Exploring sources of heterogeneity in systematic reviews of diagnostic tests

Jeroen G. Lijmer¹,*,†, Patrick M. M. Bossuyt¹ and Siem H. Heisterkamp²

¹Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, The Netherlands
²Department of RIVM, Bilthoven, The Netherlands

SUMMARY

It is indispensable for any meta-analysis that potential sources of heterogeneity are examined, before one considers pooling the results of primary studies into summary estimates with enhanced precision. In reviews of studies on the diagnostic accuracy of tests, variability beyond chance can be attributed to between-study differences in the selected cutpoint for positivity, in patient selection and clinical setting, in the type of test used, in the type of reference standard, or any combination of these factors. In addition, heterogeneity in study results can also be caused by flaws in study design. This paper critically examines some of the potential reasons for heterogeneity and the methods to explore them. Empirical support for the existence of different sources of variation is reviewed. Incorporation of sources of variability explicitly into systematic reviews on diagnostic accuracy is demonstrated with data from a recent review. Application of regression techniques in meta-analysis of diagnostic tests can provide relevant additional information. Results of such analyses will help understand problems with the transferability of diagnostic tests and to point out flaws in primary studies. As such, they can guide the design of future studies.

KEY WORDS: diagnostic tests; diagnostic accuracy; meta-analysis; systematic review

INTRODUCTION

Meta-analyses of diagnostic accuracy are becoming increasingly common. A MEDLINE search from January 1991 to January 1993 identifies 30 meta-analyses, indexed under the key word ‘sensitivity and specificity’. This number increases to 94 if the search is repeated for the period January 1999 to January 2001. The objective of meta-analyses is to provide a summary estimate of enhanced precision from a series of diagnostic test evaluations. However, merely reporting such a summary measure is of limited value. Understanding of the causes of heterogeneity in a meta-analysis will increase the scientific value and clinical relevance of the results [1, 2].

*Correspondence to: Jeroen G. Lijmer, Department of Psychiatry, University Medical Centre Utrecht, Heidelberglaan 100 3508 GA, Utrecht, The Netherlands
†E-mail: j.g.lijmer@psych.azu.nl

Copyright © 2002 John Wiley & Sons, Ltd.
Often there is considerable variation in the results of primary studies of diagnostic tests. This variation may be caused by chance alone (small sample sizes) or errors in calculating measures of accuracy statistics, but can also reflect true heterogeneity. Possible clinical sources of such heterogeneity are between-study differences in the type of test used, in the selected cutpoint for positivity, in patient selection and clinical setting, or any combination of these factors. In addition, heterogeneity in study results can also be caused by the fact that some studies were flawed by deficiencies in study design – methodological or artefactual heterogeneity [3].

Several papers have been written on the assessment of heterogeneity in meta-analysis of randomized clinical trials. These include reports of test statistics for the presence of heterogeneity [4–6], plots to detect outliers [7, 8] and meta-regression techniques to explore sources of variations in study results [9, 10]. Most of these techniques are based on summary measures for results presented in $2 \times 2$ tables of outcome versus treatment arm. As the results of studies evaluating diagnostic accuracy are often presented in a $2 \times 2$ table of index test versus reference test, many of the tools developed for meta-analyses of randomized clinical trials can also be used to explore heterogeneity in reviews of diagnostic tests.

This paper reviews some techniques to explore potential sources for heterogeneity in meta-analyses of diagnostic tests and demonstrates their use with data of a recently published meta-analysis.

**ASSESSMENT OF DIAGNOSTIC ACCURACY**

Studies on diagnostic accuracy consist of a comparison of the results from one or more index tests with those obtained from a reference test on a group of subjects. Ideally these should be patients suspected of the target condition that the test is designed to detect. For a dichotomous test, the study results can be summarized in a $2 \times 2$ table, which contains counts from the comparison of index tests versus reference test. From this $2 \times 2$ table several measures of diagnostic accuracy can be calculated (Figure 1). The most commonly reported pair of indices is sensitivity and specificity, the conditional probability of obtaining a positive result, in subjects with the target condition, and the conditional probability of obtaining a negative result, in subjects without the target condition.

