Methods Guide for Medical Test reviews

Paper 8

Meta-analysis of Test Performance Evidence When There is a "Gold Standard" Reference Standard

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

DRAFT

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

Effectiveness and Comparative Effectiveness Reviews, systematic reviews of existing research on the effectiveness, comparative effectiveness, and comparative harms of different medical tests, are intended to provide relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. In an effort to improve the transparency, consistency, and scientific rigor of the work of the Effective Health Care (EHC) Program, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and highlights particular areas where the inherently different qualities of medical tests necessitate a different or variation of the approach to systematic review compared to a review on treatments. We provide guidance in stepwise fashion for those conducting a systematic review.

The *Medical Test Methods Guide* is a living document, and will be updated as further empirical evidence develops and our understanding of better methods improves. Comments and suggestions on the *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

DRAFT

Paper 8. Meta-analysis of Test Performance Evidence When There is a "Gold Standard" Reference Standard

Meta-analysis and related methodologies are an important part of systematic reviews of medical tests. Quantitative analyses are not required for a systematic review; several very informative and well-conducted reviews do not include any such analyses. However, when possible, systematic reviewers appropriately strive to perform meaningful quantitative analyses to provide summary estimates for key quantities, or to explore and explain observed heterogeneity in the results of identified studies.

Syntheses of medical test data tend to focus on test performance; that is, on the ability to discern the presence of a particular condition or level of risk ("accuracy" studies). Thus, much of the attention on statistical issues relevant to synthesizing medical test evidence focuses on summarizing test performance data. It is important to be aware that key clinical questions driving an evidence synthesis (e.g., Is this test alone, or some test/treat strategy, likely to improve decisionmaking and patient outcomes?) are only indirectly related to test performance *per se*. Formulating an effective evaluation approach requires careful consideration of the context in which the test will be used. These framing issues are addressed in other papers in this *Medical Test Methods Guide* (*see* Papers 2, 6, and 10). For our purposes, we assume that the primary goal of synthesis is to combine and interpret test performance data.

It is beyond the scope of this *Medical Test Methods Guide* to provide a detailed description of statistical methodologies for meta-analysis of medical test performance and the assumptions they invoke, or to discuss the practical aspects of their implementation. In addition, it is expected that readers are versed in clinical research methodology and familiar with methodological issues pertinent to the study of medical tests, as well as with the common measures of medical test performance. For example, we do not review challenges posed by methodological or reporting shortcomings of test performance studies.[1] The Standards for Reporting of Diagnostic Accuracy (STARD) initiative published a 25-item checklist that aims to improve reporting of medical test studies.[2] We refer readers to Paper 5 of this *Medical Test Methods AnGuide* and to several methodological and empirical explorations that discuss the effects of bias and variation on the performance of medical tests.[3-5]

Common Challenges

Summarizing medical test performance data in general, or for a subgroup of patients, is complicated because medical test performance studies can differ on the definition of test positivity, may apply varying criteria for "truth," and often involve heterogeneous patient populations. This presents two common challenges.

First, the simplest challenge is summarizing studies of test performance when the reported sensitivities and specificities are not variable (non-heterogeneous) across studies. Non-heterogeneity can appear when there is no explicit or implicit variation in the threshold for a "positive" test result, as well as for other reasons.

DRAFT

A second challenge is summarizing test performance when there is substantial variation (heterogeneity) in sensitivity or specificity estimates across studies. This often-encountered situation can be secondary to explicit or implicit variation in the threshold for a "positive" test result; heterogeneity in populations, reference standards, or index tests; study design; chance; or bias. For example, when the test threshold changes, it is expected that sensitivity and specificity will also change, and in opposite directions.[a]

Principles for Addressing the Challenges

In this paper, we describe our approach to summarizing commonly encountered types of diagnostic data by addressing the two challenges mentioned above: (1) The case of non-heterogeneous sensitivity and specificity, which can be encountered when there is no variation in the threshold of positive tests; and (2) the case of heterogeneous sensitivity or specificity, which can appear when there is variation in the use and reporting of thresholds, among other things. We also discuss how to assess and explore heterogeneity using illustrative examples. We propose that reviewers follow two general principles.

Principle 1: Favor the Simplest Analysis That Properly Summarizes the Data

Meta-analysis of test performance data may require more complex methods than are needed for meta-analyses of intervention studies. This is because there are many metrics that describe test performance from different points of view, and one has to choose among these metrics. For most analyses, we resort to modeling sensitivity and specificity rather than other metrics, as will be discussed below. Furthermore, test performance is most often specified as a two-dimensional problem; that is, it consists of an analysis of sensitivity and specificity, or of quantities derived from them. A large number of analytical options are available to the meta-analyst, some more suitable to certain situations than others. We will discuss a general approach for choosing among the available options, which should of course be critically applied to the topic at hand.

Principle 2: Explore Any Variability in Study Results With Graphs and Suitable Analyses

One of the most important uses of meta-analysis is to quantify and explore reasons for between-study heterogeneity. Meta-regression-based analyses and subgroup analyses are naturally amenable to this purpose.

One of several possible approaches to meta-analysis of diagnostic data is presented in Box 8-1 and expanded in the section that follows.[6]

[a] Changes in the threshold of a test can be explicit when, for example, a cutoff value across a continuum of measurements is employed; or changes can be implicit when, for example, tests have a more qualitative interpretation, such as many imaging tests (e.g., chest radiography used for screening versus confirmation of tuberculosis). Notably, this is a frequent source of heterogeneity for test performance studies, and so the issue of coping with heterogeneity will be addressed here.

Box 8-1. Algorithm for meta-analysis of diagnostic studies when there is a "gold standard"

Step 1: Decide whether separate meta-analyses of sensitivity and specificity are suitable. This is true if there is no evidence of variability in the sensitivity or specificity across studies (e.g., by observing forest plots, or with standard heterogeneity assessment methods).[a] This can happen when all studies use the same explicit threshold for positive tests, but also in other cases.

Step 2: If there is substantial variation in sensitivity or specificity, consider methods that analyze them jointly. This can be the case when studies use different explicit or implicit thresholds for positive tests. Fit the bivariate model (see text). If there is evidence that the correlation between sensitivity and specificity is

a. *Negative** (i.e., as expected), present:
    i. Summary point estimates of sensitivity and specificity[a]
    ii. Summary receiver operating characteristic curve (SROC)
    iii. Both of the above
b. *Positive*** (i.e., in the opposite than expected direction), then for this collection of studies consider summarizing as per Step 1.
c. *Zero*, then the results of the bivariate model would be similar to the separate meta-analyses of Step 1.

Step 3: If more than one threshold is reported per study, this has to be taken into account in the quantitative analyses. Tentatively, we encourage both qualitative analysis via graphs and quantitative analyses via proper methods.

Step 4: Explore the impact of study characteristics on summary results in the context of the primary methodology used to summarize studies (Steps 1, 2, or 3):

a. Meta-regression-based analyses
b. Subgroup analyses

* Based on the posterior distribution of the correlation parameter, or perhaps based on other, external knowledge about the test.

** It is not uncommon to have a positive median or mean in the posterior distribution of the correlation. This is contrary to what is expected, and may be the result of confounding (omission of an important covariate).

