias had been avoided compared to those in which this information was not reported. No studies reported sufficient information to determine that disease progression bias was present. Studies in which details were provided on reasons for withdrawals had DORs that were 2.8 times higher (95% CI: 1.1, 6.9, p = 0.033) than those in which it was unclear whether there had been any withdrawals. There was no evidence of any difference in the DOR between studies that did not report on reasons for withdrawals and studies in which it was unclear whether there were any withdrawals (p = 0.502). In the unweighted analysis, only two items showed a strong evidence of an association with the DOR.

### Table 4: Results of the regression analysis for microscopy for the diagnosis of UTI

<table>
<thead>
<tr>
<th>Variable*</th>
<th>Univariate analysis</th>
<th>Multivariate model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RDOR (95% CI)</td>
<td>p-value</td>
</tr>
<tr>
<td><strong>Microscopy for pyuria: weighted (n = 28 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection criteria reported: yes vs no</td>
<td>2.4 (1.0, 5.9)</td>
<td>0.057</td>
</tr>
<tr>
<td>Reference standard details reported: yes vs no</td>
<td>3.6 (0.8, 16.1)</td>
<td>0.089</td>
</tr>
<tr>
<td>Test review bias avoided: yes vs unclear</td>
<td>4.8 (1.6, 14.7)</td>
<td>0.008</td>
</tr>
<tr>
<td>Diagnostic review bias avoided: yes vs unclear</td>
<td>5.5 (1.8, 17.1)</td>
<td>0.005</td>
</tr>
<tr>
<td>Uninterpretable results reported: no vs unclear</td>
<td>0.3 (0.0, 4.1)</td>
<td>0.364</td>
</tr>
<tr>
<td>Uninterpretable results reported: yes vs unclear</td>
<td>2.9 (0.9, 9.1)</td>
<td>0.073</td>
</tr>
<tr>
<td>Withdrawals accounted for: no vs unclear</td>
<td>0.2 (0.1, 0.7)</td>
<td>0.012</td>
</tr>
<tr>
<td>Withdrawals accounted for: yes vs unclear</td>
<td>1.9 (0.7, 5.3)</td>
<td>0.200</td>
</tr>
<tr>
<td>Europe vs North America</td>
<td>0.3 (0.1, 1.6)</td>
<td>0.190</td>
</tr>
<tr>
<td>Asia vs North America</td>
<td>0.6 (0.1, 6.6)</td>
<td>0.646</td>
</tr>
<tr>
<td>Other areas vs North America</td>
<td>0.2 (0.0, 1.3)</td>
<td>0.090</td>
</tr>
<tr>
<td>Age &lt;5 years vs &lt;2 years</td>
<td>0.1 (0.0, 0.4)</td>
<td>0.004</td>
</tr>
<tr>
<td>Age &lt;12 years vs &lt;2 years</td>
<td>0.1 (0.0, 0.3)</td>
<td>0.004</td>
</tr>
<tr>
<td>Age &lt;18 years vs &lt;2 years</td>
<td>0.2 (0.1, 0.5)</td>
<td>0.001</td>
</tr>
<tr>
<td>Sample centrifuged: yes vs no</td>
<td>0.3 (0.1, 0.6)</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>Microscopy for pyuria: unweighted (n = 28 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age &lt;5 years vs &lt;2 years</td>
<td>0.3 (0.1, 1.1)</td>
<td>0.076</td>
</tr>
<tr>
<td>Age &lt;12 years vs &lt;2 years</td>
<td>0.5 (0.1, 2.1)</td>
<td>0.337</td>
</tr>
<tr>
<td>Age &lt;18 years vs &lt;2 years</td>
<td>0.3 (0.1, 1.0)</td>
<td>0.048</td>
</tr>
<tr>
<td>Sample centrifuged</td>
<td>0.5 (0.2, 1.1)</td>
<td>0.080</td>
</tr>
<tr>
<td><strong>Microscopy for bacteriuria: weighted (n = 22 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorporation bias avoided: no vs yes</td>
<td>24.5 (1.0, 604.4)</td>
<td>0.050</td>
</tr>
<tr>
<td>Diagnostic review bias avoided: yes vs unclear</td>
<td>3.2 (0.8, 12.8)</td>
<td>0.092</td>
</tr>
<tr>
<td>Gram stain used: yes vs no</td>
<td>3.6 (1.3, 10.4)</td>
<td>0.018</td>
</tr>
<tr>
<td><strong>Microscopy for bacteriuria: unweighted (n = 22 studies)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease progression bias: yes vs unclear</td>
<td>0.05 (0.0, 1.5)</td>
<td>0.083</td>
</tr>
<tr>
<td>Incorporation bias avoided: no vs yes</td>
<td>32.5 (1.2, 895.0)</td>
<td>0.041</td>
</tr>
<tr>
<td>Uninterpretable results reported: yes vs unclear</td>
<td>7.1 (1.1, 46.9)</td>
<td>0.042</td>
</tr>
<tr>
<td>Withdrawals reported: no vs unclear</td>
<td>6.6 (1.0, 43.3)</td>
<td>0.049</td>
</tr>
<tr>
<td>Withdrawals reported: yes vs unclear</td>
<td>2.2 (0.3, 18.7)</td>
<td>0.447</td>
</tr>
<tr>
<td>Sample centrifuged: yes vs no</td>
<td>0.2 (0.0, 1.1)</td>
<td>0.058</td>
</tr>
<tr>
<td>Gram stain used: yes vs no</td>
<td>3.9 (0.9, 16.3)</td>
<td>0.062</td>
</tr>
</tbody>
</table>