The two measures are closely related to each other. Lowering the threshold for abnormality, and thereby increasing sensitivity, will always lead to a decrease in the specificity of the test. Therefore both measures are necessary to judge the discriminatory power of diagnostic test. The diagnostic odds ratio (DOR) has as an advantage that it is a single indicator of diagnostic

![Figure 1. A $2 \times 2$ table summarizing the results of a study on the diagnostic accuracy of the test with the formulae for the different measures of diagnostic accuracy.](image-url)
accuracy in contrast to most of the other measures, which have to be judged in pairs. The DOR can take values between 0 and infinity. High values indicate good test performance. A value equal to 1 means that a test does not discriminate between patients with and those without the target condition. Values less than 1 are the result of improper test interpretation or chance, if samples are small, as they indicate that normal test results are more common among patients with the target condition.

For an ordinal or continuous test result data can be presented in an ROC curve by plotting sensitivity against 1-specificity at different positivity thresholds. The data underlying a specific point on the curve can be presented in a $2 \times 2$ table.

**EXAMPLE: ACCURACY OF D-DIMER ASSAYS**

We will use as an example a set of 13 primary studies of D-dimer assay accuracy for acute venous thromboembolism [11]. D-dimer assays measure the level of D-dimer, a fragment specific for the degradation of fibrin, in blood or plasma. D-dimer assays might therefore constitute a useful diagnostic tool to refute venous thromboembolism at referral. The 13 studies included in the meta-analysis evaluated 15 Latex, 15 Elisa and 2 rapid D-dimer assays. The D-dimer assay was compared to ultrasonography or venography in patients suspected for deep leg vein thrombosis, or to ventilation-perfusion lung scanning (VQ scan) or angiography in patients suspected for pulmonary embolism. We will restrict ourselves to the subgroup of studies evaluating Latex and Elisa assays, which were methodologically appraised by two independent reviewers as described in detail elsewhere [12].

**EXPLORING HETEROGENEITY OF PRIMARY STUDIES**

The choice of a summary measure is not straightforward. In theory all of the measures outlined in Figure 1 can be pooled with random or fixed effects models. Ideally the summary measure of choice is less prone to statistical heterogeneity than its competitors, and easily interpretable. Engels et al. examined empirically in 125 meta-analyses of treatment effect if the heterogeneity differed for different summary measures [5]. They concluded that risk differences usually displayed more heterogeneity than odds ratios for the same meta-analyses.

Similar research in the field of diagnostic accuracy is still lacking. One could examine the heterogeneity of summary measures in a forest plot. Figure 2 shows forest plots of sensitivity, specificity, log diagnostic odds ratio and log-likelihood ratio of a negative test, ordered by size of the study. It shows a large variation in study results and visually the relative measures show less heterogeneity than sensitivity and specificity.

In the remaining part of this paper we will focus on the DOR as a summary measure. However, the techniques and methods described here can also be adopted to examine heterogeneity of another summary measure of diagnostic accuracy.

**Quantifying heterogeneity**

Several statistics have been developed to quantify the amount of heterogeneity of the results in a set of studies [4, 6, 13]. All of these statistics can be used with diagnostic test evaluations.
Figure 2. The log-odds ratio (ln DOR), log-likelihood ratio of a negative test (ln LR-), sensitivity and specificity of 30 evaluations of D-dimer assays for detecting venous thromboembolism. Points indicate the estimates. Horizontal lines are 95 per cent confidence intervals for the estimates.

Significant values of heterogeneity statistics should indicate a larger between-study variability than can be expected by chance alone. However, these tests have low statistical power to detect heterogeneity when the number of studies included is small and detect clinically unimportant heterogeneity when the number of studies is large [9, 14]. Newer measures of heterogeneity are being developed which might overcome some of the pitfalls of the old statistics [15].

As long as these are not available, the $Q$-statistic seems to be the most robust and therefore the best choice [14]. The $Q$-statistic of the DORs in the 30 D-dimer studies was 46.2 with a $p$-value of 0.022, indicating statistical heterogeneity. As the interpretation of the size of this statistic is not straightforward, clinical knowledge and common sense remain indispensable when deciding whether or not to pool the results of diagnostic studies.

