Statistical Modelling 2012; **12**(**4**): 347–375

Meta-analysis of diagnostic studies based upon SROC-curves: a mixed model approach using the Lehmann family

Heinz Holling[1], Walailuck Böhning[1] and Dankmar Böhning[2]
[1]Statistics and Quantitative Methods, University of Münster, Münster, Germany
[2]Southampton Statistical Sciences Research Institute, University of Southampton, Highfield Campus, Southampton, SO17 1BJ, UK

Abstract: Meta-analysis of diagnostic studies experiences the common problem that different studies might not be comparable since they have been using a different cut-off value for the continuous or ordered categorical diagnostic test value defining different regions for which the diagnostic test is defined to be positive. Hence specificities and sensitivities arising from different studies might vary just because the underlying cut-off value had been different. To cope with the cut-off value problem, interest is usually directed towards the receiver operating characteristic (ROC) curve which consists of pairs of sensitivities and false positive rate (1–specificity). In the context of meta-analysis, one pair represents one study and the associated diagram is called SROC curve where the *S* stands for 'summary'. The paper will consider—as a novel approach—modelling SROC curves with the Lehmann family that assumes log-sensitivity is proportional to the log-false positive rate across studies. The approach allows for study-specific false positive rates which are treated as (infinitely many) nuisance parameters and eliminated by means of the profile likelihood. The adjusted profile likelihood turns out to have a simple univariate Gaussian structure which is ultimately used for building inference for the parameter of the Lehmann family. The Lehmann model is further extended by allowing the constant of proportionality to vary across studies to cope with unobserved heterogeneity. The simple Gaussian form of the adjusted profile likelihood allows this extension easily as a form of a mixed model in which unobserved heterogeneity is incorporated by means of a normal random effect. Some meta-analytic applications on diagnostic studies including brain natriuretic peptides for heart failure, alcohol use disorder identification test (AUDIT) and the consumption part of AUDIT for detection of unhealthy alcohol use as well as the mini-mental state examination for cognitive disorders are discussed to illustrate the methodology.

Key words: diagnostic accuracy; Lehmann family; profile and adjusted profile likelihood; proportional hazards model; SROC modelling

Received March 2010; **revised** August 2011 & January 2012; **accepted** February 2012

Note: The original plots in colour are available at http://stat.uibk.ac.at/smij/

Address for correspondence: Dankmar Böhning, School of Mathematics & Southampton Statistical Sciences Research Institute, University of Southampton, UK. E-mail: d.a.bohning@soton.ac.uk

10.1177/1471082X1201200403

1 Introduction and Notation

We are interested in the following situation in the field of meta-analysis of diagnostic studies (Hedges and Olkin, 1985; Cooper and Hedges, 1994; Hasselblad and Hedges, 1995; Irwing *et al.*, 1995; Sutton *et al.*, 2000; Egger *et al.*, 2001; Schulze *et al.*, 2003): a variety of diagnostic studies are available providing estimates of the diagnostic measures of specificity $(1 - u) = P(T = 0|D = 0)$ as $\hat{u}_i = x_i/n_i$ (estimate of false positive rate) and of sensitivity $p = P(T = 1|D = 1)$ as $\hat{p}_i = y_i/m_i$ (estimate of sensitivity), where $D = 1$ and $D = 0$ denotes presence or absence of disease, respectively, and $T = 1$ or $T = 0$ denotes positivity or negativity of the diagnostic test, respectively. Also, x_i are the number of false positives out of n_i healthy individuals, y_i are the number of true positives out of m_i diseased individuals, for $i = 1, \ldots, k$, k being the number of studies. For more details on the statistical modelling of the diagnostic situation on the basis of a single study see Pepe (2000, 2003). In the following we will look at several examples from medicine and psychology for this special meta-analytic situation. In principle, however, applications could occur from all areas. Swets (1996) considers mainly psychological applications, but also mentions cases from engineering (quality control), manufacturing (failing parts in planes), meteorology (correctness of weather predictions), information science (correctness of information retrieval) or criminology (correctness of lie detection test). Likewise Krzanowski and Hand (2009), without having specifically the meta-analytic aspect in mind, mention applications from machine learning, atmospheric sciences, geosciences, biosciences, finances, experimental psychology and sociology. We illustrate the special meta-analytic situation mentioned above with a meta-analysis on a diagnostic test on heart failure.