*Only items that showed moderate evidence (p < 0.10) for an association with the DOR in the univariate analysis are included.

*Items were dropped where there were too few studies in one category to allow a coefficient to be calculated.
analysis there was very strong evidence that the DOR was higher in studies that used contrast enhanced ultrasound than those that used standard ultrasound (RDOR = 29.8, 95% CI: 13.5, 65.8, p < 0.001). Studies in which partial verification bias was avoided had DORs 4.1 times higher (95% CI: 1.1, 14.8) than those that did not (p = 0.034).

Discussion
The methodological quality of primary studies remains a significant issue for systematic reviews of diagnostic tests [8,205,206]. The STARD initiative has provided clear guidance for the reporting of diagnostic accuracy studies [5]. This should have a positive impact on the quality of the diagnostic literature in the future. The QUADAS tool facilitates systematic evaluation of the quality of diagnostic accuracy studies, and was specifically developed for use in systematic reviews of diagnostic tests [12]. However, where studies are poorly reported the information that can be derived from quality assessment becomes limited. We cannot know whether an unreported QUADAS item reflects a true methodological flaw or poor reporting of a study that may be methodologically sound. Many of the studies included in our review were poorly reported. Our assessment of the impact of components of methodological quality on diagnostic accuracy may therefore partially reflect completeness of reporting. Whilst poor reporting remains a widespread problem, it is almost impossible to assess the impact of components of methodological quality on the results of diagnostic meta-analyses.

The common practice of using summary quality scores in systematic reviews has been widely debated elsewhere [207-209]. Summary scores, when used to inform quality-based analyses, may mask important effects of individual quality components [210]. As we report, the numbers of QUADAS items that were adequately addressed by studies included in our review were similar between the three clinical stages assessed in the review. Had the number of QUADAS items fulfilled been used as a summary score, potentially important variations in the individual items fulfilled would have been hidden. We therefore advocate that components of quality assessment should be reported fully, and their impact on outcome measures analysed individually rather than as summary scores.

