Chapter 8: Meta-analysis of Test Performance When There is a "Gold Standard"

Thomas A. Trikalinos, MD[1,2], Cynthia M. Balion, PhD[3,4], Craig I. Coleman, PharmD[5,6], Lauren Griffith, PhD[4], Pasqualina L. Santaguida, BScPT, PhD[4], Ben Vandermeer, MSc[7], and Rongwei Fu, PhD[8]

[1]Center for Evidence-based Medicine, and Department of Health Services Policy and Practice, Brown University, Providence, RI, USA; [2]Tufts Evidence-based Practice Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA; [3]Department of Pathology and Molecular Medicine, Hamilton General Hospital, Hamilton, Ontario, Canada; [4]McMaster Evidence-based Practice Center, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada; [5]University of Connecticut/Hartford Hospital Evidence-based Practice Center (EPC), Storrs, CT, USA; [6]University of Connecticut School of Pharmacy, Storrs, CT, USA; [7]University of Alberta Evidence-based Practice Center, Edmonton, Alberta, Canada; [8]Department of Public Health and Preventive Medicine, Oregon Evidence-based Practice Center, Oregon Health and Science University, Portland, OR, USA.

Synthesizing information on test performance metrics such as sensitivity, specificity, predictive values and likelihood ratios is often an important part of a systematic review of a medical test. Because many metrics of test performance are of interest, the meta-analysis of medical tests is more complex than the meta-analysis of interventions or associations. Sometimes, a helpful way to summarize medical test studies is to provide a "summary point", a summary sensitivity and a summary specificity. Other times, when the sensitivity or specificity estimates vary widely or when the test threshold varies, it is more helpful to synthesize data using a "summary line" that describes how the average sensitivity changes with the average specificity. Choosing the most helpful summary is subjective, and in some cases both summaries provide meaningful and complementary information. Because sensitivity and specificity are not independent across studies, the meta-analysis of medical tests is fundamentaly a multivariate problem, and should be addressed with multivariate methods. More complex analyses are needed if studies report results at multiple thresholds for positive tests. At the same time, quantitative analyses are used to explore and explain any observed dissimilarity (heterogeneity) in the results of the examined studies. This can be performed in the context of proper (multivariate) meta-regressions.

KEY WORDS: gold standard; test performance; meta-analysis.
J Gen Intern Med 27(Suppl 1):S56–66
DOI: 10.1007/s11606-012-2029-1

INTRODUCTION

The series of papers in this supplement of the journal highlights common challenges in systematic reviews of medical tests and outlines their mitigation, as perceived by researchers partaking in the Agency for Healthcare Research and Quality (AHRQ) Effective Healthcare Program. Generic by their very nature, these challenges and their discussion apply to the larger set of systematic reviews of medical tests, and are not specific to AHRQ's program.

This paper focuses on choosing strategies for meta-analysis of test "accuracy", or more preferably, test performance. Meta-analysis is not required for a systematic review, but when appropriate, it should be undertaken with a dual goal: to provide summary estimates for key quantities, and to explore and explain any observed dissimilarity (heterogeneity) in the results of the examined studies.

"Summing-up" information on test performance metrics such as sensitivity, specificity, and predictive values is rarely the most informative part of a systematic review of a medical test.[1–4] Key clinical questions driving the evidence synthesis (e.g., is this test alone or in combination with a test-and-treat strategy likely to improve decision-making 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 issue of the journal.[5–7] Further, in this paper we assume that medical test performance has been measured against a "gold standard", that is a reference standard that is considered adequate in defining the presence or absence of the condition of interest. Another paper in this supplement discusses ways to summarize medical tests when such a reference standard does not exist.[8]

Syntheses of medical test data often focus on test performance, and much of the attention to statistical issues relevant to synthesizing medical test evidence focuses on summarizing test performance data; thus their meta-analysis was chosen to be the focus of this paper. We will assume that the decision to perform meta-analyses of test performance data is justified and taken, and will explore two central challenges, namely how do we quantitatively summarize medical test performance when: 1) the sensitivity and specificity estimates of

various studies do not vary widely, or 2) the sensitivity and specificity of various studies vary over a large range.

1) Briefly, it may be helpful to use a "summary point" (a summary sensitivity and summary specificity pair) to obtain summary test performance when sensitivity and specificity estimates do not vary widely across studies. This could happen in meta-analyses where all studies have the same explicit test positivity threshold (a threshold for categorizing the results of testing as positive or negative) since if studies have different explicit thresholds, the clinical interpretation of a summary point is less obvious, and perhaps less helpful. However, an explicit common threshold is neither sufficient nor necessary for opting to synthesize data with a "summary point"; a summary point can be appropriate whenever sensitivity and specificity estimates do not vary widely across studies.
2) When the sensitivity and specificity of various studies vary over a large range, rather than using a "summary point", it may be more helpful to describe how the average sensitivity and average specificity relate by means of a "summary line". This oft-encountered situation can be secondary to explicit or implicit variation in the threshold for a "positive" test result, heterogeneity in populations, reference standards, or the index tests, study design, chance, or bias.

Of note, in many applications it may be informative to present syntheses in both ways, as they convey complementary information.