Step 1: Decide whether separate meta-analyses of sensitivity and specificity are suitable. When there is no obvious variability in the sensitivity and specificity across studies, then there is no use for more advanced methods, which were specifically developed to take into account how sensitivity and specificity are related across their range. Among other cases, this situation can arise when all studies use the same explicit or implicit threshold for positive tests.

Most measures of test performance (with the exception of positive and negative predictive value) can be quantitatively combined using standard methodologies discussed in AHRQ's *General Methods Guide*.[7] Specifically, for the different measures:

Sensitivity and specificity. Sensitivity and specificity could be combined separately in the absence of variation in diagnostic thresholds when there is not substantial between-study heterogeneity. The naive approach of summing over all true positives (negatives) and diseased (non-diseased) subjects across studies and calculating "pooled" sensitivity (specificity) from the ratio of the corresponding sums should be avoided. This approach does not take into account between-study variability, or the potential bias caused by unbalanced ratios between diseased and non-diseased patients.[8-9]

Standard meta-analysis techniques using a fixed-effect or random-effects model could be used to obtain summary estimates by combining logit-transformed sensitivity and specificity and then transforming back to the original scale.[b] Such an approach assumes that the logits of sensitivity and specificity approximately follow a normal distribution. It is also possible to combine sensitivity and specificity using the number of test positives and negatives directly in a random-effects logistic model assuming a binomial distribution. There is evidence that a binomial approach performs better than the normal approximation.[10]

Positive and negative predictive values. As mentioned above, predictive values are dependent on prevalence. Because prevalence is often variable, and because many medical test studies have a case-control design (where prevalence cannot be estimated), it is rarely meaningful to combine predictive values across studies. Instead, the summary sensitivity and specificity can be used to calculate the "overall" positive and negative predictive values along a range of prevalence values. The calculations can be tabulated or plotted.

Positive and negative likelihood ratios. Positive and negative likelihood ratios could also be combined in the absence of threshold variation. Some methodological authors and medical journals give explicit guidance to combine likelihood ratios using standard fixed-effect or random-effects meta-analysis methods.[11] However, others note that this practice may result in problems under specific conditions.[12] For example, "summary" likelihood ratios calculated in this way may correspond to "summary" sensitivity and specificity values outside the 0 to 1 range.[c] Instead of summarizing likelihood ratios, these authors suggest to calculate the "summary" likelihood ratios from summary sensitivities and specificities obtained from bivariate analyses (as described in the following paragraphs). However, the actual numerical differences in the calculated quantities with alternate methods are generally small and do not result in different clinical conclusions.[13]

Diagnostic odds ratio. The synthesis of diagnostic odds ratios is straightforward and follows standard meta-analysis methods.[7,14] The diagnostic odds ratio is closely linked to sensitivity, specificity, and likelihood ratios, and can easily be included in meta-regression models to explore the impact of explanatory variables on between-study heterogeneity. In addition to challenges in interpreting diagnostic odds ratios, a disadvantage to using them is that they do not allow separate weighting of true positive and false positive rates.

DRAFT

[b] Note that neither of the above methods takes into account the correlation in sensitivity and specificity. In fact, this may not be a problem because in the absence of variation in thresholds, sensitivity and specificity may not be negatively correlated. However, even in this case, sensitivity and specificity are paired (they are still estimated from the same study), and this could be taken into account by modeling them jointly with a bivariate random effects model.[20] Although these models are developed with the goal of accounting for variation in threshold, they would equally apply to the situation of no variation in threshold. Generally, summary sensitivity and specificity estimated from separate random-effects meta-analyses of sensitivity and specificity are similar to those estimated from the bivariate model.[8] The advantage of the bivariate model is that a joint confidence region (instead of separate confidence interval) could be constructed for sensitivity and specificity to evaluate the two measures simultaneously and efficiently.

[c] This can happen not only when the correlation between LR+ and LR- is ignored, but also when one analyzes LR+ and LR- jointly with bivariate models.[12]

In such a situation:

- Reviewers can combine sensitivity and specificity in separate meta-analyses (preferably with a binomial method). However, it is also possible to combine them in a bivariate model.[d]
- It is recommended to back-calculate positive and negative predictive values from summary estimates of sensitivity and specificity (obtained with the bivariate method), rather than meta-analyzing them directly (in separate or joint models).
- It may be preferable to back-calculate "summary" positive and negative likelihood ratios from summary estimates of sensitivity and specificity (obtained with the bivariate method [see below]), rather than meta-analyzing them directly (in separate or joint models).
- One can summarize diagnostic odds ratios using standard meta-analysis methods.
- Generally, we recommend using a random-effects model as discussed in the *General Methods Guide*.[7]

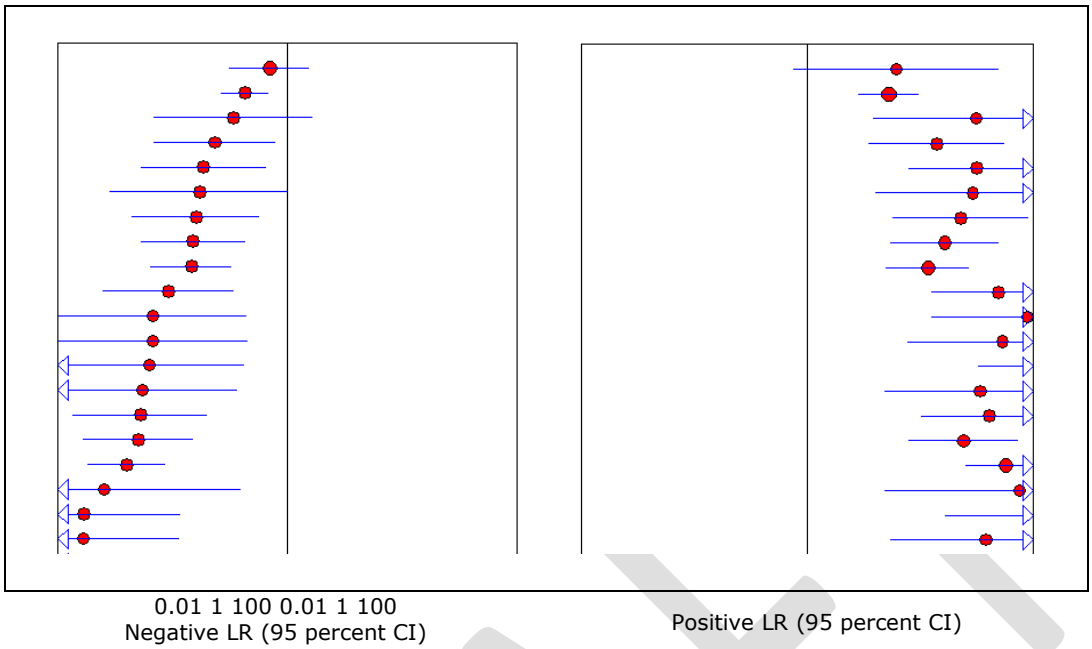
Step 2: Consider joint analysis of sensitivity and specificity when each study reports a single threshold. When there is obvious variability in the sensitivity and specificity across studies, then there is no use for more advanced methods, which were specifically developed to take into account how sensitivity and specificity are related across their range. Among other cases, this can happen when all studies use the same explicit or implicit threshold for positive tests.