Although ours was a large review, it included 187 studies reporting 487 data sets, our analysis of the impact of methodological quality on diagnostic accuracy was severely limited both by the diversity of the included studies (few tests were evaluated by sufficient studies to allow meaningful use of meta-analytic pooling and investigation of heterogeneity), and by incomplete reporting. All of

---

**Table 5: Results of the regression analysis for ultrasound for the diagnosis of reflux**

<table>
<thead>
<tr>
<th>Variable*</th>
<th>Univariate analysis</th>
<th>Multivariate analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RDOR (95% CI)</td>
<td>p-value</td>
</tr>
<tr>
<td>Ultrasound for the detection of reflux: weighted (n = 28 studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of contrast enhanced ultrasound: yes vs no</td>
<td>23.9 (9.8, 58.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ultrasound for the detection of reflux, with ultrasound type forced into the model: weighted (n = 28 studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate reference standard: yes vs unclear</td>
<td>0.2 (0.0, 1.0)</td>
<td>0.047</td>
</tr>
<tr>
<td>Disease progression bias avoided: yes vs unclear</td>
<td>3.5 (1.4, 9.2)</td>
<td>0.011</td>
</tr>
<tr>
<td>Withdrawals accounted for: yes vs unclear</td>
<td>3.2 (1.2, 8.5)</td>
<td>0.020</td>
</tr>
<tr>
<td>Withdrawals accounted for: no vs unclear</td>
<td>(0.4, 0.1, 1.7)</td>
<td>0.175</td>
</tr>
<tr>
<td>Ultrasound for the detection of reflux: unweighted (n = 28 studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of contrast enhanced ultrasound: yes vs no</td>
<td>29.8 (13.5, 65.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ultrasound for the detection of reflux, with ultrasound type forced into the model: unweighted (n = 28 studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate reference standard: yes vs unclear</td>
<td>0.2 (0.0, 1.2)</td>
<td>0.075</td>
</tr>
<tr>
<td>Partial verification bias avoided: yes vs no</td>
<td>4.1 (1.1, 14.8)</td>
<td>0.034</td>
</tr>
</tbody>
</table>

*Only items that showed moderate evidence (p < 0.10) for an association with the DOR in the univariate analysis are included.

$Only 1 unclear

$Items were dropped where there were too few studies in one category to allow a coefficient to be calculated.
the data sets used were sub-optimal, in that the numbers of observations were low in comparison to the number of variables investigated in the multivariate analyses[13]. Although different types of diagnostic tests were evaluated in the three clinical stages used by the review, generalisability is limited in that all data concerned a single condition (UTI). A number of the items found to be associated with test performance related to specific test methodologies (e.g. Gram stain and contrast-enhanced ultrasound) and have no generalisability elsewhere. These items were found to show association in both the weighted and unweighted analyses. For the individual quality items there were some differences between the results of the weighted and unweighted analyses. In general, the results of weighted analyses showed more intuitive associations. Unweighted analyses more often produced results that were difficult to explain, for example, in leukocyte esterase dipstick tests the unweighted analysis found that the test was more accurate in the group of children aged <12 years than in those aged <2 years. This might be expected and would probably reflect a higher likelihood of sample contamination in younger children, however, no difference in accuracy was found between under 18’s and children aged <2 years. For both tests on the diagnosis of and further investigation of UTI weighted analyses showed an association between a number of variables relating to quality of reporting and diagnostic accuracy (well reported studies had higher DORs). We might expect this association to extend to diagnostic accuracy studies of all types of tests, but the present study is not adequate to demonstrate this. Weighted analysis of studies of ultrasound for the detection of reflux showed that the DOR was higher where studies reported information to determine that disease progression bias had been avoided. Disease progression bias is a particular issue for imaging studies of this type where follow-up examinations (used as the reference standard of diagnosis) may be scheduled some time after ultrasound (usually the initial examination). This association was not shown in the unweighted analysis.