Example 1: Meta-Analysis of Diagnostic Accuracy of Brain Natriuretic Peptides (BNP) for Heart Failure. Doust *et al.* (2004) provide a meta-analysis on the diagnostic accuracy of the BNP as diagnostic test for heart failure. Details are provided in Table 1. According to the authors, diagnosis of heart failure is difficult with both, overdiagnosis and underdiagnosis, occurring. The BNP has been suggested as diagnostic test and the authors provide data from various studies using different reference standards (a reference standard defines the presence or absence of disease). Here we only use the eight studies using the left ventricular ejection fraction of 40% or less as reference standard.

The cut-off value problem. A separate meta-analysis of sensitivity and specificity using the meta-analytic tools for independent binomial samples is problematic when the underlying diagnostic test is continuous or ordered categorical and different cut-off values have been used in different diagnostic studies. A simple variation of the cut-off value from study to study might lead to quite different values of sensitivity and specificity without any actual change in the diagnostic accuracy of the underlying continuous test. This situation is illustrated in Figure 1 for a continuous outcome T which is normally distributed in the two populations.

SROC curve. Because of this comparability problem for sensitivity and specificity, interest is usually focused on the *summary receiver operating characteristic* (SROC)

Table 1 Meta-analysis of diagnostic accuracy of BNP for heart failure using the left ventricular ejection fraction of 40% or less as reference standard

	Diseased		Healthy		
Study *i**	y_i(TP)	$m_i - y_i$(FN)	$n_i - x_i$(TN)	x_i(FP)	$n_i + m_i$
Bettencourt (2000)	29	7	46	19	101
Choy (1994)	34	6	22	13	75
Valli (2001)	49	9	78	17	153
Vasan (2002a)	4	6	1612	85	1707
Vasan (2002b)	20	40	1339	71	1470
Hutcheon (2002)	29	2	102	166	299
Landray (2000)	26	14	75	11	126
Smith (2000)	11	1	93	50	155

Note: *Details on these studies are found in Doust *et al.* (2004).

curve consisting of the pairs $(u(t), p(t))$, where $u(t) = P(T \geq t|D = 0)$ and $p(t) = P(T \geq t|D = 1)$ for a continuous test T with potential value t. Consider k possible unknown cut-off values $t_1, \ldots, t_k$, then the pairs $(u(t_i), p(t_i))$ can be estimated by

$$(\hat{u}_i, \hat{p}_i) = (x_i/n_i, y_i/m_i),$$

for $i = 1, \ldots, k$. The SROC curve copes with the cut-off value problem. Different pairs could have quite different values of specificity and sensitivity, but still reflect identical diagnostic accuracy. The SROC diagram for the meta-analysis on BNP and heart failure is provided in Figure 2. Clearly, there is a wide range of values for specificity and sensitivity. Nevertheless, as Figure 2 shows, it cannot be excluded that the pairs might stem from a common SROC curve (symbolized by the solid line in Figure 2). Since the SROC approach copes with the cut-off value problem, it is commonly preferred to summary measures like the Youden index (Youden, 1950) or the diagnostic odds ratio (Glas *et al.*, 2003), although these measures can be considerably stable under certain conditions as pointed out in Edwards (1966), Hasselblad and Hedges (1995) or Böhning *et al.* (2008). In the following we focus our analysis on the SROC curve.

Background of SROC modelling. SROC modelling has received considerable attention in the field. A first model has been suggested by Littenberg and Moses (Moses *et al.*, 1993; Littenberg and Moses, 1993; Midgette *et al.*, 1993) and has been used in practice frequently. Deeks (2007) discusses its prominent role in modelling meta-analytic diagnostic study accuracy. Jones and Athanasiou (2005) state that the Littenberg-Moses model is one of the most commonly used regression models. Indeed, a simple med-line search reveals that the Littenberg-Moses model has numerous entries in published literature. Littenberg and Moses (1993) suggest to fit $D = \alpha + \beta S$, where $D = \log DOR = \log \frac{p}{1-p} - \log \frac{u}{1-u}$ is the *log-diagnostic odds ratio* and $S = \log \frac{p}{1-p} + \log \frac{u}{1-u}$ is a measure for a potential threshold effect. After α and β have