In the most common scenario, there is obvious variability in the sensitivity and specificity across studies. This variability may be secondary to different explicit or implicit thresholds across studies for positive tests, composition of populations, heterogeneity in the reference standard or the index test, biases, or chance. For example, one cannot be sure that there is no variation in thresholds (e.g., for qualitative tests, such as many imaging tests). If elevated values indicate disease, a high threshold leads to low sensitivity and high specificity, while a low threshold produces high sensitivity and low specificity. This variation in threshold induces a negative correlation between sensitivity and specificity. Combining sensitivity and specificity separately (ignoring the negative correlation between them) would generally underestimate test accuracy given variation in threshold.[9] The preferred method is to model sensitivity and specificity simultaneously, accounting for tradeoff variations and for the correlation between sensitivity and specificity.

Graphing paired sets of accuracy results (sensitivity/specificity or likelihood ratios) using a forest plot, ideally in descending or ascending order of accuracy, may help illustrate the extent of variability across studies (Figure 8-1).

[d] If there is no variability in sensitivity and specificity across studies, their correlation is close to zero, and the results of the bivariate model should be similar to those from separate meta-analyses.

Figure 8-1. Side-by-side forest plots of "paired" negative and positive likelihood ratios

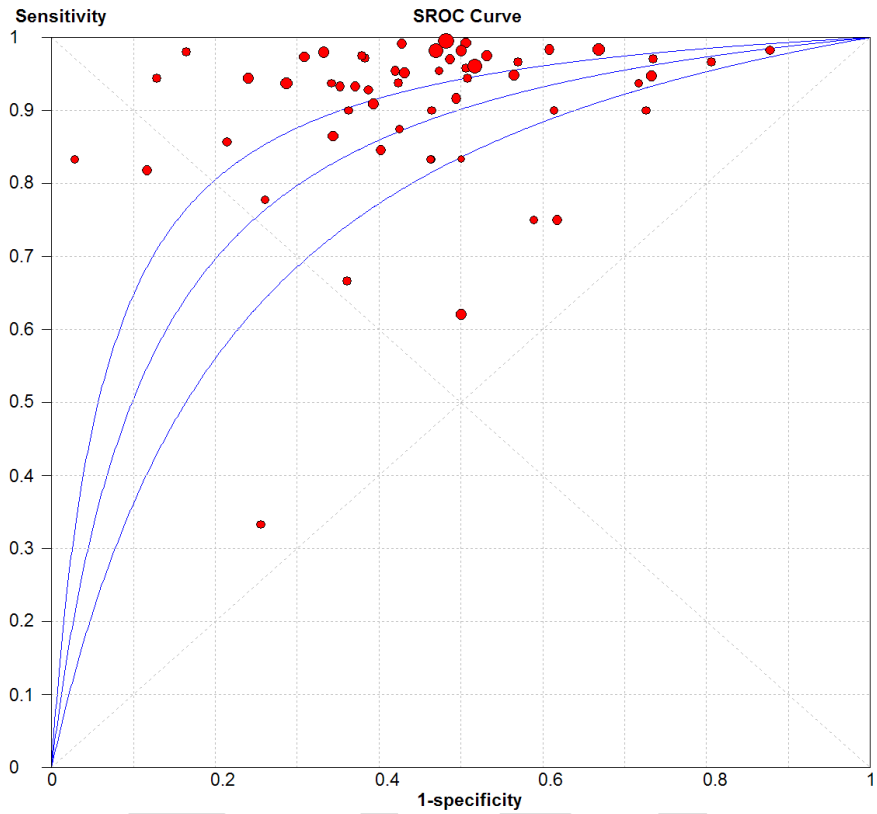


CI = confidence interval; LR = likelihood ratio. The plot represents the twenty diagnostic studies from the meta-analysis by Gupta and colleagues[32] on the ability of endometrial thickness measured by pelvic ultrasonography for diagnosing endometrial cancer. Studies shown in descending order of LR- (left panel) along with with the corresponding LR+ (right panel). Note the inverse relationship between positive and negative LRs, which is suggestive of a threshold effect.

In addition, plotting accuracy studies on a summary receiver operating characteristic (SROC) curve plane, with sensitivity on the vertical axis and 1–specificity on the horizontal, is expected to result in a "shoulder arm" pattern when a threshold effect is present. Figure 8-2 shows a typical example.

DRAFT

Figure 8-2. SROC curve with a "shoulder arm" pattern



SROC = summary receiver operating characteristics. SROC curve includes all studies (n = 57) included in the meta-analysis by Gupta and colleagues[32] evaluating the ability of endometrial thickness measured by pelvic ultrasonography for diagnosing endometrial cancer. The "shoulder arm" appearance suggests the presence of a "threshold effect."

The common case is when each study reports *a single pair of sensitivity and specificity at a given threshold* (with different studies reporting different thresholds). Several statistical models have been developed to summarize medical test performance for this case. Another, more complex situation arises when *multiple sensitivity and specificity pairs (at different thresholds)* are reported in each study. Statistical models for the latter case exist, but there is less empirical evidence on their use. We discuss both cases, emphasizing the former. Also, we discuss the synthesis of sensitivity and specificity and recommend that other measures be back-calculated from summary sensitivity and specificity estimates, as when combining studies that use the same explicit threshold.

To model the relationship between specificity and sensitivity, it is recommended always to plot the data (sensitivity versus 1–specificity) to visually assess the relationship between the two measures first. The early development of the various models has generally followed the idea to construct an SROC curve. More comprehensive approaches (random intercept, hierarchical SROC [HSROC], and bivariate sensitivity-specificity models) have been developed to address many limitations of the early models. Below is a brief comment on the strengths and limitations of the aforementioned approaches. The details of the parameterization of these models are not reviewed here.

SROC models. The most used SROC model[e] was developed by Littenberg and Moses[15-16] and attempts to explicitly model the testing threshold effect. It is essentially a regression (weighted or unweighted variations exist) of the difference of the logit-transformed true and false positive rates (sensitivity and 1–specificity, respectively) versus their sum. The back-transformed regression line, plotted in the sensitivity versus 1–specificity space, is the SROC. When there is no threshold effect, the model can also provide an estimate of the summary diagnostic odds ratio. The model is easy to implement and can be performed in most statistical packages; thus, it has been routinely used to summarize diagnostic accuracy. However, this model has major limitations, including:

- It is a fixed-effect model, which ignores unexplained variation among studies.
- It does not account for the correlation between sensitivity and specificity.
- It does not account for the variation associated with the independent variable in the model.
- It does not weight studies optimally in estimation, and therefore inferences on the effects of covariates are wrong (i.e., this model should not be used to explore effects of covariates).
- It uses an arbitrary continuity correction of 0.5 when there are empty cells in the 2-by-2 matrix.

Random intercept models. It is straightforward to extend the Littenberg and Moses model to a random-effects model, which is parameterized as a random intercept model. This model is an improvement over the Littenberg and Moses model and has been used in combining medical tests.[17-18] The random intercept model can incorporate study-level covariates. However, it does not address all the limitations listed above. It does not take into account the correlation between sensitivity and specificity and the variation associated with the independent variable in the model.