Deciding whether a "summary point" or a "summary line" is more helpful as a synthesis is subjective, and no hard-and-fast rules exist. We briefly outline common approaches for meta-analyzing medical tests, and discuss principles for choosing between them. However, a detailed presentation of methods or their practical application is outside the scope of this work. 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. We also assume familiarity with the common measures of medical test performance (reviewed in the Appendix, and in excellent introductory papers).[9] For example, we do not review challenges posed by methodological or reporting shortcomings of test performance studies.[10] The Standards for Reporting of Diagnostic accuracy (STARD) initiative published a 25-item checklist that aims to improve reporting of medical tests studies.[10] We refer readers to other papers in this issue[11] and to several methodological and empirical explorations of bias and heterogeneity in medical test studies.[12–14]

Nonindependence of sensitivity and specificity across studies and why it matters for meta-analysis

In a typical meta-analysis of test performance, we have estimates of sensitivity and specificity for each study, and seek to provide a meaningful summary across all studies. Within each study sensitivity and specificity are independent, because they are estimated from different patients (sensitivity from those with the condition of interest, and specificity from those without). According to the prevailing reasoning, across studies sensitivity and specificity are likely negatively correlated: as one estimate increases the other is expected to decrease. This is perhaps more obvious when studies have different explicit thresholds for "positive" tests (and thus the term "threshold effect" has been used to describe this negative correlation). For example, the D-dimer concentration threshold for diagnosing an acute coronary event can vary from approximately 200 to over 600 ng/mL.[15] It is expected that higher thresholds would correspond to generally lower sensitivity but higher specificity, and the opposite for lower thresholds (though in this example it is not clearly evident; see Fig. 1a). A similar rationale can be invoked to explain between-study variability for tests with more implicit or suggestive thresholds, such as imaging or histological tests.

Negative correlation between sensitivity and specificity across studies may be expected for reasons unrelated to thresholds for positive tests. For example, in a meta-analysis evaluating the ability of serial creatine kinase-MB (CK-MB) measurements to diagnose acute cardiac ischemia in the emergency department,[16, 17] the time interval from the onset of symptoms to serial CK-MB measurements (rather than the actual threshold for CK-MB) could explain the relationship between sensitivity and specificity across studies. The larger the time interval, the more CK-MB is released into the bloodstream, affecting the estimated sensitivity and specificity. Unfortunately, the term "threshold effect" is often used rather loosely to describe the relationship between sensitivity and specificity across studies, even when, strictly speaking, there is no direct evidence of variability in study thresholds for positive tests.

Because of the above, the current thinking is that in general, the study estimates of sensitivity and specificity do not vary independently, but jointly, and likely with a negative correlation. Summarizing the two correlated quantities is a multivariate problem, and multivariate methods should be used to address it, as they are more theoretically motivated.[18, 19] At the same time there are situations when a multivariate approach is not practically different from separate univariate analyses. We will expand on some of these issues.

PRINCIPLES FOR ADDRESSING THE CHALLENGES

To motivate our suggestions on meta-analyses of medical tests, we invoke two general principles

- Principle 1: Favor the most informative way to summarize the data. Here we refer mainly to choosing between a summary point and a summary line, or both.
- Principle 2: Explore the variability in study results with graphs and suitable analyses, rather than relying exclusively on "grand means".

