tapraid5/met-met/met-met/met00212/met2187d12z | xppws | S=1 | 4/17/12 | 12:56 | Art: 2010-0161

Psychological Methods

1082-989X/12/$12.00 DOI: 10.1037/a0028091

A Mixed Model Approach to Meta-Analysis of Diagnostic Studies With Binary Test Outcome

Philipp Doebler and Heinz Holling
University of Münster

Dankmar Böhning
University of Reading, Whiteknights

We propose 2 related models for the meta-analysis of diagnostic tests. Both models are based on the bivariate normal distribution for transformed sensitivities and false-positive rates. Instead of using the logit as a transformation for these proportions, we employ the t_α family of transformations that contains the log: logit and (approximately) the complementary log. A likelihood ratio test for the cutoff value problem is developed, and summary receiver operating characteristic (SROC) curves are discussed. Worked examples showcase the methodology. We compare the models to the hierarchical SROC model, which in contrast employs a logit transformation. Data from various meta-analyses are reanalyzed, and the reanalysis indicates a better performance of the models based on the t_α transformation.

Keywords: meta-analysis, diagnostic test, cutoff value, SROC curve, SROC modeling, bivariate normal distribution, transformation, linear mixed model

AQ: 1

AQ: 2

Supplemental materials: http://dx.doi.org/10.1037/a0028091.supp

The correct diagnosis of a specific condition is of eminent and prime interest in psychology and medicine. Diagnostic tests are common tools to discern the presence or absence of a condition, or to screen patients who are at risk to develop a condition. Many diagnostic tests are based on scores derived from brief questionnaires or rely on a single biomarker. Hence they will not always yield a correct diagnosis. When primary studies assessing the quality of a diagnostic test are available, conducting a diagnostic meta-analysis has become a key tool to investigate the available information on a diagnostic test (Egger, Smith, & Altman, 2001; Hasselblad & Hedges, 1995; Schulze, Holling, & Böhning, 2003; Sutton, Abrams, Jones, Sheldon, & Song, 2000). In a primary diagnostic study, the quality of a diagnostic test is often measured in terms of the sensitivity (true-positive rate) and the specificity (true-negative rate = 1 − false-positive rate) of the test; that is, parallel to a gold standard procedure, which defines the presence of a certain condition, the diagnostic test is performed and then the *sensitivity* (the ratio of the number of true-positive cases identified by the diagnostic test and the number of positive cases according to the gold standard) and the *specificity* (the ratio of the number of true negatives identified by the test and the number of true-negative cases) can be calculated.

Philipp Doebler and Heinz Holling, Statistics and Quantitative Methods, Faculty of Psychology and Sport Science, University of Münster, Münster, Germany; Dankmar Böhning, Department of Mathematics and Statistics, University of Reading, Whiteknights, Reading, England.

Dankmar Böhning is now at Southampton Statistical Sciences Research Institute and School of Mathematics, University of Southampton, Highfield Campus, Southampton, England.

This work is funded by German Research Foundation Grant Hol1286/7-1.

Correspondence concerning this article should be address to Philipp Doebler, Faculty of Psychology and Sport Science, University of Münster, Fliednerstrasse 21, D-48149 Münster, Germany. E-mail: doebler@uni-muenster.de

Heterogeneity in Diagnostic Meta-Analysis

One can expect the observed sensitivities and specificities to vary across primary studies. This is due to two main reasons:

1. Different authors will calibrate a test differently. Given a score from a questionnaire or a level of a biomarker, a researcher will have to decide which minimum value (or maximum value) should yield a positive result of the test. This value is known as the *cutoff value.* Sometimes, especially in screening tests for rare conditions, a cutoff value will be set to achieve a certain level of sensitivity such as 95%, often leading to small specificity, but mostly some kind of compromise between sensitivity and specificity is found. Note that both approaches lead to a variety of cutoff values. In general, a population-specific calibration aiming at a certain level of sensitivity will result in different cutoff values for different populations.
2. When a diagnostic test is applied to several populations, one can expect different sensitivities and specificities, even if the same cutoff value is used.

In a diagnostic meta-analysis, one evaluates the quality of a diagnostic test by integrating over data from the primary studies, which usually include sensitivities and specificities; some of these primary studies might not report the cutoff value. This challenge, dealing with inhomogeneous and typically unknown cutoff values, is known as the *cutoff value problem.*

Receiver Operating Characteristic and Summary Receiver Operating Characteristic Curves

At the primary study level, an important graphical tool for choosing a cutoff value is the *receiver operating characteristic* (ROC) curve of a test. This is the curve of sensitivity versus the

false-positive rate as the cutoff value varies (see Pepe, 2000, 2004, for more statistical background on ROC curves). Figure 1 shows an example of such a curve. In many of psychology's subfields, ROC curves have been recognized as valuable tools. Examples range from educational psychology (e.g., Kettler & Elliott, 2010) to clinical psychology (e.g., Bredemeier et al., 2010; Cornell, Peterson, & Richards, 1999; White & Grilo, 2011) and industrial and organizational psychology (e.g., Lehr, Koch, & Hillert, 2010; Stillman & Jackson, 2005). On the level of a primary study, ROC curves help to understand how the diagnostic accuracy of a test with a binary outcome depends on the cutoff value.

As information on diagnostic accuracy of various tests mounts, *summary receiver operating characteristic* (SROC) curves provide a concept on the meta-analytic level corresponding to ROC curves. To obtain an SROC curve, typically the expectation of the sensitivity is computed conditional on a given false-positive rate, a model, and parameters. SROC curves look similar to ROC curves (see Figure 1) and have a similar interpretation, but the difference is as follows. Whereas the ROC curve relates the sensitivity to the false-positive rate in a specific study, the SROC curve relates the sensitivity to the false-positive rate in a collection of studies (Jones & Athanasiou, 2005; Sutton et al., 2000; Walter, 2002). It should be noted that typically meta-analytic approaches do not aim to pool ROC curves stemming from primary studies; so even if primary studies do not report ROC curves but merely 2×2 tables of diagnostic accuracy, meta-analytic approaches that include SROC curves are nevertheless useful for the following four reasons. First, the majority of examples of real-world data that we reanalyzed show a substantial amount of heterogeneity of diagnostic accuracy data, and at least some of it is due to the variability of the underlying cutoffs. Since one of the key aims of meta-analysis is to explain observed heterogeneity, it is crucial to include the underlying SROC curve in any statistical approach. Second, one can also use the SROC curve to predict the outcome of a planned diagnostic study: Given an SROC curve, and a level of sensitivity that the diagnostic study aims at (say, 95%), one can predict the false-positive rate; this obviously also works the other way round. Even when not planning a new study, one can use the SROC, at least as a sanity check, when choosing cutoffs to obtain a certain sensitivity or false-positive rate given the other one. Third, one can compare tests in detail on the meta-analytic level using their SROCs. For example, given two tests with similar pooled pairs of sensitivity and false-positive rate, it could be that not one test is consistently better than the other (see Figure 2); one test could be better as an instrument for mass screening in low-risk settings (where a high false-positive rate is unacceptable) and the other one better suited for high-risk situations where high sensitivity is key. Fourth, the SROC curve could be used to compute measures of diagnostic accuracy that (in some sense) depend on an ROC curve-like area under the curve (Pepe, 2000), Q^* (Gatsonis & Paliwal, 2006), or Youden's index (Böhning, Böhning, & Holling, 2008; Youden, 1950); computations at the meta-analytic level then exhibit greater validity (Gatsonis & Paliwal, 2006).

Current Models for Diagnostic Meta-Analysis

A variety of models for diagnostic meta-analysis have been developed (see Hamza, Reitsma, & Stijnen, 2008, for a recent comparison of many models); we only give a short account of models with SROC curves. For a long time the recommended way to derive an SROC curve has been the approach by Littenberg and Moses (1993) and Moses, Shapiro, and Littenberg (1993). Since

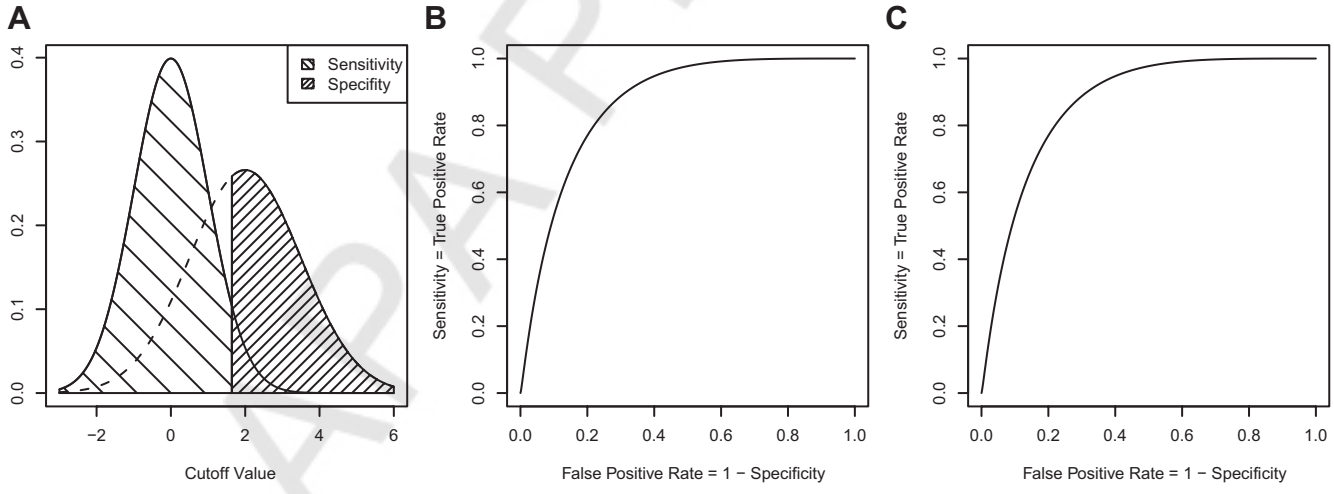

Figure 1. Cutoff values, receiver operating characteristic (ROC) curves, and summary receiver operating characteristic (SROC) curves. (A) Densities of fictional populations with a condition (continuous line, standard normal distributed) and without (dashed line, normal with mean 2 and standard deviation 1.5). On the *x*-axis the (continuous) cutoff value is varied, and every test result that is smaller than the cutoff is a positive result. The vertical line is a cutoff value that yields a sensitivity of 95%; the shaded areas are the sensitivity and specificity, respectively. (B) The ROC curve that results from the populations on the left-hand side. Note that for discrete cutoff values, the ROC curve would be a step function. (C) Fictional pairs of sensitivity and false-positive rate that could stem from diagnostic studies about a test with an ROC curve as in Figure 1B. The curve is identical to the one in Figure 1B and could serve as an SROC curve in this example.

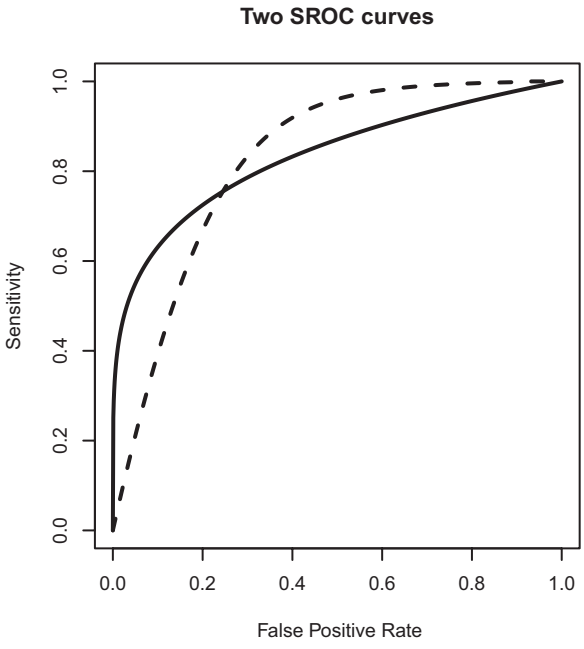

Figure 2. Summary receiver operating characteristic (SROC) curves of two tests with comparable accuracy. These two fictional SROC curves have almost identical area under the curve, but the usefulness of the test depends on the given situation. For mass screening, a test with the solid line SROC curve might be suitable, since its sensitivity is acceptable for small false-positive rate; a test with the dashed SROC curve could be used in a high-risk situation where the identification of every positive is key and higher false-positive rates are acceptable. Conducting a meta-analysis without an SROC might not reveal this difference. This underlines the usefulness of SROC curves, even when no receiver operating characteristic curves are reported in primary studies.

the Littenberg and Moses approach has several shortcomings, several authors have seen the need for alternative approaches (Arends et al., 2008; Rutter & Gatsonis, 2001). Among these are refinements of the Littenberg and Moses approach using random intercepts (Berkey, Hoaglin, Mosteller, & Colditz, 1995; van Houwelingen, Arends, & Stijnen, 2002) and the proportional hazards model (PHM; Holling, Böhning, & Böhning, 2012). The random-intercepts approach has been shown to have large bias for small sample size (Hamza et al., 2008). The other alternatives are the hierarchical SROC model (HSROC) of Rutter and Gatsonis (2001; see also Macaskill, 2004) and the bivariate normal approach of Reitsma et al. (2005). These two alternatives have been shown to lead to the same family of SROC curves, and in fact the models are reparameterizations of each other if one does not use the fully Bayesian approach of Rutter and Gatsonis (Harbord, Deeks, Egger, Whiting, & Sterne, 2007). We will therefore refer to these two models as the HSROC, though we will rather follow Reitsma et al. with respect to technical details. Chu, Guo, and Zhou (2010) extended the HSROC by studying other link functions than the logit used in the HSROC. Many of the mentioned models can be extended by performing *meta-regression*; that is, features of the primary studies are taken as covariates for the parameters of the model. For example, the type of the gold standard procedure might vary and explain part of the variation in diagnostic accuracy.