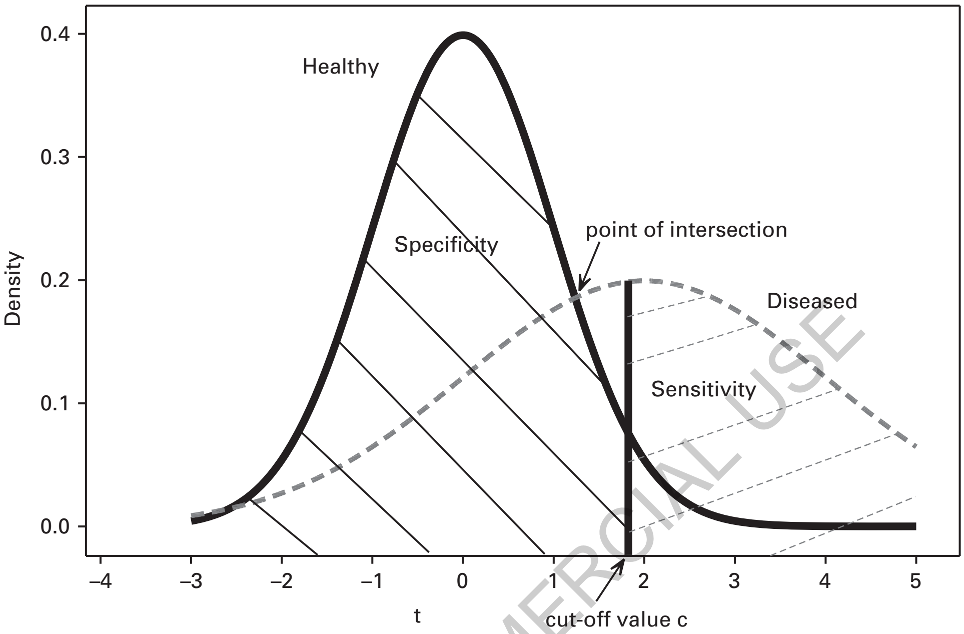

Figure 1 Diagnostic situation illustrated with two normal distributions: one has mean 0 and variance 1 (healthy population), the other has mean 2 and variance 4 (diseased population)

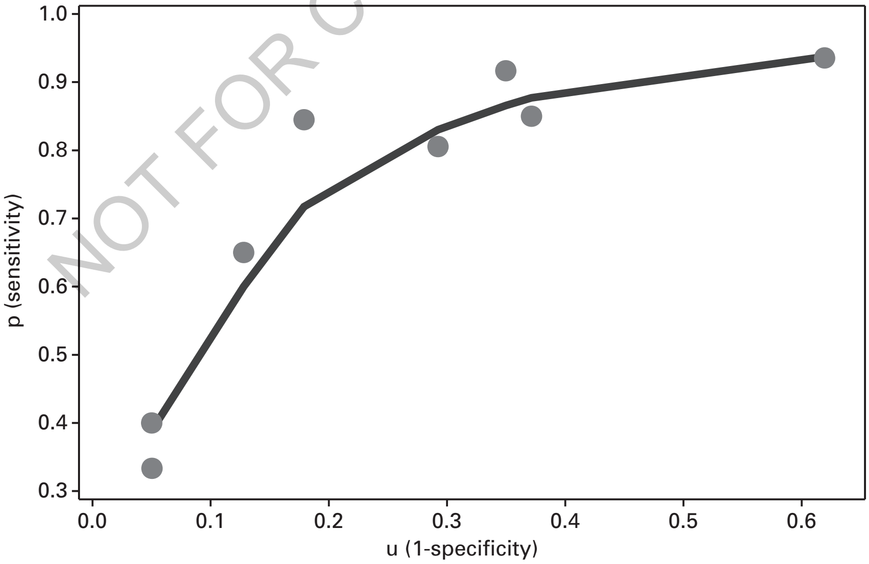

Figure 2 SROC diagram for meta-analysis of BNP and heart failure with LOWESS smoother (solid line)

been fitted from the data, the SROC curve (p vs. u) is reconstructed from the fitted values of α and β. α is interpreted as the *summary log-DOR*, which is adjusted by means of S for potential *cut-off value effect*. A two-level approach has been suggested by Rutter and Gatsonis (2001) which is typically given in the following notational form (Walter and Macaskill, 2004). Let $Y_{ij} \sim Bi(n_{ij}, \pi_{ij})$, where Y_{ij} is the number of test positives in study i for arm j ($j = 1$ is diseased, $j = 2$ is non-diseased), n_{ij} is the size of arm j in study i and π_{i1} is the sensitivity, π_{i2} is the false positive rate. Then the model is

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = (\theta_i + \alpha_i DS_{ij}) \exp(-\beta DS_{ij}),$$

where θ_i is an implicit threshold parameter for study i and α_i is the diagnostic accuracy parameter in study i. DS_{ij} represents a binary variable for the disease status. The parameter β allows for an association between test accuracy and test threshold. When $\beta = 0$, α_i is estimated by D_i and θ_i is estimated by $S_i/2$, where D_i and S_i are as in the Littenberg-Moses model. Furthermore, to account for between-study variation, a random effect is assumed for $\theta_i \sim N(\Theta, \tau_\theta^2)$ and $\alpha_i \sim N(\Lambda, \tau_\alpha^2)$. Yet, in another approach, a bivariate normal random effects meta-analysis has been suggested by van Houwelingen *et al.* (1993, 2002). See also Reitsma *et al.* (2005) and Arends *et al.* (2008). Harbourd *et al.* (2006) show that these models are closely related.