Graphical presentation of the individual study results can aid in the investigation of possible heterogeneity. A plot of the results of a diagnostic study in ROC space, sensitivity versus 1-specificity, will usually demonstrate a large variability of the study results; yet it is difficult to interpret how much can be attributed to chance. Studies with the same DOR may...
plot on different points in ROC space by mere differences in the positivity threshold used. A useful diagram was proposed by Galbraith [7]. The log-odds ratio of each study divided by its standard error (the Z-statistic) is plotted against the reciprocal of the standard error. Small studies with less precise results will appear on the left side and the largest trials will be plotted on the right end of the figure. A regression line, through the origin, represents the overall log-odds ratio. The dotted lines two units above and below the solid line represent the 95 per cent boundaries of the overall log-odds ratio. The majority of the study results are expected to lie in this area in the absence of heterogeneity. The characteristics of the studies lying near or outside these boundaries can be examined closely, keeping in mind that the results of such examinations will be post hoc explanations. In Figure 3, for example, all four non-blinded trials lie above the regression line and two of them near the upper line.

Causes of heterogeneity

One of the difficulties in summarizing diagnostic studies is that the results may vary at different positivity thresholds. In Figure 4(a) the results of the 30 D-dimer studies are plotted in ROC space. Note that the studies using a low threshold for positivity are in the right upper quadrant of the figure, indicating a high sensitivity and low specificity. It is possible that sensitivity and specificity vary with the diagnostic odds ratio remaining constant. However, one would then expect the study results to follow a symmetrical ROC curve. This may not always be the
Figure 4. (a) Receiver operating characteristic plot of 30 evaluations of D-dimer assays for detecting venous thromboembolism. The points are stratified by the reported cut-off for positivity as shown in the legend. (b) Log-odds ratio ($D$) and the sum of the logit of sensitivity and the logit of ($1 - \text{specificity}$) of these 30 studies. The fitted line shows the average log-odds ratio ($D$).

To take possible variations of the DOR at different positivity thresholds into account, Moses et al. suggested the following meta-analytic regression model [16]:

$$D = \alpha + \beta S$$

(1)

where

$$D = \logit(\text{sensitivity}) - \logit(1 - \text{specificity})$$

$$= \log\{\text{sensitivity}/(1 - \text{sensitivity})\}/[(1 - \text{specificity})/\text{specificity}]\}

$$= \log(\text{DOR})$$

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

$$= \log\{\text{sensitivity}/(1 - \text{sensitivity})\} \times [(1 - \text{specificity})/\text{specificity}\}$$

$D$ is natural logarithm of the DOR of the individual studies. $S$ describes the leniency of the test positivity criterion. A lenient criterion will amount in a high number of positive test results and thus a high sensitivity and low specificity. A high sensitivity with a low specificity will result in a large number for $S$. The intercept ($\alpha$) is equal to the DOR of the point were sensitivity matches specificity and $S$ is 0. It can be interpreted as the pooled DOR of the corresponding test in case the parameter for the slope is estimated close to 0 and non-significant, indicating no variability due to threshold differences. The parameter for the slope ($\beta$) expresses variation of the DOR across individual studies at different positivity thresholds. The parameters are directly related to the distribution of the test results in patients.
with and without the target condition of interest as follows:

\[ z = \frac{2 \pi (\mu_{nd} - \mu_d)}{\sqrt{3} (\sigma_{nd} + \sigma_d)} \] (2)

\[ \beta = \frac{(\sigma_{nd} - \sigma_d)}{(\sigma_{nd} + \sigma_d)} \] (3)

where \( \mu_{nd} \) and \( \sigma_{nd} \) are the mean and standard deviation of the test results in the non-diseased patients and \( \mu_d \) and \( \sigma_d \) the mean and standard deviation of the test results in the diseased patients, assuming logistic distributions in both populations.

From (3) one can derive that the parameter of the slope will be zero in the case where the variance of the test results is similar in patients with and without the target condition. In this unique case of so-called shift distributions the DOR will be constant across studies. In all other cases the DOR is likely to vary with different thresholds. This means that by adding the slope parameter to the model, one is examining whether the variance of the test results is different between the diseased and the non-diseased, a possible clinical source of heterogeneity, or more general if the distributions are similar.

Moses et al. also presented a formula to transform the parameters for the intercept and the slope back to ROC space and plot a summary ROC curve using the conventional axes of sensitivity against \((1 - \text{specificity})\)

\[
\text{Sensitivity} = \left[ 1 + e^{-z/(1-\beta)} \cdot \left( \frac{\text{Specificity}}{1 - \text{Specificity}} \right)^{(1+\beta)/(1-\beta)} \right]^{-1}
\] (4)

Figure 4(b) shows the results of a fixed effects analysis for the D-dimer data set. The line has an intercept of 2.64 and a slope of \(-0.081\) indicating some variation of the DOR at different positivity thresholds.