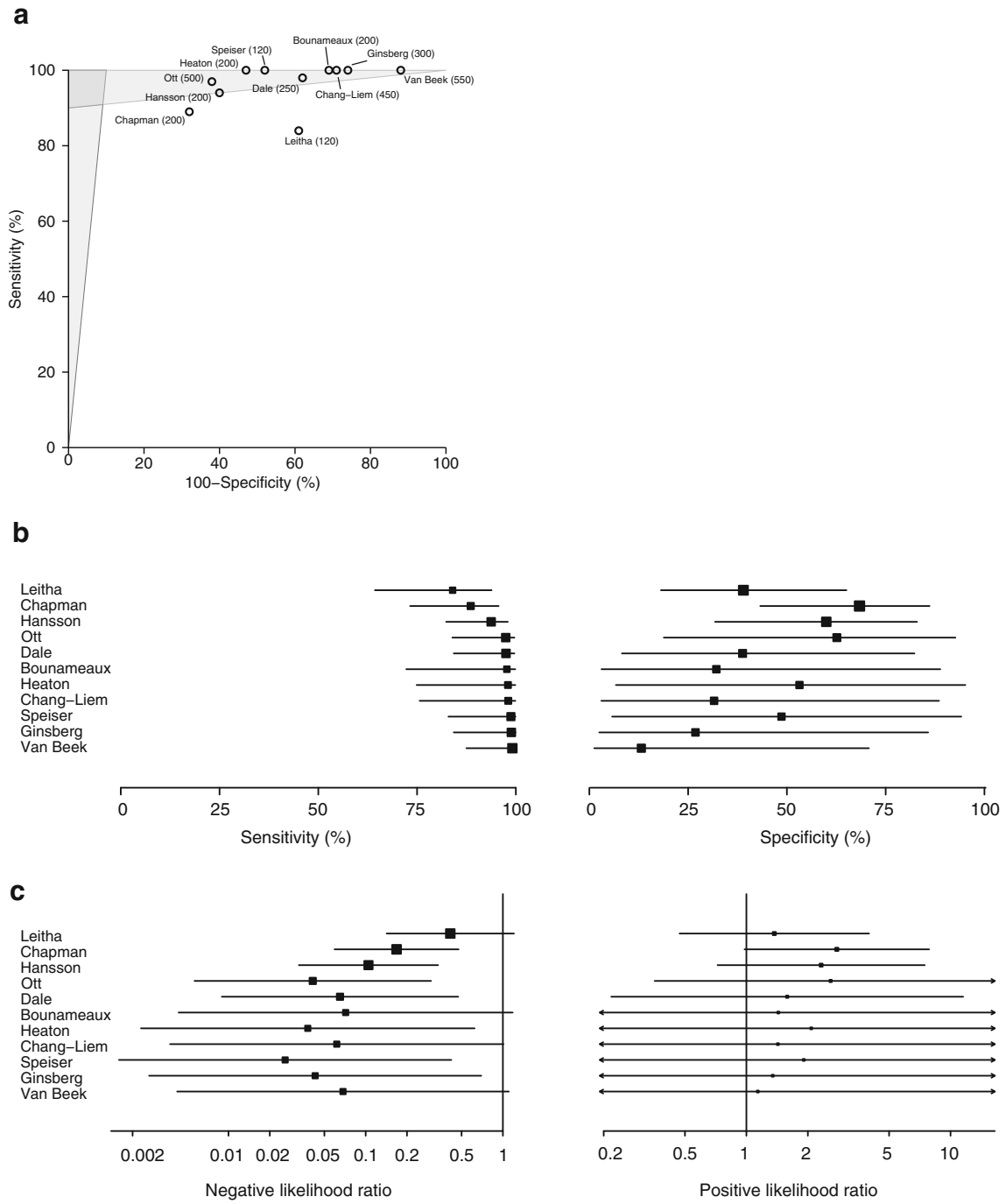

Figure 1. Typical data on the performance of a medical test (D-dimers for venous thromboembolism). Eleven studies on ELISA-based D-dimer assays for the diagnosis of venous thromboembolism.[15] The top panel (a) depicts studies as markers, labeled by author names and thresholds for a positive test (in ng/mL). Studies listed on the left lightly shaded area have a positive likelihood ratio of at least 10. Studies listed on the top lightly shaded area have a negative likelihood ratio of at most 0.1. Studies listed at the intersection of the gray areas (darker gray polygon) have both a positive likelihood ratio of at least 10 and a negative likelihood ratio of 0.1 or less. The second panel (b) shows 'paired' forest plots in ascending order of sensitivity (left) along with with the corresponding specificity (right). Note how sensitivity increases with decreasing specificity, which could be explained by a "threshold effect". The third panel (c) shows the respective negative and positive likelihood ratios.

RECOMMENDED APPROACHES

Which metrics to meta-analyze

For each study, the estimates of sensitivity, specificity, predicitive values, likelihood ratios, and prevalence are related through simple formulas (Appendix). However, if one performed a meta-analysis for each of these metrics, the summaries across all studies will generally be inconsistent: the formulas would not be satisfied *for the summary estimates*. To avoid this, we propose to obtain summaries

for sensitivities and specificities via meta-analysis, and to back-calculate the overall predictive values or likelihood ratios from the formulas in the Appendix, for a range of plausible prevalences. Figure 2 illustrates this strategy for a meta-analysis of K studies. We explain the rationale below.

Why it does make sense to directly meta-analyze sensitivity and specificity. Summarizing studies with respect to sensitivity and specificity aligns well with our understanding of the effect of positivity thresholds for diagnostic tests. Further, sensitivity and specificity are often considered independent of the prevalence of the condition under study (though this is an oversimplification that merits deeper discussion).[20] The summary sensitivity and specificity obtained by a direct meta-analysis will always be between zero and one. Because these two metrics do not have as intuitive an interpretation as likelihood ratios or predictive values,[9] we can use formulas in the Appendix to back-calculate "summary" (overall) predictive values and likelihood ratios that correspond to the summary sensitivity and specificity for a range of plausible prevalence values.

Why it does not make sense to directly meta-analyze positive and negative predictive values or prevalence. Predictive values are dependent on prevalence estimates. Because prevalence is often wide ranging, and because many medical test studies have a case-control design (where prevalence cannot be estimated), it is rarely meaningful to directly combine these across studies. Instead, predictive values can be calculated as mentioned above from the summary sensitivity and specificity for a range of plausible prevalence values.

Why directly meta-analyzing likelihood ratios could be problematic. Positive and negative likelihood ratios could also be combined in the absence of threshold variation, and in fact, many authors give explicit guidance to that effect.[21] However, this practice does not guarantee that the summary positive and negative likelihood ratios are "internally consistent". Specifically, it is possible to get summary likelihood ratios that correspond to impossible "summary" sensitivities or specificities (outside the zero to one interval).[22] Back-calculating the "summary" likelihood ratios from summary sensitivities and specificities avoids this complication. Nevertheless, these aberrant cases are not common,[23] and calculations of summary likelihood ratios by directly meta-analyzing them or from back calculation of the summary sensitivity and specificity rarely results in different conclusions.[23]

Directly meta-analyzing diagnostic odds ratios. The synthesis of diagnostic odds ratios is straightforward and follows standard meta-analysis methods.[24, 25] The diagnostic odds ratio is closely linked to sensitivity, specificity, and likelihood ratios, and it can be easily included in meta-regression models to explore the impact of explanatory variables on between-study heterogeneity. Apart from challenges in interpreting diagnostic odds ratios, a disadvantage is that it is impossible to weight the true positive and false positive rates separately.