Paper overview. In the following, we will suggest a specific model, called Lehmann model, which we believe is very suitable for the analysis of SROC curves. The model involves study-specific sensitivities and specificities and a diagnostic accuracy parameter which connects the two. Specificities are treated as nuisance parameters and eliminated by means of the profile likelihood. It is shown that this profile likelihood, if correctly adjusted, leads to a proper Gaussian likelihood. The Lehmann model receives flexibility by allowing the diagnostic accuracy parameter to become a random effect. Maximum likelihood inference is developed including a fixed point algorithm for providing maximum likelihood estimates as well as finding variance estimators. Section 3 applies the method to a number of meta-analyses and Section 4 provides comparisons to existing methods. The paper ends with a brief discussion.

2 The Lehmann Model

Le (2006) suggests to model the relationship between sensitivity and false positive rate using the Lehmann family

$$p = u^\theta. \tag{2.1}$$

The Lehmann model has a number of nice properties including that $p \in [0, 1]$ if $u \in [0, 1]$ for $\theta > 0$. Hence it represents a feasible reparameterization of the SROC curve. In addition, the parameter θ is easily interpreted as representing diagnostic accuracy. The smaller the value of θ, the higher the diagnostic accuracy. Some Lehmann models are shown in Figure 3 for different values of θ. Also, two diagnostic

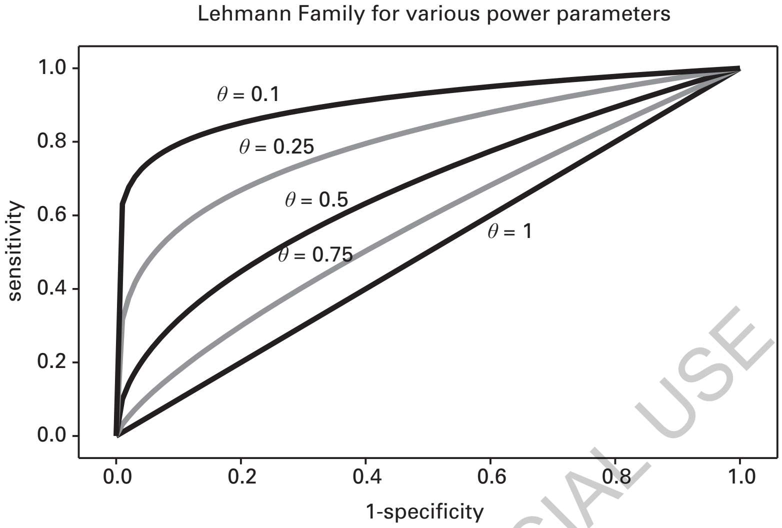

Figure 3 Lehmann model for different values of θ

tests represented by two different θ values can easily be compared. In addition, other measures of interest, such as the *area under the curve* (AUC) can easily be derived as $AUC = \int_0^1 u^\theta du = 1/(1+\theta)$. The model (2.1) has also the property that the ratio of log-true positive rate and log-false positive rate is constant: $\log p(t)/\log u(t) = \theta$. This is very similar to the proportional hazards model (PHM) used in failure time analysis. Here, for failure time t, the hazard $h(t)$ is proportional to a baseline hazard $h_0(t)$, so that the final PHM is $h(t) = h_0(t)\exp(\alpha)$ which might be extended to allow for covariates. This analogy leads Le (2006) to call the model (2.1) also *PHM*. Furthermore, Gönen and Heller (2010) point out the proportional hazards property of the Lehmann family. We might occasionally use this name as well without abstracting from the substantial difference to the failure time scenario. In the following, we are interested in inference for θ.

There are various reasons why model (2.1) is appealing. Recall that we only have *one* pair $(\hat{p}_i, \hat{u}_i)$ of sensitivity and false positive rate available from each study. In the SROC space, this pair is represented by one point. Clearly, infinitely many lines pass through this point, in other words, a straight line model (allowing for intercept and slope to be unconstrained) is *not identifiable* within study i. For this point, see also Rücker and Schumacher (2009, 2010) who also provide an illustrative example of a form of ecological fallacy, a situation where all studies show positive slopes in the logit-transformed ROC space, but if only one point per study is selected, the corresponding SROC has negative slope. Hamza *et al.* (2009) also point out:

> We conclude that in the situation where we have only one pair of sensitivity and specificity per study a calculated SROC can only be interpreted

> as a real overall ROC under an untestable assumption. The assumption is especially sensitive when the differences among the estimated between-studies variances and covariance of sensitivity and specificity are large. This issue seems to have been overlooked in the literature.