Hamza et al.[10] compared the bias and confidence interval coverage probabilities of the random intercept model and the bivariate model. So far, there is no study comparing the random intercept model and the hierarchical SROC (HSROC) model. In the Hamza et al. study,[10] the random intercept model had better coverage probabilities than the bivariate model when there are few studies (fewer than 10). However, it yields biased parameter estimates. For more than 10 studies, the bivariate model yields less biased estimates and has higher coverage probabilities.[f]

HSROC and bivariate models. The HSROC model and the closely related bivariate model address all the limitations of the simple SROC approach. Recently, Harbord et al.[8] compared the performance of the HSROC/bivariate model to several simpler models[g] and concluded that

[e] The SROC model was first proposed by Kardaun and Kardaun,[33] but it was not straightforward to implement and did not gain wide use.

[f] Hamza et al.[10] compared in simulations the performance of the bivariate random effects models with the aforementioned random intercept model estimating summary log diagnostic odds ratio examining bias, mean squared error, and coverage probabilities. When the number of studies (N) $\geq$ 10, the model using binomial distribution is always preferred. When the number of studies = 10, the model using binomial distribution provides a better unbiased estimate, but the coverage probability and mean squared error is often outperformed by the random intercept model.

[g] The random intercept model was not examined in this study.

HSROC/bivariate models are necessary to address variation in threshold. The HSROC and bivariate models can incorporate study-level covariates.

Rutter and Gatsonis[19] describe the HSROC model that is motivated by an ordinal regression approach and is constructed in terms of positivity threshold and accuracy parameters. The numbers of positive tests are evaluated using a binomial distribution, and a Bayesian approach is used to obtain model parameter estimates through Markov chain Monte Carlo (MCMC).[h] The Bayesian approach provides a flexible modeling and estimating framework but at the price of more programming, simulation, evaluation of model adequacy, and synthesis of simulation results.[19]

Reitsma et al.[20] and Arends et al.[21] advocate the approach of bivariate random effects model to analyze sensitivity and specificity jointly. These models preserve the two-dimensional nature of the data (sensitivity/specificity pairs), produce summary estimates of sensitivity and specificity, and incorporate any possible (negative) correlation between the two measures.[i]

Harbord et al.[22] showed that the HSROC model and the bivariate random-effects model are very closely related, and in the absence of study-level covariates, they are different parameterizations of the same model. The different parameterizations may reflect a difference in selecting a summary measure for medical test accuracy. Namely, the HSROC model naturally leads to an SROC curve when there is a threshold effect but little heterogeneity in the accuracy parameter; and the bivariate model parameterization naturally produces a summary sensitivity and specificity, together with a joint confidence or prediction region for both together. However, all other commonly used measures for medical test accuracy could readily be obtained from each model with proper calculations. Rutter and Gatsonis[19] provided formulas to estimate summary sensitivity, specificity, and likelihood ratios; several types of SROC curves could be produced from the bivariate model based on regressing logit sensitivity on logit specificity, regressing logit specificity on logit sensitivity, or using other options,[21] in addition to likelihood ratios and diagnostic odds ratios.[20]

Chappell et al.[6] discuss that if there is substantial evidence that the correlation between sensitivity and specificity is *positive* (opposite from what is theoretically expected), then an SROC curve will be meaningless, and perhaps an important covariate has not been taken into account.[j] In that event, they suggest that one performs independent analyses of sensitivity and specificity to try to identify a likely explanation.

---

[h] Macaskill[34] implemented the same model in a classical framework using the SAS® procedure NLMIXED and produced comparable results without the degree of complexity inherent in the MCMC simulation.

[i] The bivariate random effects model can be parameterized assuming that the logits of sensitivity and specificity follow a bivariate normal distribution,[20,21] alternatively it can model the number of positive tests using the binomial distribution.[21,35] One disadvantage of the former model is that an arbitrary correction of 0.5 is required to avoid undefined log odds when the data are sparse with zero positive tests. The model using binomial distribution generally performs better than the model using normal distribution for logits of sensitivity and specificity, though the difference is not always practically relevant when the study sample sizes are large.[10,21]

[j] Chappell et al.[6] point out that they have still to encounter an example where the correlation between sensitivity and specificity is, e.g., positive with probability >0.95. However, they comment that it is not uncommon for the median or mean of the posterior density of this correlation to be positive. In that case, the model is probably not well estimated.

Choice of models. We make the following recommendations about selecting the appropriate model:

1) There is adequate theoretical motivation to discourage the use of the Littenberg and Moses model to draw inferences (compare subgroups of studies or perform direct or indirect comparisons between two or more index tests).

2) We do not recommend the routine use of the random intercept model.

3) Based on currently available empirical evidence, we encourage the use of the HSROC model or the bivariate model (preferably using the binomial distribution). These models are quite complex. Nevertheless, the complexity of the diagnostic data demands such models.[k]

4) When there is no covariate in the model, the HSROC and bivariate models are equivalent. The primary interests of inference may help guide the choice of model. That is, if the investigators are mostly interested in an SROC curve (or the research question makes an SROC curve a better choice), then the HSROC model is more convenient to use. If the research question makes a summary sensitivity or specificity the most appropriate measure, the bivariate model would be more convenient. For example, if the reported sensitivity and specificity has a wide range of values, an SROC would be appropriate. On the other hand, if the reported sensitivity and specificity are very similar to each other and located only in a small portion of the ROC space, then it would not be appropriate to extrapolate and construct a full ROC curve, and summary sensitivity or specificity are better choices by using the bivariate model. Harbord et al.[22] also explained how the difference in design and conduct of the included diagnostic accuracy studies may affect the choice of the model. For example, "spectrum effects," where the subjects included in a study are not representative of the patients who will receive the test in practice,[23] "might be expected to affect test accuracy rather than threshold, and might therefore be most appropriately investigated using the HSROC approach. Conversely, between-study variation in disease severity will (likely) affect sensitivity but not specificity, leading to a preference for the bivariate approach."[22] Investigators are encouraged to look at study characteristics and evaluate how these study characteristics could affect the diagnostic accuracy, which in turn might affect the choice of model. Further research is also needed to evaluate how study characteristics are associated with the performance of these models.

5) When there are covariates in the model, the HSROC model allows direct evaluation of the difference in accuracy or threshold parameters or both. Bivariate models, on the other hand, allow for direct evaluation of the difference in sensitivity or specificity or both. In addition, the HSROC model can be more easily extended to

[k] If the systematic reviewers opt not to use a Bayesian approach, they could use the SAS NLMIXED procedure or the Stata® *xtmelogit* or *gllamm* commands. Note that the proper syntax for these commands can become complicated in that subtle variations may result in fitting a different model.

include a covariate to affect the degree of asymmetry of the SROC curve.[22] Investigators are encouraged to consider these factors in model choice.