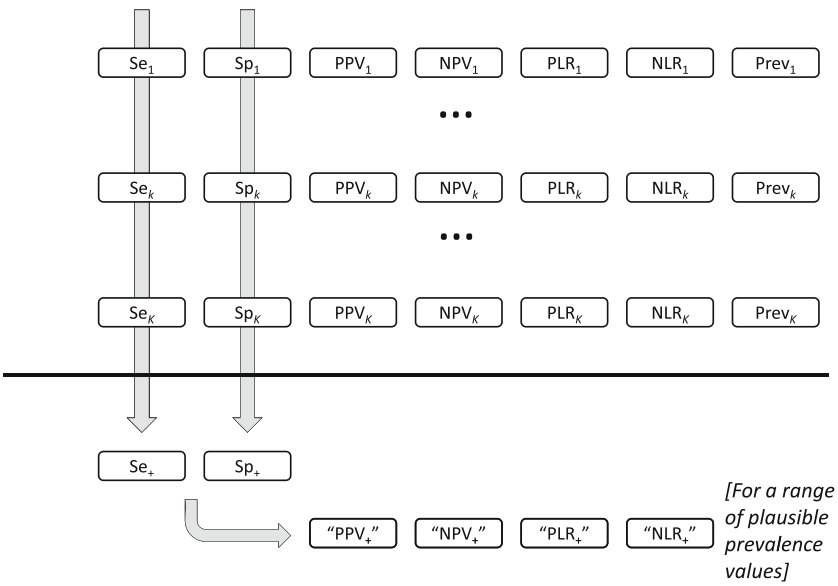

Figure 2. Obtaining summary (overall) metrics for medical test performance. PLR/NLR = positive (negative) likelihood ratio; PPV/NPV = positive (negative) predictive value; Prev = prevalence; Se = Sensitivity; Sp = specificity. The herein recommended approach is to perform a meta-analysis for sensitivity and specificity across the K studies, and then use the summary sensitivity and specificity (Se_+ and Sp_+; a row of two boxes after the horizontal black line) to back-calculate "overall" values for the other metrics (second row of boxes after the horizontal black line). In most cases it is not meaningful to synthesize prevalences (see text).

Desired characteristics of meta-analysis methods

Over several decades many methods have been used for meta-analyzing medical test performance data. Based on the above considerations, methods should be motivated by (a) respecting the multivariate nature of test performance metrics (i.e., sensitivity and specificity); (b) allowing for the nonindependence between sensitivity and specificity across studies ("threshold effect") and (c) allowing for between-study heterogeneity. Table 1 lists commonly used methods for meta-analysis of medical tests. The most theoretically motivated meta-analysis approaches are based on multivariate methods (hierarchical modeling).

We will focus on the case where each study reports a single pair of sensitivity and specificity at a given threshold (although thresholds can differ across studies). 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. These will be described briefly, as a special case.

Preferred methods for obtaining a "summary point" (summary sensitivity and specificity): two families of hierarchical models

When a "summary point" is deemed a helpful summary of a collection of studies, one should ideally perform a *multivariate meta-analysis* of sensitivity and specificity, i.e., a joint analysis of both quantities, rather than separate univariate meta-analyses. This is not only theoretically motivated,[26–28] but also corroborated by simulation analyses.[1, 27, 29]

Multivariate meta-analyses require advanced hierarchical modeling. We can group the commonly used hierarchical models in two families: The so called "bivariate model"[26] and the "hierarchical summary ROC" (HSROC) model.[30] Both use two levels to model the statistical distributions of data. At the first level, they model the counts of the 2×2 table within each study, which accounts for within-study variability. At the second level, they model the between-study variability (heterogeneity), allowing for the theoretically expected nonindependence of sensitivity and specificity across studies. The two families differ in their parameterization at this second level: the bivariate model uses parameters that are transformations of the average sensitivity and specificity—while

Table 1. Commonly Used Methods for Meta-Analysis of Medical Test Performance

Method	Description or comment	Does it have desired characteristics?
Summary point		
Independent meta-analysis of sensitivity and specificity	Separate meta-analyses per metric	Ignores correlation between sensitivity and specificity
	Within-study variability preferably modeled by the binomial distribution.[44]	Underestimates summary sensitivity and specificity and incorrect confidence intervals[26]
Joint (multivariate) meta-analysis of sensitivity and specificity based on hierarchical modeling	Based on multivariate (joint) modeling of sensitivity and specificity. Two families of models[26, 30] (see text), equivalent when there are no covariates[18] Modeling preferably using binomial likelihood rather than normal approximations[30, 37, 45, 46]	The generally preferred method
Summary line		
Moses and Littenberg model	Summary line based on a simple regression of the difference of logit-transformed true and false positive rates versus their average.[32–34]	Ignores unexplained variation between-studies (fixed effects) Does not account for correlation between sensitivity and specificity Does not account for variability in the independent variable Inability to weight studies optimally—yields wrong inferences when covariates are used
Random intercept augmentation of the Moses-Littenberg model	Regression of the difference of logit-transformed true and false positive rates versus their average with random effects to allow for variability across studies[35, 36]	Does not account for correlation between sensitivity and specificity Does not account for variability in the independent variable
Summary ROC based on hierarchical modeling	Same as for multivariate meta-analysis to obtain a summary point—hierarchical modeling[26, 30] Many ways to obtain a (hierarchical) summary ROC : Rutter-Gatsonis (most common)[30] Several alternative curves[37, 38]	Most theoretically motivated method Rutter-Gatsonis HSROC recommended in the Cochrane handbook,[47] as it is the method with which there is most experience

the HSROC model uses a scale parameter and an accuracy parameter, which are functions of sensitivity and specificity, and define an underlying hierarchical summary ROC curve.

In the absence of covariates, the two families of hierarchical models are mathematically equivalent; one can use simple formulas to relate the fitted parameters of the bivariate model to the HSROC model and vice versa, rendering choices between the two approaches moot.[18] The importance of choosing between the two families becomes evident in meta-regression analyses, when covariates are used to explore between-study heterogeneity. The differences in design and conduct of the included diagnostic accuracy studies may affect the choice of the model.[18] For example, "spectrum effects," where the subjects included in a study are not representative of the patients who will receive the test in practice,[31] "might be expected to impact test accuracy rather than the 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."[18] When there are covariates in the model, the HSROC model allows direct evaluation of the difference in accuracy or threshold parameters or both, which affect the degree of asymmetry of the SROC curve, and how much higher it is from the diagonal (the line of no diagnostic information).[18] Bivariate models, on the other hand, allow for direct evaluation of covariates on sensitivity or specificity or both. Systematic reviewers are encouraged to look at study characteristics and think through how study characteristics could affect the diagnostic accuracy, which in turn might affect the choice of the meta-regression model.