The HSROC has been applied for meta-analysis and is recommended in the current meta-analytic literature (Leeflang, Deeks, Gatsonis, & Bossuyt, 2008). At the meta-analytic level, the HSROC proposes a bivariate normal distribution of the logit-transformed sensitivities p and false-positive rates q; a coarse approximation to the HSROC is

$$(\text{logit}(p), \text{logit}(q)) \sim N(\mu, \Sigma), \quad (1)$$

for some $\mu \in \mathbb{R}^2$, and a 2×2 covariance matrix Σ. The literature on the HSROC model offers various alternative ways to model the level within a study. Typically, one assumes a binomial model at the primary study level (see Hamza et al., 2008, for more details). This leads to random effects at the within-study level, and since the variance of a binomial variable only depends on its mean, the variances of these random effects are assumed to be known and derived from the observed sensitivities and false-positive rates. One can see the model as a linear mixed model (LMM) or a nonlinear mixed model, depending on whether one incorporates the binomial error structure via an empirical logit transformation (Reitsma et al., 2005) or directly (Harbord et al., 2007). In the following presentation of the HSROC and the subsequent generalization, we adopt the LMM approach.

Disadvantages of the Logit Transformation

The following three shortcomings of the HSROC motivate us to seek generalizations. First, the choice of the logit transformation is, up to a certain degree, arbitrary. Several authors realized this. Chu et al. (2010) demonstrated how to generalize the HSROC by using the complementary log-log transformation, and a proportional hazards family of models based on the log transformation has also been developed (Holling et al., 2012). Second, the SROC curve of the HSROC is not identifiable if only a single pair of sensitivity and false-positive rate is available from each study (Arends et al., 2008; Hamza, Arends, van Houwelingen, & Stijnen, 2009; Rücker & Schuhmacher, 2009); that is, without further (implicit) assumptions the SROC curve, which is a straight line on the logit space, cannot be identified. Note that linearity on logit space is a further (implicit) assumption. Third, in many examples, observed sensitivities tend to cluster around .95, leading to situations where the logit-transformed observed sensitivities are highly skewed (i.e., nonnormal), and hence the distributional assumptions of the HSROC are doubtful. All three shortcomings are related to the logit transformation. We aim to address these shortcomings by studying models based on a family of transformations that varies between the logit and the log, the t_α family of transformations given by

$$t_\alpha(p) = \alpha \log(p) - (2 - \alpha) \log(1 - p),$$

here $p \in [0, 1]$ and $0 \leq \alpha \leq 2$. Note that for $\alpha = 1$ the logit is obtained, and for $\alpha = 0$ one obtains 2 log. Apart from the degenerate cases $\alpha = 0$ and $\alpha = 2$, the family is unbounded and sigmoid, and it is asymmetric apart from the case $\alpha = 1$. It is this asymmetry that qualifies the t_α transformation to serve as a generalization for the logit. For $\alpha = 1.4$ a good approximation of the complementary log transformation (modulo a constant) is obtained, that is, $t_{1.4}(x) \approx 1.42 \log(-\log(1 - x))$ and $t_{0.6}(x) \approx -1.42 \log(-\log(x))$ (see Figure 3). Since the t_α family incorporates the F3

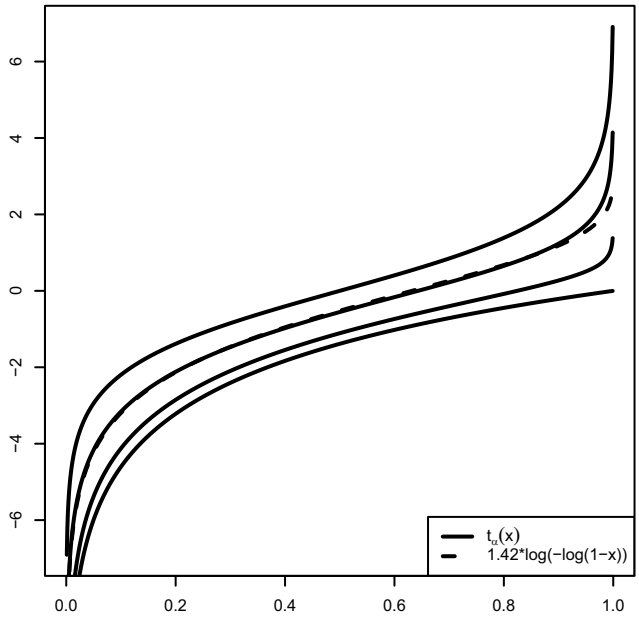

Figure 3. The t_α family and the approximation to the complementary log. The solid lines (from top to bottom) are t_1 = logit, $t_{1.4}$, $t_{1.8}$, and t_2 = 2 log. The dashed line, which almost coincides with $t_{1.4}$, is the complementary log transformation multiplied by a constant. Since t_α is symmetric ($t_\alpha(x) = -t_{2-\alpha}(1-x)$), the figure only shows examples for $\alpha \geq 1$.

AQ: 9

logit and log and approximates the complementary log well, the models we are proposing generalize three models: the HSROC, its generalization with the complementary log (Chu et al., 2010), and, with respect to the SROC curves, the PHM that builds on the log transformation (Holling et al., 2012).

Outline

We will develop two models for meta-analysis of diagnostic tests based on the t_α family, discuss the SROC curves of these models, and illuminate the cutoff value problem. As an illustration of our methods, we reanalyze data from a meta-analysis of the Mini-Mental State Examination (MMSE) by Mitchell (2009) and from a meta-analysis by Patrick et al. (1994), dealing with the accuracy of self-reports of smoking.

The MMSE, which has its origin in Folstein, Folstein, and McHugh (1975), is a short interview that screens for cognitive deficits and whose primary uses are the diagnosis of dementia, to assess the severity of cognitive impairment and to assess the effect of treatment by using the test repeatedly over time. In Mitchell (2009) the MMSE's ability to diagnose dementia and mild cognitive impairment (MCI) is analyzed. Scores in the MMSE can range from 0 to 30; scores that are 24 or less indicate some cognitive impairment, and the lower the score, the more severe the deficit. Numerous cutoff values are suggested in the literature (Crum, Anthony, Bassett, & Folstein, 1993; Folstein et al., 1975; Grigoletto, Zappalà, Anderson, & Lebowitz, 1999); the main reasons for this are that the educational background influences the score and that the MMSE can be used to diagnose mild conditions as well as severe ones. The cutoff values in Mitchell span from 17 to 28; one primary study distinguishes between the educational background of the subject, resulting in inhomogeneous cutoff values even within this primary study. Hence we expect the cutoff value problem to be present in the dementia and the MCI part of the MMSE data. We will reanalyze the data from Mitchell with our models, taking into account the cutoff value problem. We will also compute and compare the SROC curves of the dementia and MCI data and thus answer the question whether the MMSE is better suited to diagnose dementia or MCI.

The validity of self-reports of smoking has been questioned on grounds of the assumption that smokers underestimate their consumption (U.S. Department of Health and Human Services, 1990) or deny smoking at all (Murray, O'Connell, Schmid, & Perry, 1987). Although biochemical tests to detect smoking are available, they are expensive compared to questionnaires and might lead to refusal (Velicer, Prochaska, Rossi, & Snow, 1992). The meta-analysis of Patrick et al. (1994) examined primary studies on self-reported measures of smoking. The primary studies are grouped in two categories: self-administered questionnaires (SAQ) and interviewer-administered questionnaires (IAQ). The gold standard in all the primary studies was a biochemical measure, though several different ones had been used across primary studies. The original data from Patrick et al. contain much more detail about the primary studies, but for the purpose of showcasing the developed methodology, we focus on the question whether SAQ or IAQ is more reliable. As in the MMSE case, the data are highly heterogeneous, again making it plausible that a cutoff value problem is present.

The two models that we present in the following are closely related and are both meant for data from diagnostic studies with binary outcome. In the first model we assume that the t_α-transformed sensitivities and false-positive rates follow a bivariate normal distribution; this leads to a model generalizing the approximation to the HSROC (Equation 1). There are several ways to use the t_α transformation here: One can restrict its use to one of sensitivity and false-positive rate and use the logit for the other one, or one can use different values of α for sensitivity and false-positive rate. Heterogeneity among the studies is modeled with the covariance matrix of this bivariate normal distribution, the cutoff value problem being incorporated as well.

The second model refines the first and builds upon the observation that some variance of the sensitivities and false-positive rates is not due to heterogeneity but due to the fact that the number of true-positive cases and the number of false-positive cases in each study follow a binomial distribution (i.e., it incorporates random effects at the study level like the HSROC). The second model needs as data, apart from sensitivities and false-positive rates, the frequencies of people with and without the condition. For $\alpha = 1$, the second model reduces to the HSROC. We discuss SROC curves for both models and parameter estimation.

We include the first model mainly to illustrate the idea of a bivariate model and transformed proportion data in a simple case. Although it is simple, we stress that inferences obtained from it (pooled diagnostic accuracies and SROC curves) are comparable to these of the more advanced second model. It is also noteworthy that when sample sizes in the primary studies are large, the second model is essentially identical to the first. In this sense the second model asymptotically approaches the first.

The Models

In the following let N denote the number of primary diagnostic studies on a diagnostic test or instrument of interest. One ultimate interest is a condition that this test is targeting to identify (e.g., in the case of the MMSE, meta-analysis dementia or MCI). In the ith study we have m_i persons with the condition and n_i persons without the condition, of which y_i persons with the condition are diagnosed correctly by the test in question and z_i persons without the condition are diagnosed incorrectly (see Table 1). So y_i/m_i estimates the sensitivity p_i of the test, and z_i/n_i estimates the false-positive rate q_i of the test in the ith study. Clearly, conditional upon the ith study, y_i is binomial with mean p_i and variance $p_i(1 - p_i)/m_i$, and z_i is binomial with mean q_i and variance $q_i(1 - q_i)/n_i$. The meta-analytic sampling model has two levels. The first level describes the sampling of studies with parameters (p_i, q_i), and the second level describes the sampling within the ith study. As described above, in the second level we assume a binomial sampling model, whereas we assume for the first level a bivariate normal model. It is well known that in these hierarchical two-level models that the unconditional means are $E[E(y_i/m_i|p_i)] = E(p_i)$ and $E[E(z_i/n_i|q_i)] = E(q_i)$, whereas the unconditional variances have two components: the variance stemming from the variation between studies and the within-study variance.

T1

The First Model

In diagnostic problems the study sizes m_i and n_i for $i = 1, \ldots, N$ are typically large (>100), as illustrated in the examples. Hence in these situations it can be assumed that the estimators $\hat{p}_i := y_i/m_i$ and $\hat{q}_i := z_i/n_i$ yield approximations of p_i and q_i of acceptable precision, so that we assume for the first model

$$\hat{p}_i = p_i \quad \text{and} \quad \hat{q}_i = q_i. \tag{2}$$

This assumption implies that $\hat{p}_i|p_i$ and $\hat{q}_i|q_i$ have negligible variance; the second model will relax this assumption.