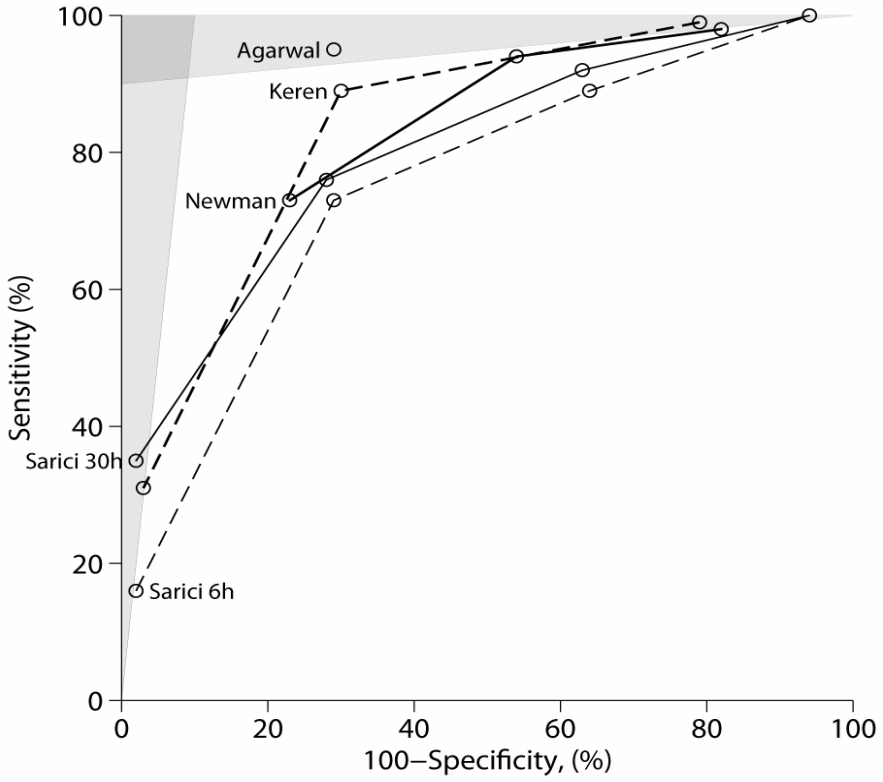
Step 3: Consider joint analysis of sensitivity and specificity when studies report multiple thresholds. It is not uncommon for some studies to report multiple sensitivity/specificity pairs. One option is to decide on a single threshold from each study and apply the aforementioned methods. To some extent, the setting in which the test is used can guide the selection of the threshold. For example, in some cases, the threshold that gives the highest sensitivity may be appropriate in medical tests to rule out disease. Another option is to use all available thresholds per study, as detailed below.

Statistical models. An extension of the HSROC model has been developed to analyze sensitivity and specificity data reported at more than one threshold.[24] This model explicitly uses latent variables and is more challenging to implement than the HSROC and bivariate models discussed above. Related models have been recently proposed in the literature.[25] Further, if each study reports enough data on sensitivity and specificity to construct a ROC curve, Kester and Buntinx[26] have proposed a so far little-used method to combine whole ROC curves.

Choice of models. Both models are theoretically motivated. The Dukic and Gatsonis model[24] is more elaborate and more technical in its implementation than the Kester and Buntinx variant.[26] There is no empirical evidence on the performance of either model in a large number of applied examples. Therefore, we refrain from providing a strong recommendation always to perform such analyses.[1] At a minimum, we suggest that systematic reviewers perform explorations in a qualitative, graphical depiction of the data in the ROC space. This will provide a qualitative summary and highlight similarities and differences among the studies. An example of such a graph is in Figure 8-3, which illustrates the diagnostic ability of early measurements of total serum bilirubin (TSB) to identify postdischarge TSB above the 95th 10-hour-specific percentile.[27]

[1] Systematic reviewers are encouraged to perform explorations, including analyses, with these models. Should they opt to do so, they should provide adequate description of the employed models and their assumptions as well as a clear intuitive interpretation of the parameters of interest in the models.

Figure 8-3. Diagnostic ability of early TSB measurements



TSB = total serum bilirubin. Ability of early TSB measurements to identify postdischarge TSB above the 95 percent 10-hour-specific percentile. Sensitivity/100 percent minus specificity pairs from the same study (obtained with different cutoffs for the early TSB measurement) are connected with lines. These lines are reconstructed based on the reported cutoffs and are not perfect representations of the actual ROC curves in each study (they lack granularity). Studies listed on the left shaded area have an LR+ of at least 10. Studies listed on the top shaded area have an LR- of at most 0.1. Studies listed at the intersection of the gray areas (darker gray polygon) have both LR+ of at least 10 and LR- of 0.1 or less.[27]

Step 4. Explore between-study heterogeneity. Heterogeneity is common in systematic reviews of medical tests due to differences in threshold or cutoff values used by different studies of an index test to define the presence of the target condition. The difference in threshold can be explicit; for example, studies that use different numerical cutoff values of blood glucose level to determine the presence of diabetes. Or the difference in threshold can be implicit, as when decision on a positive test depends on subjective judgment and interpretation; for example, the abnormality of radiographs. This "threshold effect" is valid only if all the studies use an equivalent index test; that is, in the case of a blood test, one manufacturer's test method gives the same values as another manufacturer's test method for the same sample. In practice, this is not often the case, and special attention to method differences is required before heterogeneity due to a threshold effect can be investigated.

Exploring heterogeneity using hierarchical models. Other than accounting for the presence of a threshold effect, HSROC and bivariate models provide flexible ways to test and explore between-study heterogeneity simultaneously. For either model, the formulation includes a variance parameter to test the presence and magnitude of between-study heterogeneity. Both models allow for inclusion of study characteristics to explore how study characteristics could

explain sources of heterogeneity across studies and allow for exploration in two measures at once. As mentioned above, an HSROC model allows for direct evaluation of the heterogeneity in accuracy or threshold parameters or both, and study characteristics for each parameter may be the same or different. Bivariate models, on the other hand, allow for direct evaluation of heterogeneity in sensitivity or specificity or both, and again, study characteristics for each measure could be different. After inclusion of study characteristics, the estimate of variance parameter provides an estimate of the residual variance and helps evaluate the amount of heterogeneity explained by these characteristics by comparing the estimate of variance parameter before including the study characteristics. Factors reflecting differences in patient population and method in patient selection, methods of verification and interpretation of results, clinical setting, disease severity, etc., are common considerations for the source of heterogeneity. Investigators are encouraged to use these models to test and explore heterogeneity, especially when they have chosen these methods for combining studies.