Preferred methods for obtaining a "summary line"

When a summary line is deemed more helpful in summarizing the available studies, we recommend summary lines obtained from hierarchical modeling, instead of several simpler approaches (Table 1).[32–36] As mentioned above, when there are no covariates, the parameters of hierarchical summary lines can be calculated from the parameters of the bivariate random effects models using formulas.[18, 30, 37] In fact, a whole range of HSROC lines can be constructed using parameters from the fitted bivariate model;[37, 38] one proposed by Rutter and Gatsonis[30] is an example. The various HSROC curves represent alternative characterizations of the bivariate distribution of sensitivity and specificity, and can thus have different shapes. Briefly, apart from the commonly used Rutter-Gatsonis HSROC curve, alternative curves include those obtained from a regression of logit-transformed true positive rate on logit-transformed false positive rate; logit false positive rate on logit true positive rate; or the major axis regression between logit true and false positive rates.[37, 38]

When the estimated correlation between sensitivity and specificity is positive (as opposed to the typical negative correlation) the latter three alternative models can generate curves that follow a downward slope from left to right. This is not as rare as once thought[37]– a downward slope (from left to right) was observed in approximately one out of three meta-analyses in a large empirical exploration of 308 meta-analyses (report under review, Tufts Evidence-based Practice Center). Chappell et al. argued that in meta-analyses with evidence of positive estimated correlation between sensitivity and specificity (e.g., based on the correlation estimate and confidence interval or its posterior distribution) it is meaningless to use an HSROC line to summarize the studies,[38] as a "threshold effect" explanation is not possible. Yet, even if the estimated correlation between sensitivity and specificity is positive (i.e., not in the "expected" direction), an HSROC still represents how the summary sensitivity changes with the summary specificity. The difference is that the explanation for the pattern of the studies cannot involve a "threshold effect"; rather, it is likely that an important covariate has not been included in the analysis (see the proposed algorithm below).[38]

A special case: joint analysis of sensitivity and specificity when studies report multiple thresholds

It is not uncommon for some studies to report multiple sensitivity and specificity pairs at several thresholds for positive tests. 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 which gives the highest sensitivity may be appropriate in medical tests to rule-out disease. Another option *is to use all available thresholds per study*. Specifically, Dukic and Gatsonis extended the HSROC model to analyze sensitivity and specificity data reported at more than one threshold.[39] This model represents as extension of the HSROC model discussed above. Further, if each study reports enough data on sensitivity and specificity to construct a ROC curve, Kester and Buntinx[40] proposed a little-used method to combine whole ROC curves.

Both models are theoretically motivated. The Dukic and Gatsonis model is more elaborate and more technical in its implementation than the Kester and Buntinx variant. 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 to always perform such analyses. Systematic reviewers are mildly 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

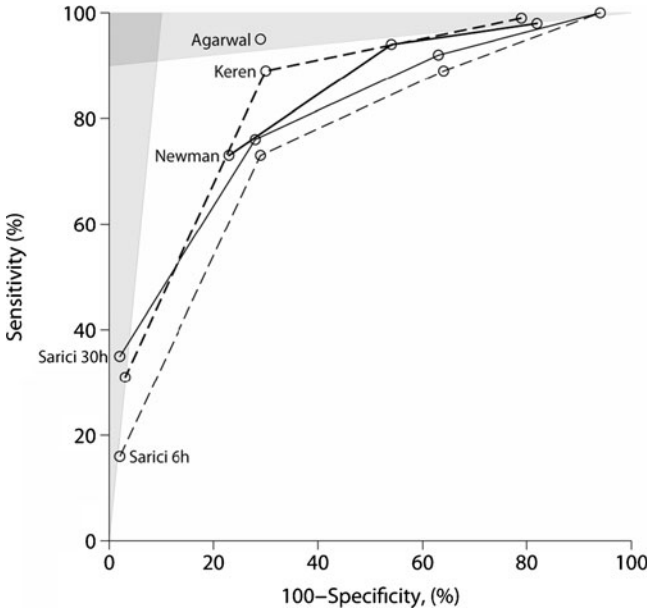

Figure 3. Graphical presentation of studies reporting data at multiple thresholds. Ability of early total serum bilirubin measurements to identify postdischarge total serum bilirubin above the 95th hour-specific percentile. Sensitivity and 100 percent minus specificity pairs from the same study (obtained with different cut-offs for the early total serum bilirubin measurement) are connected with lines. These lines are reconstructed based on the reported cut-offs, and are not perfect representations of the actual ROC curves in each study (they show only a few thresholds that could be extracted from the study). Studies listed on the left lightly shaded area have a positive likelihood ratio of at least 10. Studies listed on the top lightly shaded area have a negative likelihood ratio of at most 0.1. Studies listed at the intersection of the gray areas (darker gray polygon) have both a positive likelihood ratio of at least 10 and a negative likelihood ratio of 0.1 or less.[41]

interest in the models. At a minimum, we suggest that systematic reviewers perform explorations in a qualitative, graphical depiction of the data in the ROC space (see Algorithm section). This will provide a qualitative summary and highlight similarities and differences among the studies. An example of such a graph is Figure 3, which illustrates the diagnostic performance of early measurements of total serum bilirubin (TSB) to identify post-discharge TSB above the 95th 10- hour-specific percentile in newborns.[41]

A WORKABLE ALGORITHM

We propose using the following three step algorithm for meta-analyzing studies of medical test performance when there is a "gold standard". This algorithm should assist meta-analysts in deciding whether a summary point, a summary line, or both are helpful syntheses of the data. When reviewing the three step algorithm, keep these points in mind:

- A summary point may be less helpful or interpretable when the studies have different explicit thresholds for positive tests, and when the estimates of sensitivity vary widely along different specificities. In such cases, a summary line may be more informative.
- A summary line may not be well estimated when the sensitivities and specificities of the various studies show little variability or when their estimated correlation across studies is small. Further, if there is evidence that the estimated correlation of sensitivity and specificity across studies is positive (rather than negative, which would be more typical), a "threshold effect" is not a plausible explanation for the observed pattern across studies. Rather, it is likely that an important covariate has not been taken into account.
- In many applications, a reasonable case can be made for summarizing studies both with a summary point and with a summary line, as these provide alternative perspectives.