The information derived from these analyses is also limited by the use of the summary ROC approach to pool studies. This method takes the DOR as the dependent variable. The DOR is used as a single indicator of test performance and shows how much more frequently a positive test result occurs in a person with the condition of interest than in one without the condition, relative to how much more frequently a negative result occurs in a person without the condition than in one with the condition. Using the DOR to investigate heterogeneity means that we cannot assess whether the factors investigated are associated with paired measures of diagnostic accuracy, such as sensitivity and specificity, or positive and negative likelihood ratios. Often factors that lead to an increase in sensitivity will lead to a decrease in specificity and vice versa. Factors that lead to this pattern of change may have no effect on an overall measure such as the DOR. Using the DOR to investigate heterogeneity may thus miss relevant clinical associations. Recently a new method for pooling sensitivity and specificity has been developed. This method is known as the "bivariate model" [211]. It preserves the underlying two-dimensional nature of the data and produces direct pooled estimates of sensitivity and specificity, incorporating any correlation that might exist between these two measures. The model can be extended to include explanatory variables leading to separate effects on sensitivity and specificity. This method has two advantages over the standard methods: (1) the pooled estimates of sensitivity and specificity take into account the correlation between these two measures; (2) the effect of possible sources of heterogeneity on both sensitivity and specificity can be investigated in a single model rather than just looking at the effect of these variables on a single measure of test performance, the DOR. These methods may have potential applications in future studies of this type.

Conclusion
Given the limitations we describe, the results of this study should be treated as hypothesis generating. Further work is needed to elucidate the influence of components of the methodological quality of primary studies on the results of diagnostic meta-analyses. Large data sets of well-reported primary studies are needed to address this question. Without significant improvements in the reporting of primary studies, progress in this area will be limited. The components of quality assessment should always be reported, and their impact on summary outcome measures be investigated, individually rather than as summary quality scores. Careful consideration should be given to the choice of weighting when conducting regression analyses. Weighting by sample size appears the most appropriate method for analyses of diagnostic accuracy studies, but this area requires further investigation.

Competing interests
The author(s) declare that they have no competing interests.

Authors' contributions
All authors contributed towards the conception and design of the study and the interpretation of the data. They also read and approved the final manuscript. PW and MW participated in data extraction, the analysis of data, and drafted the article.

Acknowledgements
We would like to thank Professor Martin Bland of the University of York and Mr Roger Harbord of the University of Bristol for their statistical advice.
References


mine the reliability of culture specimen of urine collected in perineal bag. Anales Españoles de Pediatría 1996, 45:149-152.
43. Giraldes M, Perozo M, González F, Rodríguez M: Infección urinaria cinta reactiva y sedimento urinario vs. urocultivo para determi-

46. Hardy JD, Furnell PM, Brunfitt W: Comparison of sterile bag, clean catch and suprapubic aspiration in the diagnosis of uri-


51. Holland PD, Doyle CT, English L: Comparative study of semiquantitative methods (leuko-

60. Lipstak GS, Campbell J, Stewart R, Hulbert WC: Screening for uri-


64. Manson R, Scholleff J, Johnston RJ, Scott R: The screening of more than 2,000 schoolgirls for bacteriuria using an auto-

66. Marrowski FJ, Owens D, Lecron JW: Use of the leukocyte ester-

69. Morin D, Veyrac C, Kozski PO, Lopez C, Dalla Vale F, Durand MF, Azencot J, Dumas R: Comparison of ultrasound and dimercapto-


73. Villanustre Ordóñez C, Buznego Sanchez R, Rodicio Garcia M, Rodrigo Saez E, Fernandez Seara MJ, Pavon Belinchon P, Castro-Gago M: Comparative study of semiquantitative methods (leuko-

74. Pryles CV, Eliot CR: Pyuria and bacteriuria in infants and chil-

75. Pullen J, Veyrac C, Kozski PO, Lopez C, Dalla Vale F, Durand MF, Azencot J, Dumas R: Comparison of ultrasound and dimercapto-


84. Sharief N, Hameed M, Petts D: Use of rapid dipstick tests to exclude urinary tract infection in children. British Journal of Bio-


act infection: are both ultrasonography and renal cortical scintigraphy necessary?


Lindseth D, Moncrieff M: Comparison of ultrasonography and intravenous urography after a urinary tract infection. Archives of Disease in Childhood 1986, 61:81-82.


using bivariate models in meta-analysis of studies of diagnostic accuracy. Submitted.

Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/5/20/prepub