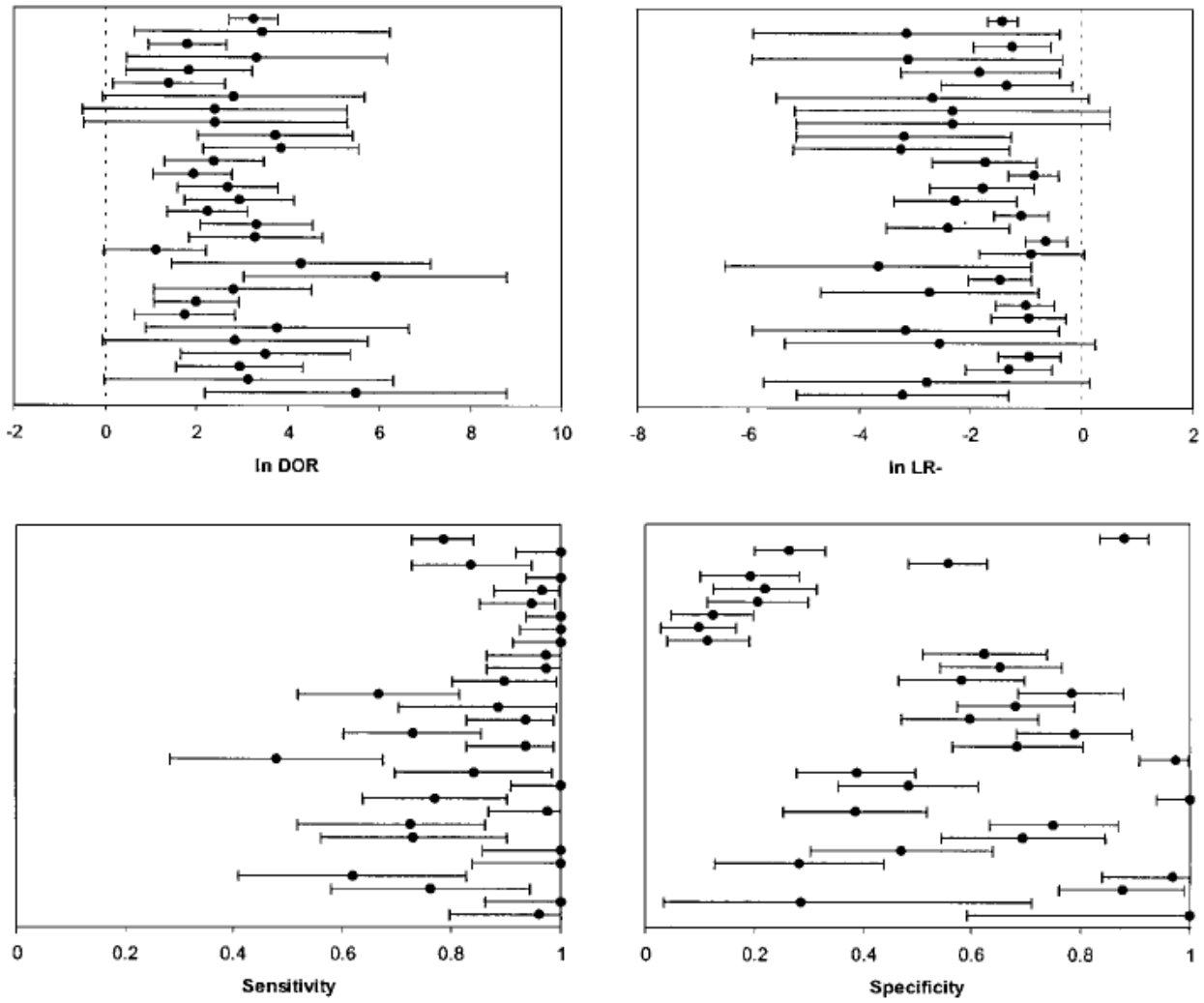
Illustrations

We briefly demonstrate the above with two applied examples. Example 1 is on D-dimer assays for the diagnosis of venous thromboembolism[28] and shows heterogeneity due to "threshold effect"; it has been discussed by Lijmer et al.[29] Example 2 is from an EPC evidence report on the ability of serial creatine kinase-MB measurements to diagnosis acute cardiac ischemia[30-31] and shows heterogeneity for another reason.

Example 1: D-dimers for diagnosis of venous thromboembolism. D-dimers are fragments specific to fibrin degradation in blood or plasma and can be used to diagnose venous thromboembolism. Figure 8-4 presents forest plots of the (log-transformed) diagnostic odds ratio, negative likelihood ratio,[m] sensitivity, and specificity for the D-dimers example.[29] As discussed above, one could, in theory, synthesize each of these measures. Note that sensitivity and specificity are more heterogeneous than the diagnostic odds ratio or the negative likelihood ratio (this can be verified by formal testing for heterogeneity). This may be due to threshold variation in these studies (from 25 to 550 ng/mL, when stated; Figure 8-5), or due to other reasons.[29]

[m] This is a diagnostic test aiming to identify the majority of patients with venous thromboembolism. Therefore, the negative likelihood ratio is arguably quite important to note.

Figure 8-4. 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



In Figure 8-4,[29] points indicate estimates; horizontal lines are 95 percent confidence intervals for the estimates. Such plots help to appreciate the extent of heterogeneity in different measures of medical test performance. Heterogeneity is much more pronounced for sensitivity and specificity rather than for the diagnostic odds ratio (log-transformed, ln DOR) or the negative likelihood ratio (lon transformed, ln LR-). As discussed in the previous section, these measures express different aspects of test performance, and one could synthesize any of them. Recall that the summary diagnostic odds ratio gives an overall indication of ability of the D-dimers to distinguish venous thromboembolism from other conditions, but it does not inform on whether it performs better in people with the disease or in people without the disease. As noted in the text above, pooling likelihood ratios may result in "summary" likelihood ratios that do not correspond to meaningful "summary" sensitivity/specificity estimates. In this example, there is large variability in the diagnostic threshold used in the different studies. Therefore, univariate (separate) syntheses of sensitivity and specificity are probably not meaningful. Heterogeneity in the thresholds is one of the potential explanations for the observed between-study heterogeneity in this example.