Step 1: Start by considering sensitivity and specificity independently

This step is probably self explanatory; it encourages reviewers to familiarize themselves with the pattern of study-level sensitivities and specificities. It is very instructive to create side-by-side forest plots of sensitivity and specificity in which studies are ordered by either sensitivity or specificity. The point of the graphical assessment is to obtain a visual impression of the variability of sensitivity and specificity across studies, as well as an impression of any relationship between sensitivity and specificity across studies, particularly if such a relationship is prominent (Fig. 1 and illustrative examples).

If a summary point is deemed a helpful summary of the data, it is reasonable to first perform separate meta-analyses of sensitivity and specificity. The differences in the point estimates of summary sensitivity and specificity with univariate (separate) versus bivariate (joint) meta-analyses is often small. In an empirical exploration of 308 meta-analyses, differences in the estimates of summary sensitivity and specificity were rarely larger than 5 % (report under review, Tufts Evidence-based Practice Center). The width of the confidence intervals for the summary sensitivity and specificity is also similar between univariate and bivariate analyses. This suggests that practically, univariate and multivariate analyses may yield comparable results. However, our recommendation is to prefer reporting the results from the hierarchical (multivariate) meta-analysis methods because of their better theoretical motivation and because of their natural symmetry with the multivariate methods that yield summary lines.

Step 2: Multivariate meta-analysis (when each study reports a single threshold)

To obtain a summary point, meta-analysts should perform bivariate meta-analyses (preferably using the exact binomial likelihood).

Meta-analysts should obtain summary lines based on multivariate meta-analysis models. The interpretation of the summary line should not automatically be that there are "threshold effects". This is most obvious when performing meta-analyses with evidence of a positive correlation between sensitivity and specificity, which cannot be attributed to a "threshold effect", as mentioned above.

If more than one threshold is reported per study and there is no strong *a priori* rationale to review only results for a specific threshold, meta-analysts should consider incorporating alternative thresholds into the appropriate analyses discussed previously. Tentatively, we encourage both qualitative analysis via graphs and quantitative analyses via one of the multivariate methods mentioned above.

Step 3. Explore between-study heterogeneity

Other than accounting for the presence of a "threshold effect", the HSROC and bivariate models provide flexible ways to test and explore between-study heterogeneity. The HSROC model allows one to examine whether any covariates (study characteristics) explain the observed heterogeneity in the accuracy and threshold parameters. One can use the same set of covariates for both parameters, but this is not mandatory, and should be judged for the application at hand. On the other hand, bivariate models allow one to use covariates to explain heterogeneity in sensitivity or specificity or both; and again, covariates for each measure can be different. Covariates that reduce the unexplained variability across studies (heterogeneity) may represent important characteristics that should be taken into account when summarizing the studies, or they may represent spurious associations. We refer to other texts for a discussion of the premises and pitfalls of metaregressions.[24, 42] Factors reflecting differences in patient populations and methods of patient selection, methods of verification and interpretation of results, clinical setting, and disease severity are common sources of heterogeneity. Investigators are encouraged to use multivariate models to explore heterogeneity, especially when they have chosen these methods for combining studies.

Illustrations

We briefly demonstrate the above with two applied examples. The first example on D-dimer assays for the diagnosis of venous thromboembolism[15] shows heterogeneity which could be attributed to a "threshold effect" as discussed by Lijmer et al..[43] The second example is from an evidence report on the use of serial creatine kinase-MB measurements for the diagnosis of acute cardiac ischemia,[16, 17] and shows heterogeneity for another reason.

D-dimers for diagnosis of venous thromboembolism. D-dimers are fragments specific for fibrin degradation in plasma, and can be used to diagnose venous thromboembolism. Figure 1 presents forest plots of the sensitivity and specificity and the likelihood ratios for the D-dimer example.[43] Sensitivity and specificity appear more heterogeneous than the likelihood ratios (this is true by formal testing for heterogeneity). This may be due to threshold variation in these studies (from 120 to 550 ng/mL, when stated; Fig. 1), or due to other reasons.[43]

Because of the explicit variation in the thresholds for studies of D-dimers, it is probably more helpful to summarize the performance of the test using a HSROC, rather than to provide summary sensitivities and specificities (Fig. 4a). (For simplicity, we select the highest threshold from two studies that report multiple ELISA thresholds.) This test has very good diagnostic ability, and it appropriately focuses on