In a next step we concentrate on modeling the bivariate distribution of the pairs (p_i, q_i). We view each pair (p_i, q_i) as a realization of a random variable (p, q). First note that if p_i grows, then q_i also increases; this will often be because a large sensitivity requires a high (or very low, depending on the poling of the test) cutoff value, thus producing also many false-positive results. Moreover, note that this fact still holds true if we transform p_i and q_i with a monotone transformation, say, log, logit, or t_α. Hence for $\alpha_p, \alpha_q \in [0, 2]$, it is reasonable to assume that $t_{\alpha_p}(p)$ and $t_{\alpha_q}(q)$ are (positively) correlated; this will be incorporated in our model. The t_α transformation is motivated by the fact that in meta-analysis of binary diagnostic tests, the relationship of p_i and q_i is usually nonlinear. One of the HSROC's implicit assumptions is that the logit-transformed p_i and q_i are roughly linear; for several real-world data sets we observed that the relationship of $\log(p_i)$ and $\log(q_i)$ is roughly linear. Since the t_α family of transformations includes the logit and (up to a fixed factor) the log, it mediates between the two transformations. To cope with study heterogeneity, we assume that the transformed study parameters $(t_{\alpha_p}(p), t_{\alpha_q}(q))$ follow a bivariate normal distribution with mean

$$\mu = (\mu_1, \mu_2)^T$$

and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma \\ \sigma & \sigma_2^2 \end{pmatrix}.$$

As we explained above on the within-study level, p_i and q_i are positively correlated when the cutoff is allowed to vary (see Figure 1). On the meta-analytic level some of this variation due to the cutoff value will still be observable, since different authors calibrate the test to their needs. So on the meta-analytic level σ models the (likely positive) covariation of $t_{\alpha_p}(p)$ and $t_{\alpha_q}(q)$ potentially induced by a cutoff value variation, whereas $\sigma_1^2 \geq 0$ and $\sigma_2^2 \geq 0$ are measures for the heterogeneity among studies. The parameters α_p and α_q could be interpreted as *shape parameters.*

Let us discuss the SROC curve of the first model. The SROC curve can be described as the relationship of the conditional mean $E(t_{\alpha_p}(p)|t_{\alpha_q}(q))$ to $t_{\alpha_q}(q)$. A standard result about bivariate normal distributions shows that $t_{\alpha_p}(p)$ is

$$N\left(\mu_1 + \frac{\sigma}{\sigma_2^2}(t_{\alpha_q}(q) - \mu_2), \sigma_1^2\left(1 - \frac{\sigma}{\sigma_1\sigma_2}\right)\right).$$

Note that $|\sigma| \leq \sigma_1\sigma_2$, following from the Cauchy–Schwarz inequality, so that the variance above involving $\frac{\sigma}{\sigma_1\sigma_2}$ is always nonnegative. Setting $\theta := \frac{\sigma}{\sigma_2^2}$, we achieve the following conditional expectation:

$$E(t_{\alpha_p}(p)|t_{\alpha_q}(q)) = (\mu_1 - \theta\mu_2) + \theta t_{\alpha_q}(q). \tag{3}$$

Hence in the special case that $(\mu_1 - \theta\mu_2) = 0$, $\alpha_p = \alpha_q = 2$, we have

$$2\log(p) = t_{\alpha_p}(p) = \theta t_{\alpha_q}(q) = 2\theta\log(q),$$

and hence

$$p = q^\theta, \tag{4}$$

which is called the Lehmann family, proposed by Le (2006) to describe ROC curves. The model (Equation 4) is also called the PHM, and its application to SROC modeling is discussed in Holling et al. (2012). Note that Equation 3 is a generalization of Equation 4. We can now easily calculate the resulting SROC curve for the first model:

$$p = t_{\alpha_p}^{-1}((\mu_1 - \theta\mu_2) + \theta t_{\alpha_q}(q)). \tag{5}$$

Note that for some values of α the inverse t_α^{-1} has a closed form; even when no closed form exists, the inverse is easily determined numerically.

Table 1
Data from the ith Study in a 2 × 2 Table

Test	With condition	Without condition
Positive	y_i	z_i
Negative	$m_i - y_i$	$n_i - z_i$
Total	m_i	n_i

The Second Model

In a nutshell, we obtain the second model from the first by discarding assumption Equation 2 and assuming that $\hat{p}_i$ and $\hat{q}_i$ are

tapraid5/met-met/met-met/met00212/met2187d12z | xppws | S=1 | 4/17/12 | 12:56 | Art: 2010-0161

realizations of binomial variables. Nevertheless we want to model the t_α-transformed study parameters as realizations of some kind of bivariate normal distribution. Again, we use p_i to denote the true sensitivity of the ith study and q_i to denote the true false-positive rate in the ith study. It is natural to model y_i and z_i as realizations of a binomial variable with parameters p_i and q_i, respectively. Then

$$\mathrm{Var}\left(\frac{y_i}{m_i}\right) = \frac{p_i(1-p_i)}{m_i} \text{ and } \mathrm{Var}\left(\frac{z_i}{n_i}\right) = \frac{q_i(1-q_i)}{n_i}.$$

Using the delta method, we obtain

$$\mathrm{Var}\left(t_{\alpha_p}\left(\frac{y_i}{m_i}\right)\right) \approx \frac{(\alpha_p(1-p_i)-(2-\alpha_p))^2}{m_i p_i(1-p_i)} =: d_{i1}^2 \qquad (6)$$

and

$$\mathrm{Var}\left(t_{\alpha_q}\left(\frac{z_i}{n_i}\right)\right) \approx \frac{(\alpha_q(1-q_i)-(2-\alpha_q))^2}{n_i q_i(1-q_i)} = d_{i2}^2. \qquad (7)$$

If the studies at hand are entirely homogeneous, that is, $p_1 = p_2 = \ldots = p_N$ and $q_1 = q_2 = \ldots = q_N$, then the above variances would explain all the observed variance. To cope with study heterogeneity, we need to incorporate additional variance terms into the modeling. We assume that the transformed observed diagnostic accuracies $(t_{\alpha_p}(\hat{p}_i), t_{\alpha_q}(\hat{q}_i))$ follow a bivariate distribution with mean $\mu = (\mu_1, \mu_2)^T$. Part of the variance of this distribution will be explained by the variances of the form (Equation 6), but the remaining variance, due to study heterogeneity, is incorporated into the model as follows: We assume that there are σ, σ_1^2, and σ_2^2 such that the covariance matrix Σ_i for the ith study is provided as

$$\Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma \\ \sigma & \sigma_2^2 \end{pmatrix} + D_i, \qquad (8)$$

where D_i is the 2×2 diagonal matrix with nonzero elements d_1^2 and d_2^2. This can be seen as an LMM with *known* variances of random effects:

$$(t_{\alpha_p}(\hat{p}_i), t_{\alpha_q}(\hat{q}_i))^T = \mu + \delta_i + \varepsilon_i,$$

with $\delta_i \sim N(0, D_i)$, $\varepsilon_i \sim N(0, \Sigma)$ and δ_i, ε_j independent for $i, j = 1, \ldots, N$. It should be noted that the full likelihood of this model (see Appendix A) contains the Jacobian of the transformation; this is a necessary complication if more than one transformation is to be compared (i.e., more than one value of α_p or α_q is to be considered). Just as in the first model, σ models the positive correlation of $\log(p_i)$ and $\log(q_i)$, and σ_1^2 and σ_2^2 are measures of the heterogeneity between studies.

As can be seen from the covariance matrix, the SROC curve of this model is identical to that of the first model if m_i and n_i are becoming large, since the expressions d_{i1}^2, d_{i2}^2, converge to 0. If study sizes are small, then the SROC curve can only be given study specifically:

$$\mathrm{E}(t_{\alpha_p}(p_i)|t_{\alpha_q}(q_i)) = (\mu_1 - \theta_i) + \theta_i t_{\alpha_q}(q_i), \qquad (9)$$

where

$$\theta_i = \frac{\sigma}{d_{i2}^2 + \sigma_2^2}.$$

An estimated SROC curve for the second model can be obtained by plugging the parameter estimates for the second model into the SROC curve of the first model (i.e., Equation 5). We call this the *analytical* SROC curve because of the convenient form of this curve. This is also in contrast to a Monte Carlo SROC curve, for which one uses random sampling from the estimated parameters of the model to obtain an SROC curve (we outline an algorithm for this in Appendix B).

A few comments are in place. Note that we have made no assumptions regarding the relationship of sensitivity and false-positive rate within a study. It follows solely from the normality assumption that the SROC curve has straight line shape on the t_α space as given in Equation 3 or 9. This also shows its close relationship to the PHM. Note that for $\alpha_p = \alpha_q = 2$, that allowing an intercept on the log scale (or a scaling factor on the SROC scale) will imply that the SROC curve will be larger or smaller than 1 for the sensitivity when the false-positive rate approaches 1. This needs to be kept in mind when doing practical implementations of the concept for this special case.

Parameter Estimation

The parameters of the t_α transformation in both models can either be fixed at interesting values (e.g., $\alpha_p = \alpha_q = 2$ leads to a model generalizing the PHM [Equation 4]) or estimated from the data. We explain the general approach before going into more details. When using parametric transformations like t_α, there are broadly two approaches with respect to the additional parameters θ introduced: The first view suggests that θ is part of the model and is thus to be incorporated into the estimation process and especially in the calculation of standard errors; the second view is to work conditional on θ and not consider it part of the model. In either case θ is typically determined by calculating the *profile likelihood* $p(\theta)$ for many different θ, that is, $p(\theta) = L(\hat{\lambda}(\theta)|\theta)$, where $\hat{\lambda}(\theta)$ is the (maximum likelihood) estimate of the remaining model parameters conditional on θ. Then θ is chosen so that it maximizes $p(\theta)$, for example, when using a grid search.

We stress that in a maximum (adjusted) likelihood approach, both views outlined above lead to the same point estimates of the model parameters; the key difference is felt when computing standard errors of the estimates. The first view leads to *variance inflation*; that is, the additional uncertainty introduced by the parameter of the transformation increases the variance of the other parameters, whereas not considering the parameters of the transformation part of the model gives smaller standard errors. This phenomenon is well understood for the Box and Cox (1964) transformation (Bickel & Doksum, 1981). In the following we adopt the second view of parameter estimation to avoid variance inflation.

For full generality, we aim to construct estimators for the parameters α_p, α_q, μ_1, μ_2, σ_1, σ_2, and σ for the first and second model, though in applications it is often desirable to fix at least one of α_p and α_q. Since both models fall into the class of LMMs, maximum likelihood estimation (MLE) of variance parameters produces biased estimates, especially for small sample size. To address the issue of small sample size, we will use a *restricted maximum likelihood* (REML; also known residual maximum likelihood) correction to the likelihoods, that is, an adjusted likelihood. Maximizing this adjusted likelihood yields unbiased estimation of variance parameters.

Likelihood of the First Model

If we omit the Jacobian, the likelihood of the ith observation $y_i = (\hat{p}_i, \hat{q}_i)$ is given by

$$L_i(\hat{p}_i,\hat{q}_i|\alpha_p,\alpha_q,\mu,\Sigma) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(t(y_i)-\mu)'\Sigma^{-1}(t(y_i)-\mu)\right),$$

where $t(y_i) = t((t_{\alpha_p}(\hat{p}_i), t_{\alpha_q}(\hat{q}_i))$. The log-likelihood function l_i for the ith study is then straightforward to calculate from the above.

If we assume independence between studies, the likelihood function for the full sample is

$$L(\hat{p}_1, \ldots \hat{p}_N,\hat{q}_1, \ldots, \hat{q}_N|\alpha_p, \alpha_q, \mu, \Sigma) = \prod_{i=1}^{N} L_i,$$

and hence the log-likelihood function is

$$l(\hat{p}_1, \ldots \hat{p}_N,\hat{q}_1, \ldots, \hat{q}_N|\alpha_p, \alpha_q, \mu, \Sigma) = \sum_{i=1}^{N} l_i. \quad (10)$$

Likelihood of the Second Model

If we write

$$\Sigma_i = \Sigma + D_i,$$

the likelihood of a single pair for the second model is

$$L_i(\hat{p}_i, \hat{q}_i|\alpha_p, \alpha_q, \mu, \Sigma) = (2\pi)^{-1}|\Sigma_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(t(y_i)-\mu)'\Sigma_i^{-1}(t(y_i)-\mu)\right).$$

We omitted the Jacobian in the above likelihood (see Appendix A). Again, it is not difficult to compute the log-likelihoods of a single pair, and so the full sample log-likelihood is

$$l(\hat{p}_1, \hat{p}_2, \ldots, \hat{q}_1, \hat{q}_2, \ldots|\alpha_p, \alpha_q, \mu, \Sigma) = \sum_{i=1}^{N} -\log(2\pi) - \frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}(t(y_i)-\mu)'\Sigma_i^{-1}(t(y_i)-\mu). \quad (11)$$

REML and MLE for the First Model

For the first model conditional on α_p and α_q, the maximum likelihood estimators have the following closed forms: The vector of the sample means

$$\bar{\mu} := (\bar{\mu}_1, \bar{\mu}_2) := \frac{1}{N}\left(\sum_{i=1}^{N} t_{\alpha_p}(\hat{p}_i), \sum_{i=1}^{N} t_{\alpha_q}(\hat{q}_i)\right)^{\mathrm{T}}$$

is the MLE for $(\mu_1, \mu_2)^{\mathrm{T}}$ and the 2×2 matrix

$$\hat{\Sigma} := \frac{1}{N}WW^{\mathrm{T}},$$

where

$$W = \begin{pmatrix} t_{\alpha_p}(\hat{p}_1) - \bar{\mu}_1 & t_{\alpha_p}(\hat{p}_2) - \bar{\mu}_1 \ldots t_{\alpha_p}(\hat{p}_N) - \bar{\mu}_1 \\ t_{\alpha_q}(\hat{q}_1) - \bar{\mu}_2 & t_{\alpha_q}(\hat{q}_2) - \bar{\mu}_2 \ldots t_{\alpha_q}(\hat{q}_N) - \bar{\mu}_2 \end{pmatrix}$$

is the MLE for Σ, as is well known (see, e.g., Anderson, 2003, for this general fact about multivariate normal distributions). Note that $\hat{\Sigma}$ is obtained by multiplying a $2 \times N$ and an $N \times 2$ matrix. We note that though $\hat{\Sigma}$ is the MLE of Σ, it is not unbiased. An unbiased estimator of the covariance matrix Σ is the *sample covariance matrix* Q, which is obtained by changing the factor $\frac{1}{N}$ to

$$Q := \frac{1}{N-1}WW^{T}.$$

Hence $\hat{\Sigma}$ will underestimate Σ by the factor $\frac{N-1}{N}$, which for small number of studies N is not close to 1. The reason for this is that $\hat{\Sigma}$ depends on $\bar{\mu}$, so the $2N$ degrees of freedom of the model are reduced by 2, leading to the factor

$$\frac{2N-2}{2N} = \frac{N-1}{N}.$$

We use Q subsequently to estimate the parameters of the first model and note that Q is in fact the REML estimate; we give more details though for the REML estimate of the second model. The variance of the MLE is also well studied (see, e.g., Lehmann and Casella, 1998, p. 472). Again, simple closed-form expressions exist that approximate the variance, though it should be noted that for large N the following approximation is considered to be better:

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{\mu}_i) \approx \frac{\hat{\sigma}_i^2}{N} \quad (12)$$

and

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{\sigma}_i^2) \approx \frac{2\hat{\sigma}_i^4}{N} \text{ for } i = 1, 2 \quad \text{and} \quad \hat{\mathrm{V}}\mathrm{ar}(\hat{\sigma}) \approx \frac{\hat{\sigma}^2 + \hat{\sigma}_i^2\hat{\sigma}_2^2}{N}. \quad (13)$$

REML and MLE for the Second Model

We discuss two approaches to parameter estimation: maximum likelihood (ML) and REML. We stress the advantage of the latter. It is well known that the ML estimates of variance parameters and more generally covariance parameters are biased, especially for small sample sizes; for a sample of N studies, the ML estimate $\hat{\Sigma}_{\mathrm{ML}}$ underestimates the true Σ roughly by a factor of $(2N - 2)/2N = (N - 1)/N$ (see the discussion of parameter estimation for the simple model). REML aims to compensate for that. There are different strategies to derive an adjusted likelihood that yields the REML estimates as its maximum (see, e.g., Lee, Nelder, & Pawitan, 2006, Section 5.2.2), but typically the first step is to derive a profile likelihood to eliminate μ.