**Methodological heterogeneity**

An important source of heterogeneity are variations in study quality. A large survey of the diagnostic literature (1990–1993) in five major journals showed that only 18 per cent of the studies satisfied five of the seven examined methodological standards [17]. These findings have been confirmed in other studies. This observation raises the problem that if one restricts a meta-analysis to studies of the highest scientific validity (studies fulfilling all criteria) only a minority of the available data can be used. A recent study empirically examined the impact of shortcomings in design, data collection and reporting on the estimates of diagnostic accuracy in 18 meta-analyses. This study confirmed that studies of lower methodological quality, particularly those including non-representative patients or applying different reference standards, tend to overestimate the diagnostic performance of a test [12]. It is possible that within a single meta-analysis these criteria might be less important and the effect of other criteria, for example not blinding the result of the index test to the reviewers, might give more bias. One could therefore consider not restricting the analysis to studies of the highest scientific validity and to explore the effect of design shortcomings within each meta-analysis instead [18]. Which validity criteria should be assessed? Several lists of items for primary studies

of diagnostic tests studies are available [12, 17, 19–21]. A very comprehensive checklist has been developed by the Cochrane Methods Working Group on Screening and Diagnostic tests [22]. A validity checklist should at least include the following elements [12, 23]:

(a) **Population of recruitment.** A relevant clinical population for recruitment 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. 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.

(b) **Method of patient selection.** 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.

(c) **Method of verification.** Partial verification bias looms if the decision to perform the reference standard is based on the result of the index test and not all patients are subjected to the reference standard. Alternatively, the estimation of diagnostic accuracy can also be biased if a subgroup of patients is verified by a different and less thorough standard (differential reference standard bias). An example would be test negative patients being verified through follow-up, and test positive patients by pathology.

(d) **Method of interpretation of tests.** 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.

(e) **Methods to avoid residual confounding.** If the reference standard is performed later in time than the index test, for example the occurrence of an event during follow-up, interventions in that period should be blinded to the index test result to avoid ‘the treatment paradox’. If the index test results are not blinded, patients with abnormal test results might receive more interventions than patients with normal test results, thereby decreasing the possible differences in event rate at the end of the study.

**Clinical heterogeneity**

Heterogeneity in study results may also be caused by clinical differences, so called clinical heterogeneity [1]. Variations in the study populations, the tests and the reference tests among the primary studies can all result in different estimates of diagnostic accuracy.

There are many examples of differences in diagnostic accuracy between subgroups of patients within primary studies of tests. Two studies showed that sensitivity and specificity of exercise electrocardiography to detect coronary lesions differed between subgroups of patients defined by age, sex, symptoms, use of medication or the extent of the disease [24, 25]. Other examples include variations in the diagnostic accuracy of the dipstick test to detect urinary tract infections, magnetic resonance imaging to detect multiple sclerosis and a tuberculosis test depending on the degree of clinical suspicion [26–28] as well as mammography for breast cancer screening depending on the use of hormone replacement therapy [29]. A biological explanation for these differences is the fact that disease status is seldom a true dichotomous classification. Often there is a spectrum of disease ranging from small limited forms, some of them even without symptoms, to more extensive conditions causing severe symptoms. The sensitivity of a test will differ depending on the severity of disease in the subgroup.
Furthermore the degree of co-morbidity of patients without the disease can vary between subgroups affecting the specificity of a test. It is reasonable to believe that these within-study differences between subgroups of patients will also be a source of variation when diagnostic accuracy is compared across study populations in meta-analyses of tests.

Differences in the thresholds for positivity of the index test are one source of variation caused by variation in the execution and/or interpretation of the test. Other differences in the testing protocol can also be a possible cause of variation. Nelemans et al. found in their meta-analysis of magnetic resonance angiography for peripheral arterial disease that the diagnostic accuracy improved in the case of three-dimensional imaging compared to two-dimensional imaging and when additional post-processing techniques were used rather than one type of projection [30]. Another example from the field of peripheral arterial disease was the superiority of colour-guided duplex over normal duplex ultrasonography, demonstrated in a systematic review [31].