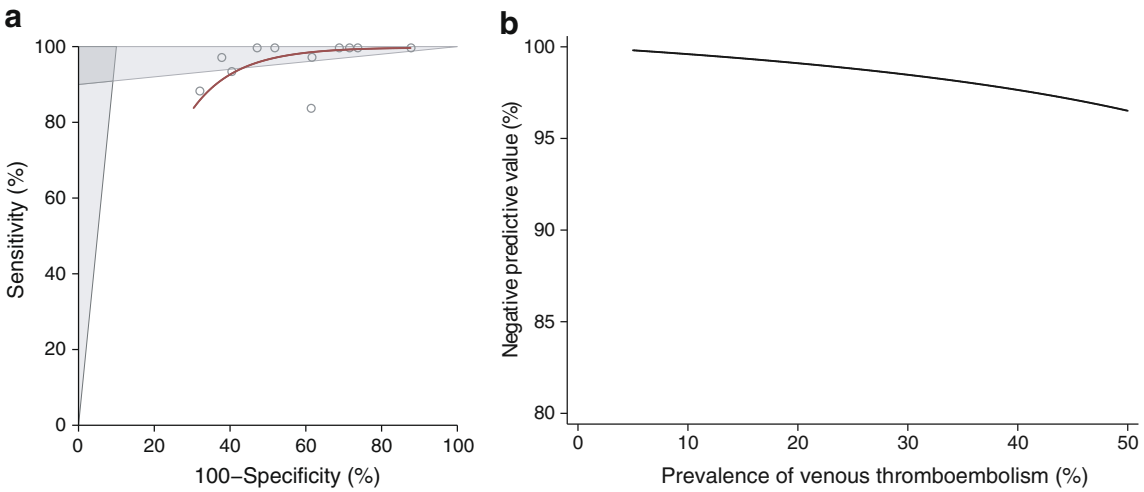

Figure 4. HSROC for the ELISA-based D-dimer tests. (a) Hierarchical summary receiver-operator curve (HSROC) of the studies plotted in Fig. 1a. (b) Calculated negative predictive value for the ELISA-based D-dimer test if the sensitivity and specificity are fixed at 80 % and 97 %, respectively, and prevalence of venous thromboembolism varies from 5 to 50 %.

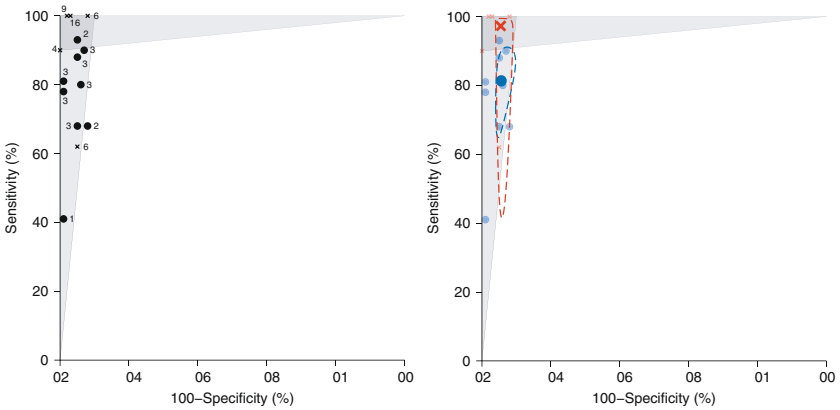

Figure 5. Sensitivity 1–specificity plot for studies of serial CK-MB measurements. The left panel shows the sensitivity and specificity of 14 studies according to the timing of the last serial CK-MB measurement for diagnosis of acute cardiac ischemia. The numbers next to each study point are the actual length of the time interval from symptom onset to last serial CK-MB measurement. Filled circles: at most 3 hours; "x" marks: longer than 3 hours. The right panel plots the summary points and the 95 % confidence regions for the aforementioned subgroups of studies (at most 3 hours: filled circles; longer than 3 hours— "x"s). Estimates are based on a bivariate meta-regression using the time interval as a predictor. The predictor has distinct effects for sensitivity and specificity. This is the same analysis as in Table 2.

minimizing false negative diagnoses. It is also informative to estimate "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 4b shows such an example using the summary sensitivity and specificity of the 11 studies of Figure 4a.

Second example: 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.[16, 17] Figure 5 shows the 14 eligible studies along with how many hours after symptom onset the last measurement was taken. It is evident that there is between-study heterogeneity in the sensitivities, and that sensitivity increases with longer time from symptom onset.

For illustrative purposes, we compare the summary sensitivity and specificity of studies where the last measurement was performed within three hours of symptom onset versus greater than three hours from symptom onset (Table 2). We used a bivariate multilevel model with exact binomial likelihood. In the fixed effects part of the model, we include a variable that codes whether the last measurement was earlier than three hours from symptom onset or not. We allow this variable to have different effects on the summary sensitivity and on the summary specificity. This is essentially a bivariate meta-regression.

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

OVERALL RECOMMENDATIONS

We summarize:

- Consider presenting a "summary point" when sensitivity and specificity do not vary widely across studies, and studies use the same explicit or "implicit threshold".
  - To obtain a summary sensitivity and specificity use the theoretically motivated bivariate meta-analysis models.
  - Back-calculate overall positive and negative predictive values from summary estimates of sensitivity and specificity, and for a plausible range of prevalence values rather than meta-analyzing them directly.
  - Back-calculate overall positive and negative likelihood ratios from summary estimates of sensitivity and specificity, rather than meta-analyzing them directly.

Table 2. Meta-Regression-Based Comparison of Diagnostic Performance

Meta-analysis metric	≤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

Results based on a bivariate meta-regression that effectively compared the summary sensitivity and summary specificity according to the timing of the last serial CK-MB measurement for diagnosis of acute cardiac ischemia. The meta-regression is on a variable that takes the value 1 if the time from the onset of symptoms to testing was 3 hours or less, and the value 0, when the respective time interval was more than 3 hours. The bivariate meta-regression model allows for different effects of timing on sensitivity and specificity. To facilitate interpretation, we present the summary sensitivity and specificity in each subgroup, calculated from the parameters of the meta-regression model, which also gave the p-values for the effect of timing on test performance.

- If the sensitivity and specificity vary over a large range, it may be more helpful to use a summary line, which best describes the relationship of the average sensitivity and specificity. The summary line approach is also most helpful when different explicit thresholds are used across studies. To obtain a summary line use multivariate meta-analysis methods such as the HSROC model.
  - Several SROC lines can be obtained based on multivariate meta-analysis models, and they can have different shapes.
  - If there is evidence of a positive correlation, the variability in the studies cannot be secondary to a "threshold effect"; explore for missing important covariates. Arguably, the summary line is a valid description of how average sensitivity relates to average specificity.
- If more than one threshold is reported per study, this has to be taken into account in the quantitative analyses. We encourage both qualitative analysis via graphs and quantitative analyses via proper methods.
- One 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.