To obtain a profile likelihood, we first work conditional on Σ and α_p, α_q. Note the following: The variances of the random effects d_{i1}^2 and d_{i2}^2 depend only on the data $y_i = (\hat{p}_i, \hat{q}_i)$, the sample sizes in the ith study m_i, n_i, and α_p and $_{\alpha q}$. One easily checks that

tapraid5/met-met/met-met/met00212/met2187d12z | xppws | S=1 | 4/17/12 | 12:56 | Art: 2010-0161

the partial derivative $\frac{\partial}{\partial \mu} l(\mu, \Sigma)$ of the full log-likelihood $l(y_i|\mu, \Sigma)$ of the ith observation is (in vector notation) proportional to

$$\Sigma_i^{-1}(t(y_i) - \mu),$$

where $t(y_i) = t(t_{\alpha_p}(\hat{p}_i), t_{\alpha_q}(\hat{q}_i))$. Setting the partial derivative of the full likelihood equal to 0 yields

$$\sum_i \Sigma_i^{-1} t(y_i) = \sum_i \Sigma_i^{-1} \mu = \left(\sum_i \Sigma_i^{-1} \right) \mu,$$

and hence

$$\hat{\mu} = \left(\sum_i \Sigma_i^{-1} \right)^{-1} \sum_i \Sigma_i^{-1} t(y_i)$$

maximizes the full likelihood conditional on α_p, α_q, and Σ. By plugging $\hat{\mu}$ into the full likelihood, one obtains the profile log-likelihood

$$p(\vec{y}|\alpha_p, \alpha_q, \Sigma) = \sum_i l(y_i|\alpha_p, \alpha_q, \hat{\mu}, \Sigma). \quad (14)$$

Since $\hat{\mu}$ is estimated from the data, intuitively one loses 2 degrees of freedom in the process. This can be compensated for by studying the following *adjusted profile log-likelihood*:

$$p_{\text{REML}}(\vec{y}|\alpha_p, \alpha_q, \Sigma) = p(\vec{y}|\alpha_p, \alpha_q, \Sigma) - \frac{1}{2} \log |\sum_i \Sigma_i^{-1}/(2\pi)|. \quad (15)$$

Maximizing $p(\vec{y}|\alpha_p, \alpha_q, \Sigma)$ yields the ML estimates, and the REML estimates are obtained by maximizing $p_{\text{REML}}(\vec{y}|\alpha_p, \alpha_q, \Sigma)$. For small sample sizes the difference in the variance components is substantial, and the REML estimates offer an unbiased estimate of Σ.

The log-likelihood (Equation 11), the profile likelihood (Equation 14), and the adjusted profile log-likelihood (Equation 15) of the second model are well behaved in the sense that they allow numerical maximization. Such a numerical maximization was carried out with the function mle and the package mvtnorm in R (Genz et al., 2010; R Development Core Team, 2010). The maximization was found to be more stable when working with the Cholesky decomposition of Σ rather than with Σ itself. Numerical maximization was feasible for all data tested, though we relied on the SANN algorithm of the optim function for some data sets, especially for those with small σ.

Confidence Intervals and Standard Errors

To obtain confidence intervals, we use different strategies for the parameters α_p, α_q, and the remaining parameters. For α_p and α_q it is not difficult to invert the appropriate likelihood ratio test;

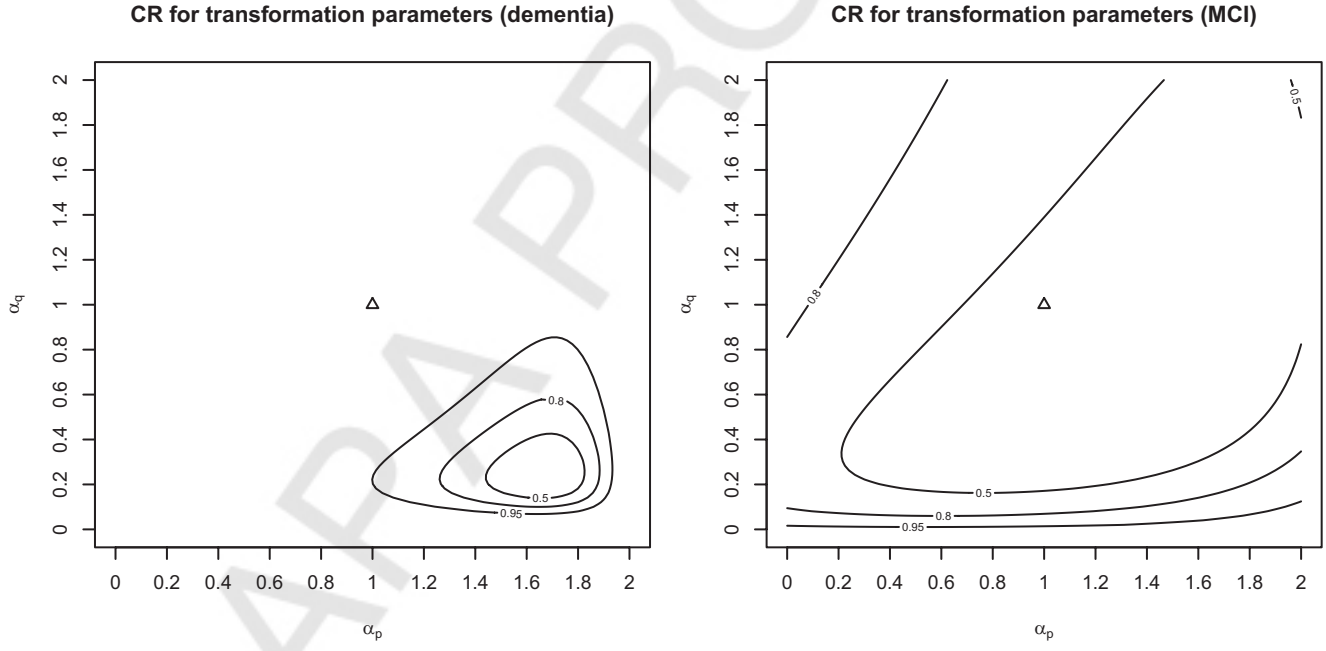

Figure 4. Confidence regions (CRs) for transformation parameters. Estimate of (α_p, α_q) is marked by a circle; the triangle amounts $\alpha_p = \alpha_q = 1$ (i.e., to the logit transformation). The plot was obtained by calculating the log-likelihood of the first model l at the restricted maximum likelihood (REML) estimates for all α on a fine grid and computing the statistic $D = 2(l^* - l)$, where l^* is the log-likelihood at the (REML) estimates. Then D is approximately chi-square distributed with 2 degrees of freedom, so the contours are obtained by comparing D to the quantiles of this distribution. For the dementia subset of the Mini-Mental State Examination data, the transformation parameters are clearly bounded away from the logit. The small sample size of the mild cognitive impairment (MCI) data ($N = 5$) makes the CR far wider than the one obtained for the dementia data; it cannot be concluded that $(\alpha_p, \alpha_q) \neq (1, 1)$. The asymmetric shape of the CRs also underlines that confidence intervals for α_p or α_q should not be based on a normal approximation.

that is, if $\hat{\alpha}_p$ and $\hat{\alpha}_q$ maximize the log-likelihood—say, l^* is the maximum—then one can obtain a confidence region for the pair by computing the likelihood l' for the (RE)ML estimates conditional on other values α'_p, α'_q, in the neighborhood. Then the test statistic $-2(l' - l^*)$ is known to be approximately chi-square distributed with 2 degrees of freedom. If one of α_p or α_q is considered fixed, then a confidence interval for the other is obtained by the same method via the chi-square distribution with 1 degree of freedom this time. Since we give more details on likelihood ratio tests below, we do not give further details here (see also Figure 4).

F4, AQ: 3

There are several ways to approximate the variances of the remaining parameters of which we discuss the bootstrap method first. We explored the bootstrap method for the dementia data from Mitchell (2009), using ordinary case resampling in this case. The standard deviations obtained in this fashion were used to calculate confidence intervals with a normal approximation. Besides using a normal approximation, one could use various bootstrap confidence intervals (see, e.g., Carpenter & Bithell, 2000). Bootstrapping was found feasible, but too cumbersome. The first alternative to bootstrap methods is numerical covariance matrices, given, for example, by mle; the key disadvantage is that these covariance matrices are unreliable for ill-conditioned optimization problems, which we encountered in data sets with small underlying σ. Also, when α_p and α_q are to be estimated from the data, then a numerical covariance matrix cannot be calculated if α_p and α_q are on one of the boundaries of [0, 2]. This is no problem, though, if working conditional on α_p and α_q. For these parameters profiling the likelihood is nevertheless feasible. The third alternative for the variance components is the approximate formulae for the first model in Equation 12.

The performance of the estimation process was evaluated with the help of simulation experiments. We report these in the supplemental materials to this article. The simulations, especially for small sample size, underline the importance of the REML correction in the estimation of the variance components.

The Presence of a Cutoff Value Variation as a Likelihood Ratio Test

In a meta-analysis of a diagnostic test, we can expect different authors to use different cutoff values. This intrinsic diffi-

Table 2
Data from Mitchell (2009) on the Mini-Mental State Examination

Study	Condition	True positive	False negative	False positive	True negative
1	Dementia	65	3	240	870
2	Dementia	117	12	10	110
3	Dementia	48	19	63	989
4	Dementia	134	8	28	152
5	Dementia	24	5	44	292
6	Dementia	67	15	48	153
7	Dementia	64	17	0	71
8	Dementia	281	64	20	286
9	Dementia	13	1	44	286
10	Dementia	262	20	29	177
11	Dementia	143	18	29	123
12	Dementia	183	33	33	51
13	Dementia	22	0	152	140
14	Dementia	112	0	590	2,091
15	Dementia	152	81	126	1,009
16	Dementia	29	26	26	236
17	Dementia	31	6	3	247
18	Dementia	10	3	12	333
19	Dementia	707	88	1438	10,447
20	Dementia	181	108	17	184
21	Dementia	59	29	23	74
22	Dementia	74	23	16	143
23	Dementia	27	12	26	209
24	Dementia	40	6	75	528
25	Dementia	317	52	173	578
26	Dementia	387	116	16	54
27	Dementia	118	65	1	44
28	Dementia	44	7	34	396
29	Dementia	123	46	98	309
30	Dementia	25	43	3	171
31	Dementia	73	32	2	225
32	Dementia	37	45	0	440
33	Dementia	78	34	45	376
34	MCI	72	12	53	214
35	MCI	106	23	410	379
36	MCI	37	36	22	118
37	MCI	67	30	22	75
38	MCI	17	77	0	90

Note. MCI = mild cognitive impairment.

tapraid5/met-met/met-met/met00212/met2187d12z | xppws | S=1 | 4/17/12 | 12:56 | Art: 2010-0161

Table 3
Meta-Analysis for Mini-Mental State Examination Dementia Data

	First model			Second model		
Parameter	*M*	*SD*	95% CI	*M*	*SD*	95% CI
μ_1	0.1220	0.1113	[−0.0961, 0.3401]	1.4161	0.1567	[1.1090, 1.7233]
μ_2	−0.3174	0.0923	[−0.4983, −0.1364]	−1.7707	0.1637	[−2.0916, −1.4498]
σ	0.1911	0.0678	[0.0583, 0.3239]	0.4360	0.1704	[0.1020, 0.7700]
σ_1^2	0.4087	0.1006	[0.2115, 0.6058]	0.6889	0.2122	[0.2730, 1.1047]
σ_2^2	0.2813	0.0693	[0.1456, 0.4170]	0.7961	0.2438	[0.3183, 1.2738]
α_p [a]	1.6746		[1.2062, 1.8975]	0.9544		[0.0000, 1.8311]
α_q [a]	0.2438		[0.0914, 0.6380]	0.8696		[0.1730, 2.0000]
$t_{\alpha_p}^{-1}(\mu_1)$ [b]	0.7924		[0.7314, 0.8488]	0.7915		[0.7377, 0.8365]
$t_{\alpha_p}^{-1}(\mu_2)$ [b]	0.1139		[0.0743, 0.1612]	0.1119		[0.0809, 0.1523]

Note. CI = confidence interval.
[a] No standard deviations for α_p and α_q were computed, since CIs were obtained from the highly asymmetric profile likelihood (see Figure 4). [b] $t_{\alpha_p}^{-1}(\mu_1)$ and $t_{\alpha_q}^{-1}(\mu_2)$ are given for pooled sensitivities and false-positive rates; since these are nonlinear transformations of μ_1 and μ_2, no standard deviation is calculated.

culty of such a meta-analysis, the cutoff value problem, is incorporated in both models by the covariance σ of the (t_α-transformed) false-positive rate and sensitivity. Nevertheless, an obvious question is whether the cutoff value problem is a relevant factor in the meta-analysis at hand. This question can be reformulated by asking if σ is (close to) 0. Note that the following answer to this question also applies to models beyond ours. The answer is provided by a likelihood ratio test that compares the fit of the model with a model with σ restricted to 0. Our null model is this restricted model; that is, we test our null-hypothesis $H_0 : \sigma = 0$ against $H_1 : \sigma \neq 0$. Note that no boundary problem exists here as it typically occurs, for example, when testing variance components being zero; that is, when $H_0 : \sigma_1^2 = 0$, the tested value is on the boundary (Greven, Crainiceanu, Küchenhoff, & Peters, 2008).

