Evaluation of diagnostic tests: from accuracy to outcome
Lijmer, J.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.

Download date: 08 Dec 2018
Exploring sources of heterogeneity in systematic reviews of diagnostic tests

Jeroen G. Lijmer, Patrick M.M. Bossuyt, Siem H. Heisterkamp

Submitted
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 in results beyond chance can be attributed to between study differences in the selected cut points 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.

We demonstrate that application of regression techniques in meta-analysis of diagnostic tests can provide relevant additional information. Results of such analyses will help us to better 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.
Exploring sources of heterogeneity in systematic reviews of diagnostic tests

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 keyword “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 As is recognized by the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group who recommend an assessment of heterogeneity in their reporting checklist.2

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) but can also reflect true heterogeneity. Possible sources of such heterogeneity beyond chance are between study differences in the type of test used, in the selected cut point 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.3

Several papers have been written on the assessment of heterogeneity in meta-analysis of randomised clinical trials. These include reports of test statistics for the presence of heterogeneity4–6, plots to detect outliers7, 8 and meta-regression techniques to explore sources of variations in study results9, 10. Most of these techniques are based on summary measures for results presented in 2x2 tables of outcome versus treatment arm. As the results of studies evaluating diagnostic accuracy are often presented in a 2x2 table of index test versus reference test, many of the tools developed for meta-analyses of randomised clinical trials can also be used to explore heterogeneity in reviews of diagnostic tests.

This paper reviews some of the 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 summarised in a
2x2 table, which contains counts from the comparison of index tests versus reference test. From this 2x2 table several measures of diagnostic accuracy can be calculated (fig 1). The most popular pair of indices is sensitivity and specificity, the conditional probability of obtaining a true positive result, in subjects with the target condition, and the conditional probability of obtaining a true 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 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 as they indicate that normal test results are more common among patients with the target condition.

For a continuous test result data can be presented as a 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 2x2 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 cross-linked fibrin, in blood or plasma. D-dimer assays might therefore constitute a useful diagnostic tool to refute venous

**Figure 1 Measures of diagnostic accuracy**

<table>
<thead>
<tr>
<th>Target condition (reference test result)</th>
<th>Present</th>
<th>Absent</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Normal</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Totals</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{a}{a+c} \)

Specificity = \( \frac{d}{b+d} \)

Positive Predictive value = \( \frac{a}{a+b} \)

Negative predictive value = \( \frac{c}{c+d} \)

Likelihood ratio abnormal test =

\[ \frac{\text{Sensitivity}}{1 - \text{Specificity}} \]

Likelihood ratio normal test =

\[ \frac{1 - \text{Sensitivity}}{\text{Specificity}} \]

Diagnostic Odds Ratio =

\[ \frac{\text{Odds(Sensitivity)}}{\text{Odds(1-Specificity)}} = \frac{(a*d)}{(c*b)} \]

92
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 as described in detail elsewhere.12 Two reviewers independently appraised the study characteristics of the individual studies. Disagreements were resolved by consensus.

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 and colleagues 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.

We are not aware of similar research in the field of diagnostic accuracy. One could examine the heterogeneity of summary measures by plotting them against the size of the study. This is outlined in figure 2 for sensitivity, specificity, diagnostic odds ratio and likelihood ratio of a positive test. It shows a large variation in study results.

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. 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. 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 inadmissible when deciding whether or not to pool the results of diagnostic studies.

Figure 2 Heterogeneity of different measures of diagnostic accuracy
Exploring sources of heterogeneity in systematic reviews of diagnostic tests

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. 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% 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 should be examined closely. In figure 3, for example, all four non-blinded trials lie above the regression line and two of them near the upper line.

**Figure 3** Galbraith plot of log odds ratios (ln D) of D-Dimer assays for the detection of venous thromboembolism.
Causes of heterogeneity

One of the difficulties in summarizing diagnostic studies is that the results may vary due to threshold differences. In figure 4a 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 case. To take possible variations of the DOR due to threshold differences in account Moses et al. suggested the following meta-analytic model\(^6\)

\[
D = \alpha + \beta S
\]

where

\[
D = \logit(\text{sensitivity}) - \logit(1-\text{specificity}) = \log\left\{\text{odds(sensitivity)}/\text{odds(1-specificity)}\right\} = \log(\text{DOR})
\]

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

The intercept (\(\alpha\)) can be interpreted as the common DOR of the corresponding test. The parameter for the slope (\(\beta\)) expresses variation of the DOR across individual studies due to positivity threshold differences. The parameters are directly related to the distribution of the test results in patients with and without the target condition of interest as follows:

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

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

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.
Exploring sources of heterogeneity in systematic reviews of diagnostic tests

Figure 4: Threshold differences

- **Sensitivity**
- **1-Specificity**

Legend:
- ○ 0-200 ng/ml
- □ 201-400 ng/ml
- ● 401-600 ng/ml
- ■ >600 ng/ml
From (3) one can derive that the parameter of the slope will be zero in case 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. Moses et al also presented a formula to transform the parameters for the intercept and the slope back ROC space and plot a summary ROC curve using the conventional axes of sensitivity against (1-specificity). Figure 4b shows the results of a fixed effects analysis for the D-Dimer dataset. The line has an intercept of 2.64 and a slope of -0.081 indicating some variation due to threshold differences.

Another 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% of the studies satisfied five of the seven examined methodological standards. 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. 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 give more bias. We therefore recommend not to use validity criteria as an exclusion condition for meta-analysis of diagnostic test, but to explore the effect of design shortcomings within each meta-analysis. Which validity criteria should be assessed? Several lists of items for primary studies of diagnostic tests studies are available. A very comprehensive checklist has been developed by the Cochrane Methods Working Group on Screening and Diagnostic tests. A validity checklist should at least include the following elements:

- **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.
Exploring sources of heterogeneity in systematic reviews of diagnostic tests

- **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.

- **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, if a subgroup of patients for example those with negative test results are verified by a different less thorough, standard, for example follow-up, the estimation of diagnostic accuracy might also be biased (differential reference standard bias).

- **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.

- **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’.

Heterogeneity in study results may also be caused by clinical differences, so called clinical heterogeneity. 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. 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 as well as mammography for
breast cancer screening depending on the use of hormone replacement therapy. 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.

We already discussed differences in the thresholds for positivity of the index test as a source of variation. 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 case of 3D imaging compared to 2D imaging and when additional post processing techniques were used rather than one type of projection. 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.

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 co-variates. The resulting parameter estimates of the co-variates can be interpreted, after anti-logarithm transformation, as relative DORs (rDOR). For example consider the model

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

where
BLIND = 1 in case the reference test is interpreted independent of the index test and
BLIND = 0 in case 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 $r\text{DOR (exp(}\gamma\text{))}$ 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.

**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 in- and outpatients), indication of the test (pulmonary embolism or deep venous thrombosis), and type of test (Elisa or Latex). Table 1 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.

We added four parameters 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 2 shows the results of the multivariate analysis. It can be seen that none of the examined characteristics explained the heterogeneity in the primary studies. A second model without covariates 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.
Table 1. Results of the scoring of study characteristics (n=30)

<table>
<thead>
<tr>
<th>Study characteristic</th>
<th>Score</th>
<th>No. 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>out 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>Assay</td>
<td>elisa</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>latex</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 2. Results of regression analysis

<table>
<thead>
<tr>
<th>Study characteristic</th>
<th>B</th>
<th>Relative DOR (95% CI)</th>
<th>B*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.49</td>
<td></td>
<td>-0.054</td>
</tr>
<tr>
<td>S</td>
<td>-0.084</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation (blinded vs not blinded)</td>
<td>-0.405</td>
<td>0.667 (0.206-2.16)</td>
<td></td>
</tr>
<tr>
<td>Patients (outpatients vs in-and out patients)</td>
<td>-0.144</td>
<td>0.866 (0.317-2.36)</td>
<td></td>
</tr>
<tr>
<td>Indication (deep venous thrombosis vs pulmonary embolism)</td>
<td>-0.057</td>
<td>0.945 (0.401-2.22)</td>
<td></td>
</tr>
<tr>
<td>Test type (Latex vs Elisa)</td>
<td>-0.546</td>
<td>0.579 (0.275-1.22)</td>
<td></td>
</tr>
<tr>
<td>$r^2$</td>
<td>0.34</td>
<td>0.34</td>
<td>0.27</td>
</tr>
</tbody>
</table>

* model without covariates
Exploring sources of heterogeneity in systematic reviews of diagnostic tests

**Discussion**

We have shown several ways to 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 recognise that there are several causes of heterogeneity: statistical, methodological as well as clinical.

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 pre-specified, 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. This might improve in the near future as reporting guidelines are developed for the primary studies of diagnostic tests.

We feel that, 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 us 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 synthesise the available evidence.

References
Exploring sources of heterogeneity in systematic reviews of diagnostic tests

Chapter 7