Acknowledgment: *This manuscript is based on work funded by the Agency for Healthcare Research and Quality (AHRQ). All authors are members of AHRQ-funded Evidence-based Practice Centers. The opinions expressed are those of the authors and do not reflect the official position of AHRQ or the U.S. Department of Health and Human Services.*

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Corresponding Author: *Thomas A. Trikalinos, MD; Thomas A. Trikalinos, Center for Evidence-based Medicine, and Department of Health Services Policy and Practice, Brown University, G-S121-7, Providence, RI, 02912, USA (e-mail: thomas_trikalinos@brown.edu).*

REFERENCES

1. **Lau J, Ioannidis JP, Schmid CH.** Summing up evidence: one answer is not always enough. Lancet. 1998;351(9096):123–127.
2. **Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, et al.** Challenges in systematic reviews of diagnostic technologies. Ann Intern Med. 2005;142(12 Pt 2):1048–1055.
3. **Lijmer JG, Leeflang M, Bossuyt PM.** Proposals for a phased evaluation of medical tests. Med Decis Making. 2009;29(5):E13–E21.
4. **Lord SJ, Irwig L, Bossuyt PM.** Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making. 2009;29(5):E1–E12.
5. **Trikalinos TA, Kulasingam S, Lawrence WF.** Chapter 10: Deciding Whether to Complement a Systematic Review of Medical Tests with Decision Modeling. J Gen Intern Med 2012. doi:10.1007/s11606-012-2019-3.
6. **Matchar DB.** Introduction to the methods guide for medical test reviews. J Gen Intern Med. 2011. doi:10.1007/s11606-011-1798-2.
7. **Hartmann K.** Chapter 6: Assessing applicability of medical test studies in systematic reviews. J Gen Intern Med. 2011. doi:10.1007/s11606-011-1961-9.
8. **Trikalinos TA, Ballion CM.** Chapter 9: Options for summarizing medical test performance in the absence of a "gold standard.". J Gen Intern Med. 2012. doi:10.1007/s11606-012-2031-7.
9. **Loong TW.** Understanding sensitivity and specificity with the right side of the brain. BMJ. 2003;327(7417):716–719.
10. **Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al.** The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138(1): W1–12.
11. **Santaguida PL, Riley CM, Matchar DB.** Chapter 5: Assessing risk of bias as a dojmain of quality in medical test studies. J Gen Intern Med. 2012. doi:10.1007/s11606-012-2030-8.
12. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al.** Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282(11):1061–1066.
13. **Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM.** Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174(4):469–476.
14. **Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J.** Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189–202.
15. **Becker DM, Philbrick JT, Bachhuber TL, Humphries JE.** D-dimer testing and acute venous thromboembolism. A shortcut to accurate diagnosis? Arch Intern Med. 1996;156(9):939–946.
16. **Balk EM, Ioannidis JP, Salem D, Chew PW, Lau J.** Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. Ann Emerg Med. 2001;37(5):478–494.
17. **Lau J, Ioannidis JP, Balk E, Milch C, Chew P, Terrin N et al.** Evaluation of technologies for identifying acute cardiac ischemia in emergency departments. *Evid Rep Technol Assess (Summ)* 2000;(26):1-4.
18. **Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA.** A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239–251.
19. **Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al.** Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994;120(8):667–676.
20. **Leeflang MMG, Bossuyt PM, Irwig L.** Diagnostic accuracy may vary with prevalence: Implications for evidence-based diagnosis. *J Clin Epidemiol* 2009;62(1):5–12.
21. **Deeks JJ, Altman DG.** Diagnostic tests 4: likelihood ratios. BMJ. 2004;329(7458):168–169.
22. **Zwinderman AH, Bossuyt PM.** We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med. 2008;27(5):687–697.
23. **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.
24. **Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ, et al.** Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011;64(11):1187–1197.
25. **Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM.** The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56(11):1129–1135.
26. **Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH.** Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982–990.
27. **Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR.** Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Med Res Methodol. 2007;7:3.
28. **Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR.** An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med. 2007;26(1):78–97.
29. **Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, 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.

30. **Rutter CM, Gatsonis CA.** Regression methods for meta-analysis of diagnostic test data. Acad Radiol. 1995;2(Suppl 1):S48–S56.
31. **Mulherin SA, Miller WC.** Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med. 2002;137 (7):598–602.
32. **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.
33. **Littenberg B, Moses LE.** Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making. 1993;13(4):313–321.
34. **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):1293–1316.
35. **Oei EH, Nikken JJ, Verstijnen AC, Ginai AZ, Myriam Hunink MG.** MR imaging of the menisci and cruciate ligaments: a systematic review. Radiology. 2003;226(3):837–848.
36. **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.
37. **Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T.** Bivariate random effects meta-analysis of ROC curves. Med Decis Making. 2008;28(5):621–638.
38. **Chappell FM, Raab GM, Wardlaw JM.** When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med. 2009;28(21):2653–2668.
39. **Dukic V, Gatsonis C.** Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003;59 (4):936–946.
40. **Kester AD, Buntinx F.** Meta-analysis of ROC curves. Med Decis Making. 2000;20(4):430–439.
41. **Trikalinos TA, Chung M, Lau J, Ip S.** Systematic review of screening for bilirubin encephalopathy in neonates. Pediatrics. 2009;124(4):1162–1171.
42. **Thompson SG, Sharp SJ.** Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med. 1999;18(20):2693–2708.
43. **Lijmer JG, Bossuyt PM, Heisterkamp SH.** Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med. 2002;21(11):1525–1537.
44. **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.
45. **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):1331–1332.
46. **Hamza TH, Reitsma JB, Stijnen T.** Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary ROC approaches. Med Decis Making. 2008;28 (5):639–649.
47. Cochrane Diagnostic Test Accuracy Working Group. Handbook for diagnostic test accuracy reviews. 2011. The Cochrane Collaboration. Ref Type: Report