For the first model, one can proceed as follows: First, recall the fact that if the true σ equals 0, then the false-positive rate and the sensitivity are independent variables. Hence in this case we can fit sensitivity and false-positive rate data separately to normal models; that is, the transformed sensitivity is $N(\mu_1, \sigma_1^2)$ distributed, and the transformed false-positive rate is $N(\mu_2, \sigma_2^2)$ distributed. So if we restrict σ to 0, we have an easy way of fitting this restricted model. Let $(\hat{\alpha}_p, \hat{\alpha}_q, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ be the (RE)ML estimates of the parameters for the (full) first model and let $(\tilde{\alpha}_p, \tilde{\alpha}_q, \tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2)$ denote the (RE)ML estimates of the parameters of the first model with $\sigma = 0$. Below l denotes the log-likelihood of the model as in Equation 10. It is well known that the statistic

$$D = -2[l(\hat{p}_1, \ldots, \hat{q}_i, \ldots | \tilde{\alpha}_p, \tilde{\alpha}_q, \tilde{\mu}_1, \tilde{\mu}_2, 0, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2) - l(\hat{p}_1, \ldots, \hat{q}_i, \ldots | \hat{\alpha}_p, \hat{\alpha}_q, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)]$$

is asymptotically (i.e., for large N) chi-square distributed with 1 degree of freedom (see, e.g., Shao, 2003). One now compares D to the quantiles of the chi-square distributions or calculates p values as usual. One rejects the null model for high values of D.

For the second model, we use the same idea and notation: Along the lines of the case for the first model, one straightforwardly obtains a model for σ restricted to 0. In our calculations we used

Table 4
Meta-Analysis for Mini-Mental State Examination Mild Cognitive Impairment Data

	First model			Second model		
Parameter	*M*	*SD*	95% CI	*M*	*SD*	95% CI
μ_1	−0.0063	0.5771	[−1.1373, 1.1248]	0.6739	0.5707	[−0.4445, 1.7924]
μ_2	−0.9630	0.6277	[−2.1932, 0.2672]	−3.6494	1.0487	[−5.7048, −1.5940]
σ	1.6419	1.0932	[−0.5008, 3.7846]	2.4773	2.0478	[−1.5363, 6.4909]
σ_1^2	1.6651	1.0531	[0.0000, 3.7291][c]	1.5621	1.1533	[0.0000, 3.8226][c]
σ_2^2	1.9698	1.2458	[0.0000, 4.4115][c]	4.7749	4.5880	[0.0000, 13.7671][c]
α_p [a]	1.2990		[0.0000, 2.0000]	0.8934		[0.0000, 2.0000]
α_q [a]	0.6249		[0.0676, 2.0000]	2.0000		[0.0899, 2.0000]
$t_{\alpha_p}^{-1}(\mu_1)$ [b]	0.6038		[0.3345, 0.8510]	0.6269		[0.3539, 0.8297]
$t_{\alpha_p}^{-1}(\mu_2)$ [b]	0.1498		[0.0281, 0.4356]	0.1613		[0.0577, 0.4507]

Note. CI = confidence interval.
[a] No standard deviations for α_p and α_q were computed, since CIs were obtained from the highly asymmetric profile likelihood (see Figure 4). [b] $t_{\alpha_p}^{-1}(\mu_1)$ and $t_{\alpha_q}^{-1}(\mu_2)$ are given for pooled sensitivities and false-positive rates; since these are nonlinear transformations of μ_1 and μ_2, no standard deviation is calculated. [c] CIs for variance components had to be truncated at 0.

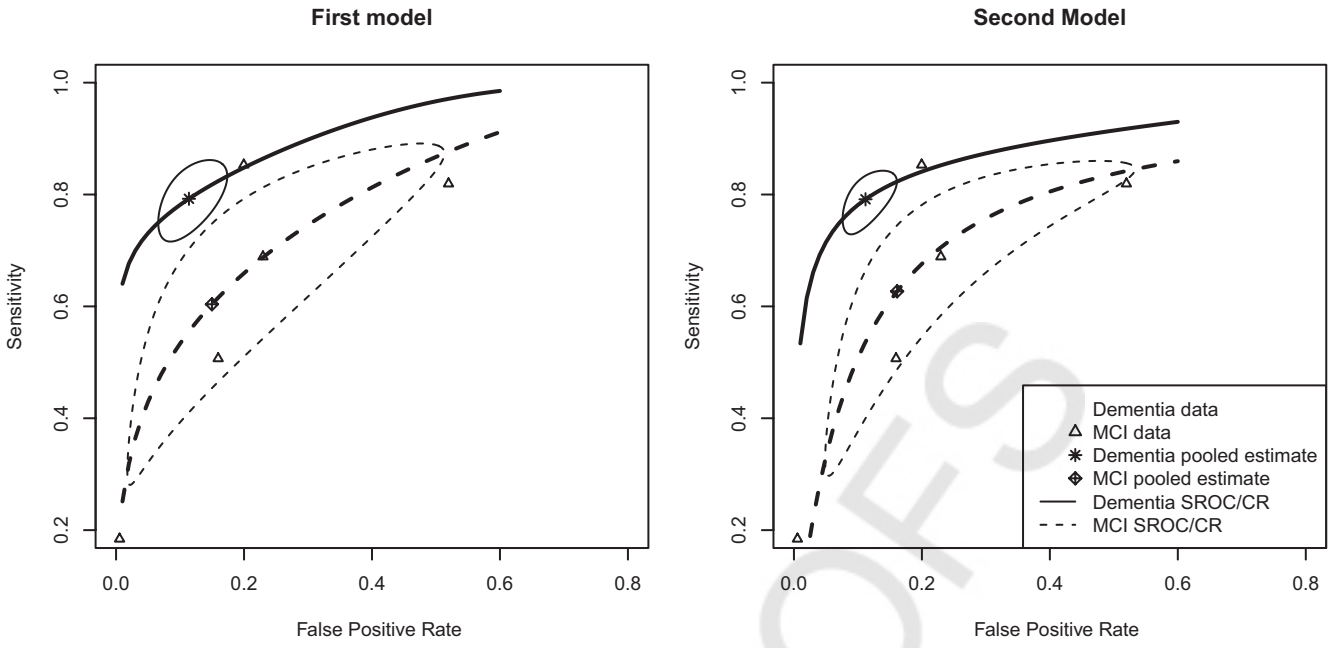

Figure 5. Data, summary receiver operating characteristic (SROC) curves, and pooled diagnostic accuracy for the Mini-Mental State Examination (MMSE) meta-analysis. Pooled estimates computed as $(t_{\alpha_p}{}^{-1}(\mu_1), t_{\alpha_q}{}^{-1}(\mu_1))$, where μ_1 and μ_2 are from Tables 3 and 4, respectively. Confidence regions (CRs) obtained by applying the appropriate inverse transformation to a 95% confidence ellipsoid on t_α space (see Harbord et al., 2007). Data are from Mitchell (2009). SROC curves were truncated to avoid extrapolation beyond the data. Graphical comparison of the SROC curves and CRs suggests that the MMSE is suited better to diagnose dementia than it is to diagnose MCI; note, though, that the mild cognitive impairment (MCI) data set is rather small ($N = 5$). The CRs for the diagnostic accuracy are smaller in the second model, since the random effects explain part of the variation. The SROC curves are less dependent on outliers. It is also noteworthy that though the estimates of α_p and α_q are far from each other, the SROC curves are surprisingly similar.

the resulting likelihood and the mle function of R to fit such a model. Again, one calculates the statistic

$$D = -2[l(\hat{p}_1, \ldots, \hat{q}_i, \ldots | \tilde{\mu}_1, \tilde{\mu}_2, 0, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2) - l(\hat{p}_1, \ldots, \hat{q}_i, \ldots | \tilde{\mu}_1, \tilde{\mu}_2, \hat{\sigma}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)]$$

where l is the log-likelihood or adjusted log-likelihood of the second model.

Apart from these likelihood ratio tests, there is another way to test $\sigma = 0$ by using Pearson's r. For this one calculates Pearson's r (also known as the Pearson product–moment correlation coefficient) from the transformed observed sensitivities and false-positive rates. As in our first model, a bivariate normal distribution is assumed on the t_α space. If $\sigma = 0$, then $\sqrt{N-3}F(r)$ is approximately standard normal, where F is the Fisher transformation. One can now calculate p values or a (two-sided) confidence interval as usual (see, e.g., Anderson, 2003). This approach, though, does not take into account the random effects of the second model.

Worked Examples

In the following examples we showcase the use of the model with both transformation parameters as free parameters. In applications it might be more appropriate to fix one of α_p or α_q, or even both, especially when sample size is small. The parameters of the first model in the following examples have been estimated with $\bar{\mu}$ and Q with a continuity correction where necessary. For the second model, we used the REML estimates, again adding a continuity correction to all cells if zero cells were present. A customary value for such a correction is .5; when estimating an odds ratio this reduces bias from small cells (see, e.g., Gart & Zweifel, 1967; Sutton et al., 2000). In our situation we found that it leads to overly large random effects in our second model. These random effects with known variances are based on the variance approximation (Equation 6) derived by the delta method, that is, based on a second-order Taylor approximation. Since Equation 6 approaches ∞ for p near 0 or 1, the random effects become large if zero cells are present and a small continuity correction is used. The resulting confidence intervals on t_α space were found unreasonably large when compared to transformed Wilson score intervals for binomial proportions. We used a continuity correction of 1 instead.

MMSE for Dementia and MCI

We continue the first example from the introduction. In the meta-analysis (Mitchell, 2009) the MMSE was surveyed as a diagnostic test for dementia and MCI. Table 2 shows the original data; note that we excluded one study, since we could not calculate **T2, AQ: 4**

Table 5
Data from Patrick et al. (1994) on the Validity of Self-Reported Smoking, Self-Administered Questionnaire Subset

Study	True positive	False negative	False positive	True negative
1	21	15	28	324
2	90	10	120	969
3	104	8	26	232
4	333	18	92	673
5	3	0	2	77
6	437	23	78	901
7	23	13	18	333
8	350	35	77	155
9	397	34	32	154
10	3	45	6	198
11	59	5	5	227
12	27	47	25	1,233
13	81	7	5	170
14	103	5	13	169
15	81	24	7	213
16	103	8	11	209
17	120	5	54	329
18	21	25	22	177
19	15	0	35	198
20	17	2	27	200
21	120	1	39	148
22	132	52	110	889
23	158	8	95	960
24	55	0	32	180
25	57	1	29	180
26	163	3	24	178
27	177	4	9	178
28	141	3	45	180
29	25	11	3	40
30	56	33	2	49
31	19	2	1	96

the frequencies of false positives, true positives, true negatives, and false negatives.

We reanalyzed the dementia part of the data as well as the MCI part of the data using REML estimation. The parameters α_p and α_q were determined by maximizing the adjusted profile log-likelihood and then considering α_p and α_q fixed for the purpose of calculation of standard errors. It should be noted that the MCI data contain only five studies. For the second model, we used a numeric covariance matrix obtained from mle to obtain estimates for the standard deviations of the parameters; we preferred this over a bootstrap because the numerical maximization of the likelihood was unstable for some of the bootstrap samples. Also, the admissibility of the bootstrap method is questionable for the small sample size of the MCI data.

T3,T4 Tables 3 and 4 show the parameters and their standard deviations for both models for the dementia and MCI data, respectively.