Figure 8-5. Variation in threshold in the D-dimer example

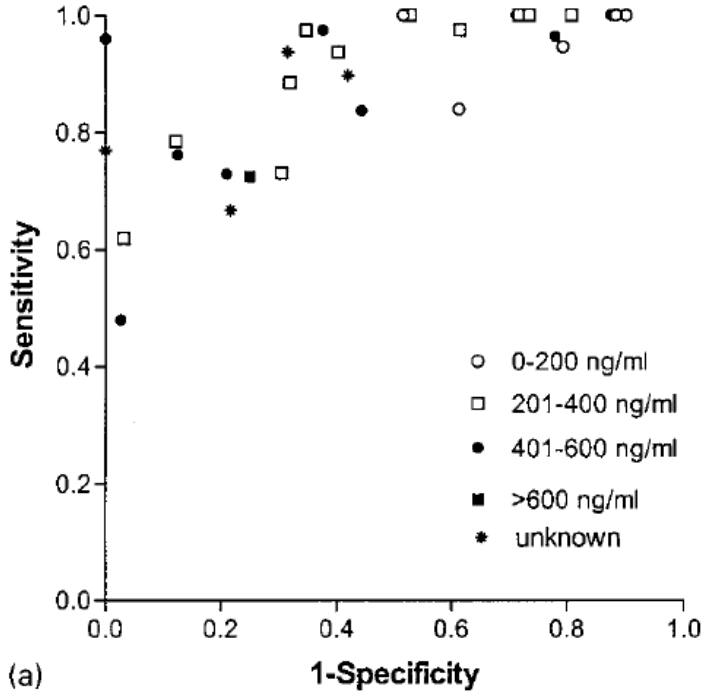


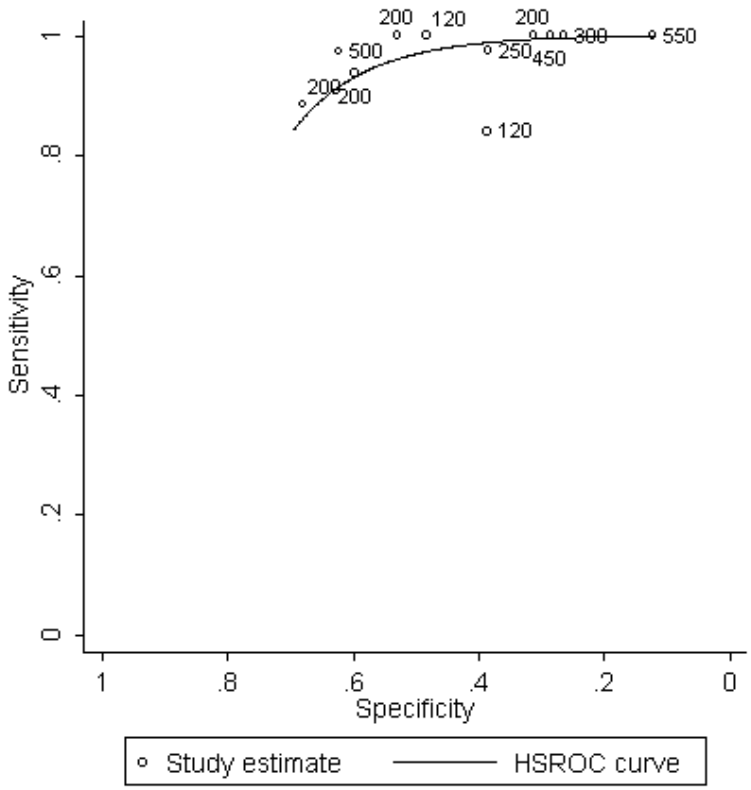
Figure 8-5 is from Lijmer et al.,[29] with data from Becker et al.[28] Shown is a sensitivity/1–specificity plot for the D-dimers example. Different markers denote different thresholds for the test. The "shoulder" image that the distribution of studies follows is seen in many meta-analyses where there is threshold-induced heterogeneity.

Since there is evident variation in the thresholds for studies of D-dimers, it is more appropriate to summarize the performance of the test using an HSROC curve rather than provide summary sensitivities and specificities.[n] The HSROC describes the tradeoff between sensitivity and specificity in the various studies. Shown below is the HSROC curve for the ELISA-based D-dimer tests in 11 studies (Figure 8-6). (For simplicity, we selected the highest threshold from two studies that reported multiple ELISA thresholds.)

Interpreting the HSROC curve (or the summary estimates) in this case is straightforward. This test has very good diagnostic ability, and it appropriately focuses on minimizing false negative diagnoses. It is also informative to "summary" negative (or positive) predictive values for this test. As described previously, we can calulate them based on the summary sensitivity and specificity estimates and over a range of plausible values for the prevalence**.** Figure 8-7 shows such an example using the summary sensitivity and specificity of the 11 studies of Figure 8-6.

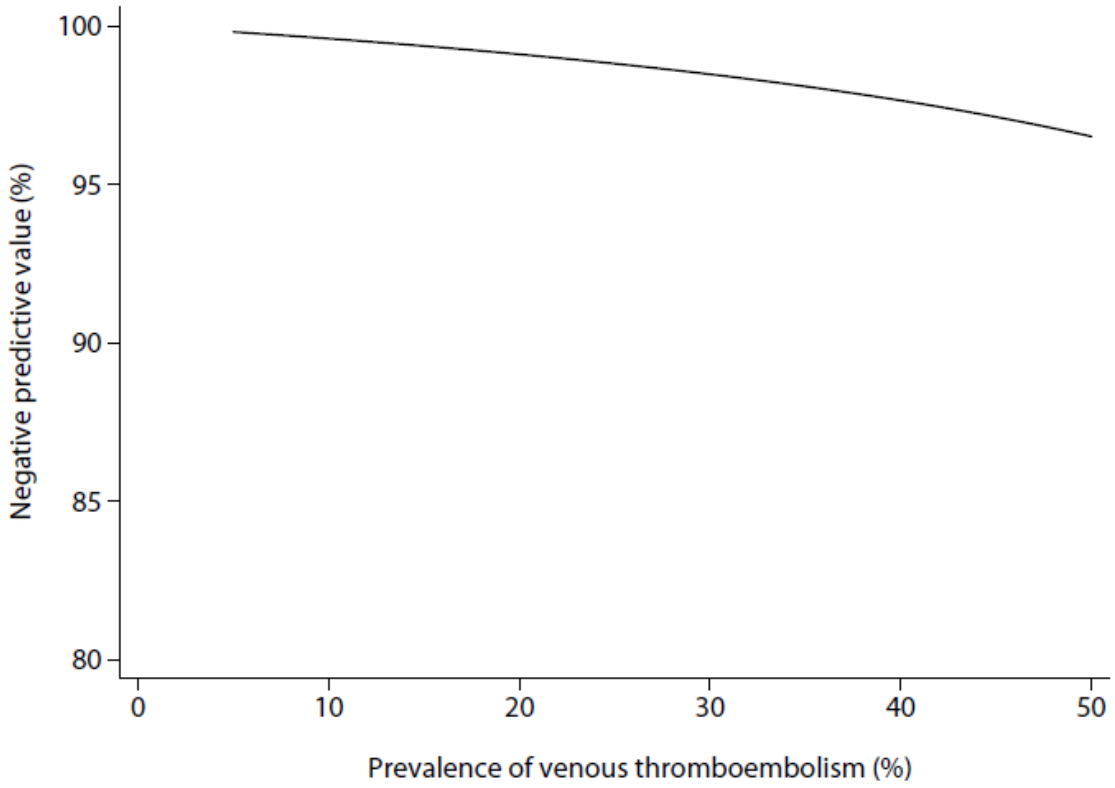
[n] Because sensitivity and specificity change across thresholds, it is unclear what a "summary" sensitivity and specificity would mean.

Figure 8-6. HSROC for the ELISA-based D-dimer tests



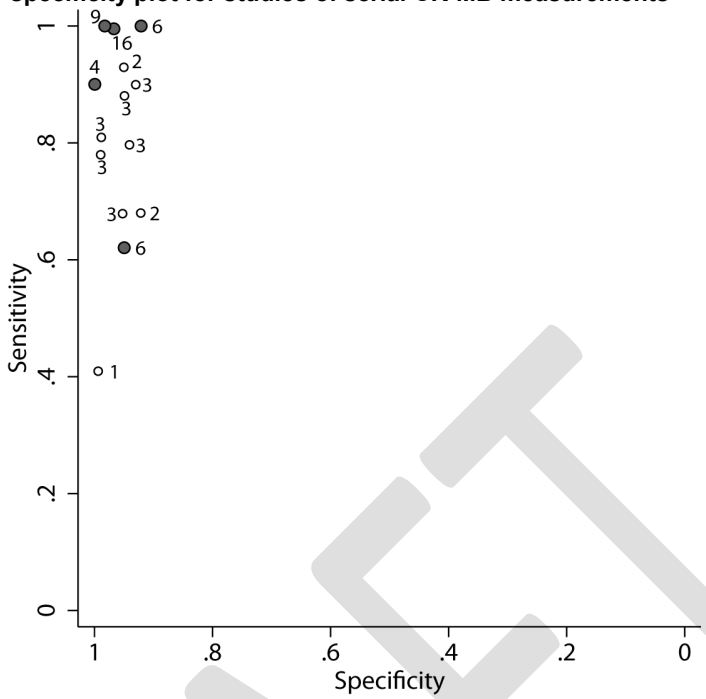
HSROC = hierarchical summary receiver-operator curve. Note that the horizontal axis is essentially 1–specificity (because it is reversed). Also shown is the threshold (in ng/mL) that was used in each study. This analysis used the *metandi* module for Stata (Roger Harbord's wrapper command for the more general *xtmelogit* and *gllamm* commands in Stata). This analysis is not included in the methodological paper by Lijmer et al.[29]

Figure 8-7. Calculated negative predictive value for the ELISA-based D-dimer test if the sensitivity and specificity are fixed at a summary value (x and y, respectively) and prevalence of venous thromboembolism is between 5 and 50 percent



Example 2: Serial creatine kinase-MB measurements for diagnosing acute cardiac ischemia. An evidence report examined the ability of serial creatine kinase-MB (CK-MB) measurements to diagnose acute cardiac ischemia in the emergency department.[30-31] Figure 8-8 shows the 14 eligible studies along with how many hours after symptom onset the last measurement was taken. It is evident that there is heterogeneity in the sensitivity, and that sensitivity increases with longer time from symptom onset.