It is likely that similar variations in the protocol of the reference test are an additional possible cause of heterogeneity in meta-analyses of diagnostic tests. The reference standard for deep venous thrombosis, for example, is venography. In recent evaluations of D-dimer assays this reference standard has been replaced by serial compression ultrasonography, which has a very high agreement with venography. However, some evaluations have used one single ultrasound examination as reference standard, which may detect fewer patients with deep venous thrombosis. One can therefore expect that the accuracy of D-dimer assays depends on the type of reference standard used.

To examine whether heterogeneity in study results can be explained by methodological and/or clinical differences, the SROC model introduced earlier can be extended to include covariates [16]. The resulting parameter estimates of the covariates can be interpreted, after antilogarithm transformation, as relative DORs (RDOR). For example consider the model

\[ D = \alpha + \beta S + \gamma \text{BLIND} \]  

(5)

where

\[ \text{BLIND} = 1 \text{ in the case where the reference test is interpreted independent of the index test} \]

and

\[ \text{BLIND} = 0 \text{ in the case where the reference is interpreted with information of the index test.} \]

The estimate of \( \gamma \) indicates the diagnostic performance of a test in blinded studies relative to the performance of the same test in studies lacking blinding. If the RDOR (exp(\( \gamma \))) is smaller than 1, it indicates that studies with blinding yield smaller estimates of diagnostic accuracy than studies failing it. When estimating the parameters it is important to incorporate an estimation of the between-study variance by using a random effects estimation method. Ignoring this variance will underestimate the standard errors of the estimated parameters, possibly resulting in incorrect conclusions on the importance of a covariate [15].

Table I. The characteristics of the primary studies evaluating a D-dimer assay (n = 30).

<table>
<thead>
<tr>
<th>Study characteristic</th>
<th>Score</th>
<th>Number of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recruitment</td>
<td>Clinical population</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Case-control</td>
<td>0</td>
</tr>
<tr>
<td>Patient selection</td>
<td>Consecutive</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Non-consecutive</td>
<td>0</td>
</tr>
<tr>
<td>Verification</td>
<td>Complete</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Different reference tests</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Partial</td>
<td>0</td>
</tr>
<tr>
<td>Interpretation of test results</td>
<td>Blinded</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Not blinded</td>
<td>4</td>
</tr>
<tr>
<td>Patients</td>
<td>Outpatients</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Outpatients and inpatients</td>
<td>25</td>
</tr>
<tr>
<td>Indication</td>
<td>Pulmonary embolism</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Deep venous thrombosis</td>
<td>19</td>
</tr>
<tr>
<td>Test type</td>
<td>Elisa</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Latex</td>
<td>15</td>
</tr>
</tbody>
</table>

**Application**

In our example of 30 D-dimer assays, the following methodological characteristics were assessed: recruitment method; patient selection; verification method, and interpretation of test results. In addition, the following study characteristics were scored: threshold (in ng/ml); study population (outpatients or inpatients and outpatients); indication of the test (pulmonary embolism or deep venous thrombosis), and type of test (Elisa or Latex). Table I contains a summary of the characteristics of the included studies. The methodology of the included studies was high as all studies recruited patients consecutively from a clinical population and verified all patients with a single standard.

Four parameters were added to the standard SROC model (1) to evaluate variations in blinding, study population, indication and test type as potential sources of heterogeneity. A weighted linear regression analysis was performed to estimate the parameters, where weights proportional to the reciprocal of the variance of the log DOR represented the within-study variation, while random effects between studies were estimated using restricted maximum likelihood estimation, taking into account the correlation between $S$ and $D$ within each study. Table II shows the results of the multivariate analysis. It can be seen that none of the examined characteristics or the slope parameter explained the heterogeneity in the DOR of the primary studies. A second model without covariates and slope parameter was used to calculate a single SROC curve (Figure 5). Note that a considerable amount of the heterogeneity remained unexplained, as the residual variance was 0.27.

Although test type does not seem to explain the observed heterogeneity in the DOR, Figure 5 clearly shows that it is an explanation for the observed heterogeneity in $S$. The results of the studies evaluating Elisa assays have on average higher sensitivities and lower specificities, which is likely to be caused by the use of less stringent criteria for positivity in these studies compared to those evaluating Latex assays. The minimal detection limit of Latex assays is higher than that of Elisa assays which restricts the range of possible thresholds in these studies. This has clinical implications as the test is used to refute venous thromboembolism

Table II. Estimates of the SROC parameters with and without covariates.