F5 Figure 5 shows the SROC curves for both applications of the MMSE and the empirical sensitivities and false-positive rates. Figure 4 shows confidence regions for α_p and α_q for the first model. The main conclusion that can be drawn from Figure 5 is that the MMSE is an appropriate instrument to screen for dementia and even better suited to screen for the absence of it, whereas the MMSE seems to be unsuitable to detect MCI.

The cutoff value problem in the MMSE data. We first discuss the cutoff value problem for the dementia data. The second model with σ restricted to 0 was fitted to the dementia data, yielding $\mu_1 = 2.3705$, $\mu_2 = -1.9429$, $\sigma_1^2 = 1.0037$, $\sigma_2^2 = 0.8779$, $\alpha_p = 0.4306$, and $\alpha_q = 0.9461$. This results in a log-likelihood for the restricted parameter space of 54.6608 compared with 59.5678 for the second model with all five parameters. So for the dementia part of the data, the likelihood ratio test statistic for the cutoff problem (second model) is $D = 9.8140$. One can compare this to 3.8415, the 95% quantile of the chi-square distribution, or compute the p value, which is less than 10^{-2}. Either way, it is clear that the sensitivities and false-positive rates are (positively) correlated, since we reject the null hypothesis that $\sigma = 0$. In fact, Mitchell (2009) reported various cutoff values for the primary studies. For the MCI data the result is similar ($D = 5.0768$).

Comparison of inferences from the HSROC. The HSROC was fitted to both data sets. The pooled pair of sensitivity and false-positive rate was (0.7910, 0.1113) for the dementia data and (0.6212, 0.1658) for the MCI data—both pairs very similar to our second model—and the confidence region was similar for the dementia data but not for the MCI data. SROC curves were inspected for the HSROC and compared to the SROCs of the second model. For the center of the data, the SROC curves agreed, but for false-positive rates greater than .5, the SROC curve of the HSROC predicted large sensitivities for the MCI data. The similarities of confidence region and SROC curve for the dementia data is not surprising, since the point estimates of α_p and α_q are close to 1, the value that yields the HSROC.

Summary of findings. In sum, the MMSE has greater diagnostic power with respect to dementia than it has with respect to MCI; this conclusion is in line with Mitchell (2009). Using the SROC curves and confidence regions to graphically compare the two subsets of the MMSE data reinforces this conclusion; in fact, the MCI SROC curve is uniformly below the dementia SROC curve for both models.

Table 6
Data from Patrick et al. (1994) on the Validity of Self-Reported Smoking, Interviewer-Administered Questionnaire Subset

Study	True positive	False negative	False positive	True negative
32	380	38	10	854
33	480	72	46	1,078
34	312	24	46	594
35	28	4	2	164
36	346	4	14	78
37	336	4	26	76
38	150	8	16	370
39	214	48	2	76
40	206	28	10	96
41	214	0	2	20
42	188	6	30	14
43	208	0	10	20
44	126	12	0	182
45	116	8	2	180
46	78	2	2	44
47	84	2	12	44
48	72	4	8	42
49	76	4	18	42
50	1,358	186	68	3,322
51	1,650	18	424	6,632

tapraid5/met-met/met-met/met00212/met2187d12z | xppws | S=1 | 4/17/12 | 12:56 | Art: 2010-0161

Table 7
Meta-Analysis for Smoking Data

	Complete data			SAQ subset			IAQ subset		
Parameter	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI
μ_1	4.388	0.289	[3.821, 4.954]	3.508	0.353	[2.816, 4.199]	5.923	0.435	[5.071, 6.775]
μ_2	−4.949	0.252	[−5.443, −4.456]	−4.851	0.244	[−5.329, −4.373]	−4.628	0.505	[−5.617, −3.638]
σ_1^2	3.705	0.846	[2.047, 5.362]	3.374	0.972	[1.469, 5.278]	3.095	1.241	[0.663, 5.526]
σ_2^2	2.855	0.647	[1.586, 4.124]	1.553	0.473	[0.625, 2.480]	4.642	1.661	[1.387, 7.897]
σ	1.570	0.563	[0.466, 2.674]	1.281	0.530	[0.241, 2.320]	2.315	1.129	[0.103, 4.528]
α_p[a]	0.234		[0.000, 0.908]	0.358		[0.000, 1.099]	0.000		[0.000, 1.965]
α_q[a]	2.000		[0.655, 2.000]	2.000		[0.339, 2.000]	1.819		[0.493, 2.000]
$t_{\alpha_p}^{-1}(\mu_1)$[b]	0.918		[0.887, 0.940]	0.885		[0.827, 0.924]	0.948		[0.921, 0.966]
$t_{\alpha_q}^{-1}(\mu_1)$	0.084		[0.066, 0.108]	0.088		[0.070, 0.112]	0.078		[0.045,0.1330]

(μ_2)[b]

Note. SAQ = self-administered questionnaire; IAQ = interviewer-administered questionnaire; CI = confidence interval.
[a] No standard deviations for α_p and α_q were computed, since CIs were obtained from the highly asymmetric profile likelihood. [b] $t_{\alpha_p}^{-1}(\mu_1)$ and $t_{\alpha_q}^{-1}(\mu_2)$ are given for pooled sensitivities and false-positive rates; since these are nonlinear transformations of μ_1 and μ_2, no standard deviation is calculated.

Diagnosing Smoking: Comparing Self-Administered and Interviewer-Administered Questionnaires

The second example is based on data from the meta-analysis of Patrick et al. (1994), which examines the accuracy of self-reported measures of smoking. We will focus on the question whether SAQ or IAQ is more reliable; Tables 5 and 6 show the original data. We will only present an analysis based on the second model, since inferences from the first model are very similar. Again, REML estimation was preferred over ML, even though the sample sizes in this example are adequate for ML.

T5,T6

Table 7 shows the parameters and their standard deviations for the complete data and the SAQ and IAQ subsets. Figure 6 shows profile plots for α_p and α_q for the complete smoking data and also the SROC curves for both subsets of the data. The main conclusion that can be drawn from Figure 6B and Table 7 is that self-reported measures of smoking are reliable, regardless of whether SAQ or IAQ is used. The second conclusion is that the confidence region for the point estimate of the diagnostic accuracy parameters for the IAQ data as well as the SROC curves implies that IAQ is more reliable than SAQ. The 95% confidence regions for the pairs of sensitivity and false-positive rate

T7 F6

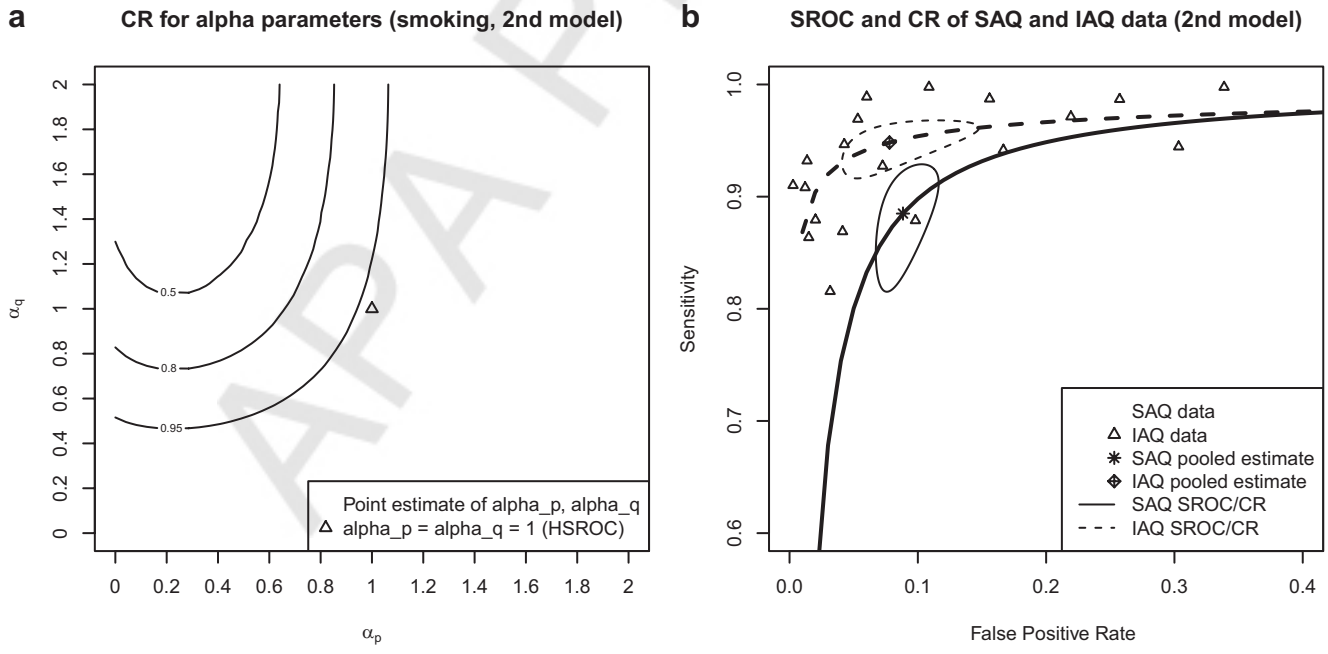

Figure 6. Smoking meta-analysis: transformation parameters and comparison of self-administered questionnaires (SAQ) and interviewer-administered questionnaires (IAQ). CR = confidence region; SROC = summary receiver operating characteristic; HSROC = hierarchical summary receiver operating characteristic.

from SAQ and IAQ do not overlap, so that this rigorous conclusion can be drawn from Figure 6.

The cutoff value problem in the smoking data. The likelihood ratio test for the cutoff value problem was calculated for the complete smoking data, the SAQ and IAQ subsets. All three statistics were significant ($p < .01$ in all three cases), leading to the conclusion that some sort of cutoff value problem is present in the smoking data. In fact, since data from comparable but not identical questionnaires were examined by Patrick et al. (1994), the wording in each primary study differed and was in many cases not reported. Nevertheless, we are reluctant to conclude that difference in wording explains the significance of the likelihood ratio test, since Patrick et al. also reported that a biochemical gold standard procedure used in some of the primary studies led to increased false-positive rates (i.e., people self-reported to be smokers, but the gold standard test was negative).

Comparison of inferences from the HSROC. The contours shown in Figure 6A are based on an inverted likelihood ratio test for the hypothesis $H_0 : \alpha_p = \alpha_q = 1$. Since H_0 amounts to the HSROC, the likelihood ratio test rejects the HSROC for the smoking data. For the SAQ and IAQ subsets, though, the same test is not significant at 5%. Point estimates from the HSROC as well as SROC curves are comparable in all three cases. The reason for the superior fit of the second model is the nonnormality of the observed diagnostic accuracy: Patrick et al. (1994) reported that sensitivities and specificities are negatively skewed; it is mentioned that a log transformation of both diagnostic accuracies led to approximate normality. One conclusion from the HSROC, though, was not in line with the findings from our second model: The confidence regions for the mean parameters overlap for the HSROC, so the rigorous conclusion drawn from the second model is not possible.

Summary of findings. The main conclusion drawn in Patrick et al. (1994) is reinforced by our analysis: Self-reported measures of smoking are reliable (pooled sensitivity = 0.918, pooled false-positive rate = 0.084). Also, the observation that the use of IAQ leads to improved accuracy compared with SAQ is made by Patrick et al., but the logistic fixed-effect model used in the original article only allows to conclude that the false-positive rates are significantly better for IAQ, not the sensitivities. The refined analysis building on the second model and the transformed confidence ellipses enable us to report a significant improved accuracy of IAQ. We did not control for the additional covariates, though, like Patrick et al. did in their original analysis.

The Models in Context

Comparison of the Two Models

The models we propose share a key advantage with existing bivariate models: Our approach is very natural in this symmetric problem; that is, we could have exchanged the roles of the sensitivity and false-positive rate and nevertheless obtained the same SROC curve. One difference between the two models is that σ_1 and σ_2 can be expected to be smaller in the second model due to the variance explained by the binomial model assumed (i.e., Equation 6). In this sense the first model overestimates the heterogeneity among the studies. The confidence regions of the pooled diagnostic accuracy for the second model are somewhat smaller. Note that the area of the confidence regions, which are confidence ellipses on t_α space, depends solely on the estimate of Σ, the parameters used to describe the heterogeneity. The confidence regions depicted in Figure 2 show that this overestimation is slight for the MMSE data. In general, the reduction would be considerable if the primary studies were very small, which is unlikely in a diagnostic meta-analysis. For all data we reanalyzed (see Table 8), the estimates of diagnostic accuracy from the two models were rather close; also, the SROC curves were similar, as in Figure 5. T8, AQ: 5

The random effects of the second model widen the confidence intervals for α_p and α_q considerably, a wide range being plausible for the MMSE data. This is very much in contrast to the first

Table 8
Comparison of Fit of the Second Model With Free and Fixed α_p, α_q, and the Hierarchical Summary Receiver Operating Characteristic ($\alpha_p = \alpha_q = 1$) for Various Data Sets

α_p	α_q	AIC
MMSE: Dementia ($N = 33$)		
Free	Free	−105.1
1	1	*−109.1*
0.6	0.6	*−108.8*
Free	1	−107.1
1	Free	−107.1
1	0.6	*−109.0*
MMSE: MCI ($N = 5$)		
Free	Free	−0.8
1	1	*−4.5*
1	2	*−4.8*
0.6	2	*−4.7*
1	Free	−2.8
AUDIT−C ($N = 14$)		
Free	Free	−48.0
1	1	*−50.7*
1.4	1.4	*−51.3*
Smoking ($N = 51$)		
Free	Free	−230.2
Free	1	−230.5
1	Free	−227.1
1	1	−227.4
0	2	*−233.5*
0.6	2	*−233.0*
Smoking: IAQ ($N = 20$)		
Free	Free	−102.8
Free	1	−103.9
1	1	−105.3
0	1.4	*−106.1*
0	2	*−106.7*
Smoking: SAQ ($N = 31$)		
Free	Free	−133.6
1	1	−133.6
0	2	*−136.4*
0.6	2	*−137.2*
1	2	−134.8
0.6	1.4	*−136.6*

Note. Mini-Mental State Examination (MMSE) for dementia and mild cognitive impairment (MCI) from Mitchell (2009); Alcohol Use Disorders Identification Test (AUDIT-C) from Kriston et al. (2008); smoking for interviewer-administered questionnaire (IAQ) and self-administered questionnaire (SAQ) from Patrick et al. (1994). N = number of primary studies in this data set; AIC = Akaike information criterion.