Figure 8-8. Sensitivity 1–specificity plot for studies of serial CK-MB measurements



Note that since it is reversed, the horizontal axis effectively represents 1–specificity as usual for an ROC plot. Numbers are the number of hours after symptom onset that the last CK-MB measurement was taken. Filled circles are studies with a last measurement taken more than 3 hours after symptom onset.

For illustrative purposes, we compare the summary sensitivity and specificity of studies where the last measurement was performed within 3 hours of symptom onset versus the remaining studies(Table 8-1).[o] We use a bivariate multilevel model with exact binomial likelihoods, as discussed above. In the fixed-effect part of the model, we include an indicator variable that codes whether the last measurement was earlier than 3 hours from symptom onset. We allow this variable to have different effects on the summary sensitivity and specificity.[p] This is essentially a bivariate meta-regression.

Table 8-1. Comparison of diagnostic performance of studies according to timing of the last serial CK-MB measurement for diagnosis of acute cardiac ischemia

	≤ 3 hours	> 3 hours	P value for the comparison across subgroups
Summary sensitivity (percent)	80 (64 to 90)	96 (85 to 99)	0.036
Summary specificity (percent)	97 (94 to 98)	97 (95 to 99)	0.56

The corresponding meta-regression can be specified for the HSROC model, in which case the results will be expressed with respect to the accuracy and threshold parameters (more difficult to interpret for a clinical audience).

[o] This analysis was not performed in the evidence report. It is performed here for illustration.

[p] Analyses are performed in Intercooled Stata 10, using the *xtmelogit* command. The same results are obtained with the *gllamm* command in earlier versions of Stata. The same model can be run in other statistical languages. Care has to be taken to specify the model correctly.

Note that properly specified bivariate meta-regressions (or HSROC-based meta-regressions) can be used to compare two or more index tests. The specification of the meta-regression models will be different when the comparison is indirect (different index tests are examined in independent studies) or direct (the different index tests are applied in the same patients in each study).

Notably, the evidence report described above included systematic reviews of several technologies (tests), including the aforementioned example. To interpret and contextualize the findings of these systematic reviews, the researchers used decision modeling analyses that compared 17 technologies and 4 test combinations.[31]

Summary

Key points are:

- Summarizing the performance of diagnostic tests is a multidimensional problem, as there are several quantities of interest that are, in principle, correlated and contribute complementary information.
- Separate meta-analyses of sensitivity and specificity are suitable if there is no evidence of variability in the sensitivity or specificity across studies. This can happen when all studies use the same explicit threshold for positive tests, but also in other cases.
- If there is substantial variation in sensitivity or specificity, one should consider methods that analyze these quantities jointly. The bivariate meta-analysis of sensitvity and specificity and the closely related HSROC method are theoretically motivated.
- If more than one threshold is reported per study, this has to be taken into account in the quantitative analyses. Tentatively, we encourage both qualitative analysis via graphs and quantitative analyses via appropriate methods.
- Reviewers should explore the impact of study characteristics on summary results in the context of the primary methodology used to summarize studies using meta-regression-based analyses or subgroup analyses.

References

1. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138(1):W1-12.

2. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003;138(1):40-4.

3. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282(11):1061-6.

4. Rutjes AW, Reitsma JB, Di NM, et al. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174(4):469-76.

DRAFT

5. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140(3):189-202.

6. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med 2009;28(21):2653-68.

7. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed September 20, 2010.

8. Harbord RM, Whiting P, Sterne JA, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol 2008;61(11):1095-1103.

9. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120(8):667-76.

10. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. J Clin Epidemiol 2008;61(1):41-51.

11. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004;329(7458):168-9.

12. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med 2008;27(5):687-97.

13. Simel DL, Bossuyt PM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. J Clin Epidemiol 2009;62(12):1292-300.

14. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56(11):1129-35.

15. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13(4):313-21.

16. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12(14):1293-316.

17. Oei EH, Nikken JJ, Verstijnen AC, et al. MR imaging of the menisci and cruciate ligaments: a systematic review. Radiology 2003;226(3):837-48.

18. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US--a meta-analysis. Radiology 2000;216(1):67-77.

DRAFT

19. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol 1995;2 Suppl 1:S48-S56.

20. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58(10):982-90.

21. Arends LR, Hamza TH, van Houwelingen JC, et al. Bivariate random effects meta-analysis of ROC curves. Med Decis Making 2008;28(5):621-38.

22. Harbord RM, Deeks JJ, Egger M, et al. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007;8(2):239-51.

23. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002;137(7):598-602.

24. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics 2003;59(4):936-46.

25. Hamza TH, Arends LR, van Houwelingen HC, et al. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. BMC Med Res Methodol 2009;9:73.

26. Kester AD, Buntinx F. Meta-analysis of ROC curves. Med Decis Making 2000;20(4):430-9.

27. Trikalinos TA, Ip S, Raman G, et al. Home diagnosis of obstructive sleep apnea-hypopnea syndrome. Technology Assessment. Rockville, MD: Agency for Healthcare Research and Quality. August 2007. Available at: http://www.cms.hhs.gov/determinationprocess/downloads/id48TA.pdf. Accessed July 9, 2010.

28. Becker DM, Philbrick JT, Bachhuber TL, et al. D-dimer testing and acute venous thromboembolism. A shortcut to accurate diagnosis? Arch Intern Med 1996;156(9):939-46.

29. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002;21(11):1525-37.

30. Balk EM, Ioannidis JP, Salem D, et al. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. Ann Emerg Med 2001;37(5):478-94.

31. Lau J, Ioannidis J, Balk E, et al. Evaluation of Technologies for Identifying Acute Cardiac Ischemia in Emergency Departments. Evidence Report/Technology Assessment Number 26. (Prepared by The New England Medical Center Evidence-based Practice Center under Contract No. 290-97-0019) AHRQ Publication No. 01-E006, Rockville, MD: Agency for Healthcare Research and Quality. May 2001. Available at:

DRAFT

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.37233. Accessed July 9, 2010.

32. Gupta JK, Chien PF, Voit D, Clark TJ, Khan KS. Ultrasonographic endometrial thickness for diagnosing endometrial pathology in women with postmenopausal bleeding: a meta-analysis. Acta Obstet Gynecol Scand. 2002 Sep;81(9):799-816.

33. Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf Med 1990;29(1):12-22.

34. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J Clin Epidemiol 2004;57(9):925-32.

35. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol 2006;59(12):1331-2.

DRAFT