<table>
<thead>
<tr>
<th>Study characteristic</th>
<th>$B$</th>
<th>(95% CI)</th>
<th>Relative DOR</th>
<th>$B^*$</th>
<th>(95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.49</td>
<td>(1.95; 5.04)</td>
<td>2.61</td>
<td>(2.27–2.94)</td>
<td></td>
</tr>
<tr>
<td>$S$</td>
<td>-0.084</td>
<td>(-0.24; 0.073)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation (blinded versus not blinded)</td>
<td>-0.41</td>
<td>(-1.58; 0.77)</td>
<td>0.67</td>
<td>(0.21;2.16)</td>
<td></td>
</tr>
<tr>
<td>Patients (outpatients versus inpatients and outpatients)</td>
<td>-0.14</td>
<td>(-1.14; 0.86)</td>
<td>0.87</td>
<td>(0.32;2.36)</td>
<td></td>
</tr>
<tr>
<td>Indication (deep venous thrombosis versus pulmonary embolism)</td>
<td>-0.057</td>
<td>(-0.91; 0.80)</td>
<td>0.95</td>
<td>(0.40;2.23)</td>
<td></td>
</tr>
<tr>
<td>Test type (Latex versus Elisa)</td>
<td>-0.55</td>
<td>(-1.29; 0.20)</td>
<td>0.58</td>
<td>(0.28;1.22)</td>
<td></td>
</tr>
</tbody>
</table>

* Model without covariates.

Figure 5. Receiver operating characteristic plot of 30 evaluations of D-dimer assays for detecting venous thromboembolism. The open circles show studies evaluating a Latex assay, the filled circles show the results of studies evaluating an Elisa assay. The fitted line shows the summary ROC curve.

at referral, which requires a high sensitivity. Differences in $S$ can be tested for statistical significance by comparing the mean $S$ across subgroups or explored with regression analysis. The latter can be useful in those situations were the explanation of the possible differences is less straightforward.

DISCUSSION

In this paper we have shown how one can explore sources of heterogeneity in meta-analyses of diagnostic tests. Our example demonstrated the applicability of these techniques, which can
provide relevant additional information on the topic of interest. It is important to recognize that there are several causes of heterogeneity: artefactual, methodological as well as clinical.

In this paper we demonstrated the exploration of heterogeneity of diagnostic tests using the DOR as summary measure. An advantage of this method is that it can control for variation due to threshold differences. Being a single measure, the DOR can be a suitable method to compare the overall diagnostic accuracy of different diagnostic tests [30, 31]. Ideally one would like to have comparative studies, limiting the amount of variation due to population differences. A disadvantage of the DOR is that it cannot be used directly by physicians to calculate the probability of disease associated with a specific test outcome, as it corresponds to a range of linked sensitivities and specificities [32]. It is only possible to obtain a summary estimate of sensitivity, or specificity, by specifying the value of the other. Where the goal of the meta-analysis is to obtain accuracy measures for clinical practice it is more useful to focus first on summarizing sensitivity and specificity. If there is no heterogeneity in one of the two measures, one can calculate a summary point estimate for this measure and explore the sources of heterogeneity for the other. However, if both measures show a large variability in their results, threshold differences are a very likely source and the SROC approach outlined here is more suitable.

It will not always be possible to examine all sources of clinical heterogeneity. The number of primary studies that meet the inclusion criteria of a review might be small. In such a situation there may be several alternative explanations of the heterogeneity found up to the point where the possible explanations for the between study differences outnumber the available data points. These problems are similar to performing subgroup analyses in small trials. The strength of conclusions based on the results of such explorations will depend whether or not they were prespecified, as well as on their magnitude, precision and plausible (biological) explanations.

Another problem is that the information will often not be reported in enough detail to assess the validity and clinical characteristics of the primary studies. Several diagnostic reviews have noticed this problem when appraising the primary studies [12, 30]. This might improve in the near future as reporting guidelines are developed for the primary studies of diagnostic tests.

Despite these restrictions, it is still useful to explore sources of heterogeneity and to incorporate it explicitly in the analysis. The results of such analyses will help to understand problems with the transferability of diagnostic tests and may help to point out deficiencies in primary studies. As such, heterogeneity offers opportunities for increasing our knowledge, rather than threats to our efforts to synthesize the available evidence.

REFERENCES