Table 9
Fitting the First Model With α_q Restricted to 1 to Real-World Data

	Free α_p and $\alpha_q = 1$				$\alpha_p = \alpha_q = 1$		
Data set (N)	α_p	95% CI for α_p	AIC	p_{Shapiro}	AIC	p_{LR}	p_{Shapiro}
Dementia (33)	1.711	[1.283, 1.915]	−98.3	0.689	−93.2	0.008	0.021
AUDIT-C (14)	1.869	[1.522, 1.971]	−55.9	0.981	−50.3	0.006	0.050
SAQ (31)	0.424	[0.000, 1.056]	−138.2	0.719	−137.1	0.079	0.109
IAQ (20)	1.921	[1.586, 1.984]	−112.3	0.867	−108.0	0.012	0.041

Note. Mini-Mental State Examination for dementia from Mitchell (2009); Alcohol Use Disorders Identification Test (AUDIT-C) from Kriston et al. (2008); smoking for interviewer-administered questionnaire (IAQ) and self-administered questionnaire (SAQ) from Patrick et al. (1994). N = number of primary studies in this data set; CI = confidence interval; AIC = Akaike information criterion; p_{Shapiro} = p value of Shapiro test for normality of transformed sensitivities; p_{LR} = p value of likelihood ratio test for H_0: $\alpha_p = 1$.

model; here the confidence intervals are often small enough not to contain 1, the value that amounts to the logit (see Table 9). The increased uncertainty of α_p and α_q due to random effects can be explained as follows: The random effects shift some of the observed pairs of sensitivities and false-positive rates toward the mean. Since the function $p \mapsto t_\alpha(p)$ is approximately linear for many values of α and $p \in [0.1, 0.9]$, estimation of α is most precise when information about the most extreme values is available, that is, p in the intervals [0, 0.1] and [0.9, 1]. The shift toward the mean hence makes the parameter of the transformation harder to estimate, so confidence intervals are wider.

All in all, the advantage of the first model is its simple parameter estimation; the advantage of the second model is the refined estimation of the heterogeneity parameters by random effects.

Comparison With the HSROC

The main difference between the HSROC and the models we propose is the choice of the transformation. We compare the HSROC with our second model, which contains the HSROC as a special case ($\alpha_p = \alpha_q = 1$). First, we discuss SROC curves and normality assumptions, then we compare the model fit.

The SROC curve in the t_α space (i.e., Equation 3) is a linear function, since it is the conditional expectation of the bivariate normal distribution; the same holds for the SROC curve in the logit space. For the first model, the conditional expectation of the bivariate normal distribution coincides with the ordinary least squares regression line (see, e.g., DasGupta, 2010, Theorem 12.6); this is a general fact not related to SROC curves. In our setup this means that the linear regression line of the t_{α_p}-transformed sensitivities on the t_{α_q}-transformed false-positive rates is identical to the t_α-transformed SROC curve. This also means that our SROC curves, and more generally all our inferences, depend on the normality assumption of linear models (first model) and LMM (second model), respectively. For LMMs, independent normal errors of the fixed and random effects are assumed, leading to a normal variation in the whole sample; in our situation, this boils down to assuming that the transformed sensitivities and false-positive rates are assumed to be normal in both models.[1] Normality of a data set can be tested, a well-known omnibus test for departures from normality being the Shapiro Test. Table 9 shows that the normality assumption for the logit-transformed sensitivities is frequently violated in real-world data sets and that the first model, even when α_q is restricted to 1, outperforms a bivariate normal model for the logit-transformed sensitivities and false-positive rates. The proper use of transformations like t_α ensures that normality assumptions of models are met.

We compare the fit of models with Akaike's information criterion (AIC) given by

$$\text{AIC} := 2k - 2\ln(L),$$

where k is the number of parameters and $\ln(L)$ is the log-likelihood of the model. Smaller values of the AIC indicate a better model. The HSROC has five parameters, two for the vector of means μ and three for the covariance matrix Σ; the comparison based on the AIC and our two models has two more parameters (α_p and α_q). If the transformation parameters are considered free, they have to be taken into account when calculating the AIC. In some of the model variants we look at below, we fix both αs at natural values, reducing the number of parameters in AIC calculations by one or two. Values we considered are 0, 0.6, 1, 1.4, and 2; here 0.6 and 1.4 are chosen, since t_α approximates the (complementary) log-log transformation for these values. Also, note that the log-likelihoods are frequently positive, mainly due to the contribution from the Jacobian, resulting in negative AIC.

For our comparison, we fitted our second model and the HSROC to various data from diagnostic meta-analyses from psychology. Many of these data sets provide more than just the sensitivities and specificities or the frequencies of true positives, false positives, true negatives, and false negatives; our point here is not to reanalyze them in detail, but to show how the log-likelihood behaves for a variety of data. Apart from the data discussed in the examples from Mitchell (2009) and Patrick et al. (1994), we studied the data from Kriston, Hölzel, Weiser, Berner, and Härter (2008), a meta-analysis of the Alcohol Use Disorders Identification Test (AUDIT-C), a short version of the AUDIT, an established test to detect unhealthy alcohol use. Tables containing the frequencies for all the data sets are part of the supplemental materials to this article.

In Table 8 we report the values of the AIC statistic. With respect to AIC data, sets with no or few zero cells or moderate sensitivities in the majority of primary studies (MMSE, AUDIT-C) seem not to

[1] Cramér's Theorem states that if ε and δ are independent random variables and $\delta + \varepsilon$ is normal, then ε and δ are normal. Gurka et al. (2006) argued that to ensure that the fixed and random effects in an LMM are normal, one can focus to check the normality of the (transformed) outcome variable, that is, on a transformation to normality.

profit from using the second model with free parameters. The smoking data set and its subsets SAQ and IAQ (with many of the primary studies reporting high sensitivity and low false-positive rate) profit substantially in terms of model fit. Since the HSROC is nested within the second model, one can also compute a likelihood ratio test for $H_0 : \alpha_p = \alpha_q = 1$. This test is significant at a 5% confidence level for the smoking data ($p = .034$).

Discussion

It could be argued that our models have the following shortcoming: If one of α_p or α_q equals 0 or 2, then the inverse transformation is $1 - \exp(-x/2)$ or $\exp(x/2)$, respectively; since the SROC curve is a straight line on t_α space (i.e., $\mathbb{R}^2$), it can well take values outside the unit square $[0, 1]^2$; that is, if the factor $\exp(\mu_1 - \frac{\sigma}{\sigma_2^2}\mu_2)$ in Equation 5 is greater than 1, then values of the SROC curve can lie outside the unit square. For example, for second model for the MCI data, the estimate is $\alpha_q = 2$ (see Table 4). Nevertheless, if we restrict the domain of the SROC curve to the interval $[\min\{\hat{q}_i; i = 1, \ldots, N\}, \max\{\hat{q}_i; i = 1, \ldots, N\}]$ (i.e., if we only plot the SROC curve where there are observed false-positive rates), then the curve is plausible for the MCI data. This restriction is reasonable in general to avoid extrapolation beyond the data. If the aim is to draw an SROC on the whole of the unit cube, then one could also truncate the SROC at 1.

One potential way of dealing with this minor defect of the SROC curves would be to use a truncated normal distribution on the log space, that is, to restrict the bivariate normal distribution on the log space to $[-\infty, 0]^2$. Using the package tmvtnorm (Wilhelm & Manjunath, 2010b), we fitted such a model to several data sets. We encountered well-known problems with MLE of the truncated normal distribution (Wilhelm & Manjunath, 2010a) but nevertheless obtained ML estimates. From our somewhat limited exploration, it seems that using the truncated normal distribution is rather detrimental than beneficial with respect to the shape of the SROC curve; in fact, we are not aware of a closed-form expression of the SROC curve, so one relies on Monte Carlo SROC curves.

Our work could be generalized by implementing a meta-regression on the parameters μ_1, μ_2; that is, if additional information is available on the primary studies—say, the type of population, the study setting, the cutoff value, or the gold standard procedure—then it is possible to include these as covariates for μ. This helps to explain the variation among the primary studies.

An alternative to t_α is given by the Guerrero and Johnson (1982) transformation

$$x \mapsto \begin{cases} \left(\left(\frac{x}{1-x}\right)^{\phi} - 1\right)/\phi & (\phi \neq 0) \\ \operatorname{logit}(x) & (\phi = 0) \end{cases}, \qquad (16)$$

which is obtained from the Box and Cox (1964) transformation by substituting $\frac{x}{1-x}$ for x. This allows for a transformation to quasinormality like t_α; since Equation 16 is bounded for a wide range of ϕ, we deemed it less suitable for this purpose. Gurka, Edwards, Muller, and Kupper (2006) discussed use of the Box–Cox transformations in LMMs, and since Equation 16 is closely related to this transformation, we assume that it is a feasible alternative.

Note that all R code that was written for this publication is available from the first author upon request. We plan to compile an R package with functions for diagnostic meta-analysis.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). Hoboken, NJ: Wiley-Interscience.

Arends, L. R., Hamza, T. H., van Houwelingen, J. C., Heijenbrok-Kal, M. H., Hunink, M. G. M., & Stijnen, T. (2008). Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making, 28,* 621–638. doi:10.1177/0272989X08319957

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine, 14,* 395–411. doi:10.1002/sim.4780140406

Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association, 76,* 296–311. doi:10.2307/2287831

Böhning, D., Böhning, W., & Holling, H. (2008). Revisiting Youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. *Statistical Methods in Medical Research, 17,* 543–554. doi:10.1177/0962280207081867

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B. Methodological, 26,* 211–252.

Bredemeier, K., Spielberg, J. M., Silton, R. L., Berenbaum, H., Heller, W., & Miller, G. A. (2010). Screening for depressive disorders using the Mood and Anxiety Symptoms Questionnaire Anhedonic Depression Scale: A receiver-operating characteristic analysis. *Psychological Assessment, 22,* 702–710. doi:10.1037/a0019915

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine, 19,* 1141–1164. doi:10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

Chu, H., Guo, H., & Zhou, Y. (2010). Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Medical Decision Making, 30,* 499–508. doi:10.1177/0272989X09353452

Cornell, D. G., Peterson, C. S., & Richards, H. (1999). Anger as a predictor of aggression among incarcerated adolescents. *Journal of Consulting and Clinical Psychology, 67,* 108–115. doi:10.1037/0022-006X.67.1.108

Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *Journal of the American Medical Association, 269,* 2386–2391. doi:10.1001/jama.1993.03500180078038

DasGupta, A. (2010). *Fundamentals of probability: A first course.* New York, NY: Springer.

Egger, M., Smith, G. D., & Altman, D. G. (Eds.). (2001). *Systematic reviews in health care: Meta-analysis in context* (2nd ed.). Hoboken, NJ: Wiley-Interscience. doi:10.1002/9780470693926

Folstein, M., Folstein, S., & McHugh, P. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189–198. doi:10.1016/0022-3956(75)90026-6

Gart, J. J., & Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika, 54,* 181–187. doi:10.1093/biomet/54.1-2.181

Gatsonis, C., & Paliwal, P. (2006). Meta-analysis of diagnostic and screening test accuracy evaluations: Methodologic primer. *American Journal of Roentgenology, 187,* 271–281. doi:10.2214/AJR.06.0226

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2010). mvtnorm: Multivariate normal and *t* distributions (R package Version 0.9-92) [Computer software]. AQ: 6

Greven, S., Crainiceanu, C. M., Küchenhoff, H., & Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics, 17,* 870–891. doi:10.1198/106186008X386599

Grigoletto, F., Zappalà, G., Anderson, D. W., & Lebowitz, B. D. (1999). Norms for the Mini-Mental State Examination in a healthy population. *Neurology, 53,* 315–320.

Guerrero, V. M., & Johnson, R. A. (1982). Use of the Box–Cox transformation with binary response models. *Biometrika, 69,* 309–314. doi:10.1093/biomet/69.2.309

Gurka, M. J., Edwards, L. J., Muller, K. E., & Kupper, L. L. (2006). Extending the Box–Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A. Statistics in Society, 169,* 273–288. doi:10.1111/j.1467-985X.2005.00391.x

Hamza, T. H., Arends, L. R., van Houwelingen, H. C., & Stijnen, T. (2009). Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Medical Research Methodology, 9,* 73–87. doi:10.1186/1471-2288-9-73

Hamza, T. H., Reitsma, J. B., & Stijnen, T. (2008). Meta-analysis of diagnostic studies: A comparison of random intercept, normal–normal, and binomial–normal bivariate summary ROC approaches. *Medical Decision Making, 28,* 639–649. doi:10.1177/0272989X08323917

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P., & Sterne, J. A. C. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics, 8,* 239–251. doi:10.1093/biostatistics/kxl004

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117,* 167–178. doi:10.1037/0033-2909.117.1.167

Holling, H., Böhning, W., & Böhning, D. (2012). Likelihood-based clustering of meta-analytic SROC curves. *Psychometrika, 77,* 106–126. doi:10.1007/s11336-011-9236-2

Jones, C. M., & Athanasiou, T. (2005). Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. *Annals of Thoracic Surgery, 79,* 16–20. doi:10.1016/j.athoracsur.2004.09.040

Kettler, R. J., & Elliott, S. N. (2010). A brief broadband system for screening children at risk for academic difficulties and poor achievement test performance: Validity evidence and applications to practice. *Journal of Applied School Psychology, 26,* 282–307. doi:10.1080/15377903.2010.518584

Kriston, L., Hölzel, L., Weiser, A.-K., Berner, M. M., & Härter, M. (2008). Meta-analysis: Are 3 questions enough to detect unhealthy alcohol use? *Annals of Internal Medicine, 149,* 879–888.

Le, C. T. (2006). A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research, 15,* 571–584. doi:10.1177/0962280206070637

Lee, Y., Nelder, J. A., & Pawitan, Y. (2006). *Generalized linear models with random effects: Unified analysis via H-likelihood.* Boca Raton, FL: Chapman & Hall/CRC.

Leeflang, M. M. G., Deeks, J. J., Gatsonis, C., & Bossuyt, P. M. M. (2008). Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine, 149,* 889–897.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York, NY: Springer.

Lehr, D., Koch, S., & Hillert, A. (2010). Where is (im)balance? Necessity and construction of evaluated cut-off points for effort–reward imbalance and overcommitment. *Journal of Occupational and Organizational Psychology, 83,* 251–261. doi:10.1348/096317909X406772

Littenberg, B., & Moses, L. E. (1993). Estimating diagnostic accuracy from multiple conflicting reports: A new meta-analytic method. *Medical Decision Making, 13,* 313–321. doi:10.1177/0272989X9301300408

Macaskill, P. (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology, 57,* 925–932. doi:10.1016/j.jclinepi.2003.12.019

Mitchell, A. J. (2009). A meta-analysis of the accuracy of the Mini-Mental State Examination in the detection of dementia and mild cognitive impairment. *Journal of Psychiatric Research, 43,* 411–431. doi:10.1016/j.jpsychires.2008.04.014

Moses, L. E., Shapiro, D., & Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine, 12,* 1293–1316. doi:10.1002/sim.4780121403

Murray, D. M., O'Connell, C. M., Schmid, L. A., & Perry, C. L. (1987). The validity of smoking self-reports by adolescents: A reexamination of the bogus pipeline procedure. *Addictive Behaviors, 12,* 7–15. doi:10.1016/0306-4603(87)90003-7

Patrick, D. L., Cheadle, A., Thompson, D. C., Diehr, P., Koepsell, T., & Kinne, S. (1994). The validity of self-reported smoking: A review and meta-analysis. *American Journal of Public Health, 84,* 1086–1093. doi:10.2105/AJPH.84.7.1086

Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association, 95,* 308–311. doi:10.2307/2669554

Pepe, M. S. (2004). *The statistical evaluation of medical tests for classification and prediction.* Oxford, England: Oxford University Press.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology, 58,* 982–990. doi:10.1016/j.jclinepi.2005.02.022

Rücker, G., & Schuhmacher, M. (2009). Letter to the editor. *Biostatistics, 10,* 806–807. doi:10.1093/biostatistics/kxp021

Rutter, C. M., & Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine, 20,* 2865–2884. doi:10.1002/sim.942

Schulze, R., Holling, H., & Böhning, D. (Eds.). (2003). *Meta-analysis: New developments and applications in medical and social sciences.* Cambridge, MA: Hogrefe & Huber.

Shao, J. (2003). *Mathematical statistics* (2nd ed.). New York, NY: Springer.

Smyth, G. K., & Verbyla, A. P. (1996). A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *Journal of the Royal Statistical Society: Series B. Methodological, 58,* 565–572.

Stillman, J. A., & Jackson, D. J. R. (2005). A detection theory approach to the evaluation of assessors in assessment centres. *Journal of Occupational and Organizational Psychology, 78,* 581–594. doi:10.1348/096317905X26147

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research.* Hoboken, NJ: Wiley-Interscience.

U.S. Department of Health and Human Services. (1990). *The health benefits of smoking cessation: A report of the Surgeon General* (DHHS Publication No. (CDC) 90-8416). Rockville, MD: Public Health Service.

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine, 21,* 589–624. doi:10.1002/sim.1040

Velicer, W. F., Prochaska, J. O., Rossi, J. S., & Snow, M. G. (1992). Assessing outcome in smoking cessation studies. *Psychological Bulletin, 111,* 23–41. doi:10.1037/0033-2909.111.1.23

Walter, S. D. (2002). Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Statistics in Medicine, 21,* 1237–1256. doi:10.1002/sim.1099

White, M. A., & Grilo, C. M. (2011). Diagnostic efficiency of *DSM–IV* indicators for binge eating episodes. *Journal of Consulting and Clinical Psychology, 79,* 75–83. doi:10.1037/a0022210

Wilhelm, S., & Manjunath, B. G. (2010a). tmvtnorm: A package for the truncated multivariate normal distribution. *The R Journal, 2,* 25–29.

Wilhelm, S., & Manjunath, B. G. (2010b). tmvtnorm: Truncated multivariate normal distribution (R package Version 1.0-2) [Computer software]. **AQ: 7**

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3,* 32–35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

(*Appendices follow*)

tapraid5/met-met/met-met/met00212/met2187d12z | xppws | S=1 | 4/17/12 | 12:56 | Art: 2010-0161

Appendix A

Detailed likelihoods

The following abbreviation is useful:

$$\rho = \frac{\sigma}{\sigma_1\sigma_2}.$$

Note that ρ is the correlation coefficient.

Likelihood of the First Model

We begin by stating the expanded likelihood of the first model: By our distribution assumption for $(t_{\alpha_p}(p), t_{\alpha_q}(q))$, the likelihood function for the ith study is

$$L_i(\hat{p}_i, \hat{q}_i|\alpha_p, \alpha_q, \mu, \Sigma) = \frac{J_{\alpha_p}(\hat{p}_i)J_{\alpha_q}(\hat{q}_i)}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}}\cdot$$
$$\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{t_{\alpha_p}(\hat{p}_i)-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{t_{\alpha_p}(\hat{p}_i)-\mu_1}{\sigma_1}\right)\left(\frac{t_{\alpha_q}(\hat{q}_i)-\mu_2}{\sigma_2}\right) + \left(\frac{t_{\alpha_q}(\hat{q}_i)-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

Here

$$J_\alpha(x) = \frac{\partial}{\partial x}t_\alpha(x) = \frac{\alpha}{x} - \frac{2-\alpha}{1-x}$$

is the Jacobian of t_α.

Likelihood of the Second Model

For the second model, we set

$$\sigma_{1i}^2 := d_{i1}^2 + \sigma_1^2,\ \sigma_{2i}^2 := d_{i2}^2 + \sigma_2^2,$$

and

$$\rho_i := \frac{\sigma}{\sigma_{1i}\sigma_{2i}}.$$

By our distributional assumption for $(t_{\alpha_p}(p), t_{\alpha_q}(q))$, the likelihood function for the ith study is

$$L_i(\hat{p}_i, \hat{q}_i|\mu, \Sigma_i) = \frac{J_{\alpha_p}(\hat{p}_i)J_{\alpha_q}(\hat{q}_i)}{2\pi\sigma_{1i}\sigma_{2i}(1-\rho_i^2)^{\frac{1}{2}}}\cdot$$
$$\exp\left\{\frac{-1}{2(1-\rho_i^2)}\left[\left(\frac{t_{\alpha_p}(\hat{p}_i)-\mu_1}{\sigma_{1i}}\right)^2 - 2\rho_i\left(\frac{t_{\alpha_p}(\hat{p}_i)-\mu_1}{\sigma_{1i}}\right)\left(\frac{t_{\alpha_p}(\hat{q}_i)-\mu_2}{\sigma_{2i}}\right) + \left(\frac{t_{\alpha_q}(\hat{q}_i)-\mu_2}{\sigma_{2i}}\right)^2\right]\right\}.$$

We also state the second model in matrix notation for mixed models: Let

$$Y = (t_{\alpha_p}(\hat{p}_1), t_{\alpha_q}(\hat{q}_1), t_{\alpha_p}(\hat{p}_2), t_{\alpha_q}(\hat{q}_2), \ldots, t_{\alpha_p}(\hat{p}_N), t_{\alpha_q}(\hat{q}_N))^{\mathrm{T}},$$

let δ and ε denote the vectors obtained by concatenating the δ_i and ε_i, respectively, and let

$$X = \begin{pmatrix} 1 & 0 & 1 & 0 & \ldots & 1 & 0 \\ 0 & 1 & 0 & 1 & \ldots & 0 & 1 \end{pmatrix}^{\mathrm{T}}$$

denote the $2N \times 2$ design matrix. Then the model can be stated as

$$Y = X\mu + \delta + \varepsilon.$$

With this notation, it is straightforward to check that the restricted maximum likelihood (REML) likelihood p_{REML} is just a special case of the REML likelihood obtained in the literature (e.g., in Lee et al., 2006, Section 5.2.2; Smyth & Verbyla, 1996).

(*Appendices continue*)

tapraid5/met-met/met-met/met00212/met2187d12z | xppws | S=1 | 4/17/12 | 12:56 | Art: 2010-0161

Appendix B

A Monte Carlo Summary Receiver Operating Characteristic Curve

We outline an algorithm for a Monte Carlo summary receiver operating characteristic (SROC) curve for the second model. The key steps are random sampling and using a LOWESS smoother. The random sampling can be realized as follows: Say the parameters μ_1, μ_2, σ, σ_1, and σ_2 have been estimated. Set

$$\lambda_m := \frac{1}{N}\sum_{i=1}^{N} m_i \quad \text{and} \quad \lambda_n := \frac{1}{N}\sum_{i=1}^{N} n_i.$$

First, generate two natural numbers m and n using Poisson distributions with parameter λ_m and λ_n, respectively. From m, n, and the parameters compute a covariance matrix as in Equation 8 and generate a random sample. Repeat this process until the desired number of random samples is available. Then fit a curve to these random samples, for example, by LOWESS smoothing.

We now compare the analytical and Monte Carlo SROC curve for the dementia data. Table 3 contains the parameters we used to calculate the analytical SROC curve. With the above algorithm we obtained 10,000 random samples and used R's lowess function with default parameters. The theoretical SROC curve and the Monte Carlo SROC curve are shown in Figure B1. The curve obtained from the LOWESS smoother is very similar to the theoretical SROC curve obtained from Equation 9. Also, for other data sets we studied, the Monte Carlo curve is very close to the theoretical curve. Since the effort of random sampling is clearly higher, we believe the theoretical SROC curve should be preferred not only in this example.

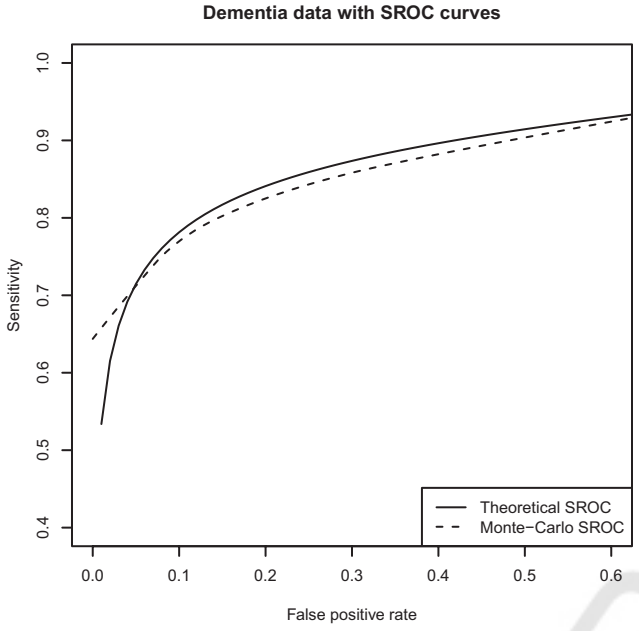

Figure B1. Primary study data from the meta-analysis of the Mini-Mental State Examination to detect dementia with resulting analytical and Monte Carlo summary receiver operating characteristic (SROC) curves.

Received October 4, 2010
Revision received August 30, 2011
Accepted January 19, 2012 ■