However, a straight line that passes through the origin is uniquely characterized by the pair of observations we have from the study. Hence, the model $\log p = \theta \log u$ is identifiable within each study. This is an important property which makes the model preferable to other models, in particular those, which are not identifiable. It is also clear that in this case, the ecological fallacy described in Rücker and Schumacher (2010) cannot occur to the expense that an untestable assumption (line through the origin) is made. However, it is also clear that it is not the only identifiable model in this situation. These issues are further discussed in the final section of this paper.

2.1 Profile Likelihood

In the following, we consider the profile likelihood method. For one, it is a widely used method to eliminate a nuisance parameter. For two, it has an invariance property that we illuminate further below after we have developed profile likelihood for the case here. Consider the product-binomial likelihood $\binom{m}{y} p^y(1-p)^{m-y} \times \binom{n}{x} u^x(1-u)^{n-x}$ as the joint distribution of Y_i and X_i for the ith study (index is suppressed for notational convenience), which we replace by the normal approximation for $\log Y_i$ and $\log X_i$

$$\frac{1}{\sqrt{2\pi s^2}} \exp\left\{-\frac{1}{2}\frac{(\log y - \log(mp))^2}{s^2}\right\} \times \frac{1}{\sqrt{2\pi t^2}} \exp\left\{-\frac{1}{2}\frac{(\log x - \log(nu))^2}{t^2}\right\},$$

with the Taylor-series variance estimates $s^2 = \frac{1}{y} - \frac{1}{m}$ and $t^2 = \frac{1}{x} - \frac{1}{n}$. We have used that the associated, estimated variances for the log-proportions $\log(y_i/m_i)$ and $\log(x_i/n_i)$ are provided as

$$\widehat{Var}(\log \hat{p}_i) = \widehat{Var}(\log(y_i/m_i)) = s_i^2 = \frac{1}{y_i} - \frac{1}{m_i}, \tag{2.2}$$

$$\widehat{Var}(\log \hat{u}_i) = \widehat{Var}(\log(x_i/n_i)) = t_i^2 = \frac{1}{x_i} - \frac{1}{n_i}, \tag{2.3}$$

assuming that $y_i > 0$ and $x_i > 0$ for $i = 1, \dots, k$. Furthermore, let $z_i = \log y_i - \log m_i$ and $w_i = \log x_i - \log n_i$, so that z_i is the *log-true positive rate* and w_i is the *log-false positive rate*.

The normal approximation is justified if the sizes per study are not small (which is typically the case in diagnostic studies) and matches well with the Lehmann family. Consider now the relevant part of the log-likelihood for study i

$$-\frac{1}{2s^2}(\underbrace{\log y - \log m}_{z} - \log p)^2 - \frac{1}{2t^2}(\underbrace{\log x - \log n}_{w} - \log u)^2,$$

which can be further written as

$$\ell(\theta, u') = -\frac{1}{2s^2}(z - \overbrace{\theta u'}^{\log p})^2 - \frac{1}{2t^2}(w - u')^2,$$

with $u' = \log u$. Maximizing $\ell(\theta, u')$ in u' for *fixed* θ leads to

$$\hat{u}'_\theta = \frac{\theta t^2 z + s^2 w}{t^2\theta^2 + s^2}$$

and plugging in $\hat{u}'_\theta$ provides the *profile log-likelihood*

$$\ell(\theta) = \ell(\theta, \hat{u}'_\theta) = -\frac{1}{2s^2}(z - \theta\hat{u}'_\theta)^2 - \frac{1}{2t^2}(w - \hat{u}'_\theta)^2 = -\frac{1}{2}\frac{(z - w\theta)^2}{t^2\theta^2 + s^2}$$

resulting in a profile log-likelihood of *remarkable simplicity*.

In addition, the profile log-likelihood has the following *invariance property*. Note that there are two forms of the ROC model:

$$\log p = \theta \log u \text{ or } \log u = \frac{1}{\theta}\log p.$$

In the first model, we can think of regressing the log-sensitivity on the log-false positive rate, whereas in the second model, the log-false positive rate is regressed on the log-sensitivity. It is well known in classical regression inference that both problems can have different solutions. Now, the profile maximum likelihood is *invariant* to the choice of the nuisance parameter, e.g., if u or p is chosen to be the nuisance parameter: $\ell(\theta, \hat{u}'_\theta) = \ell(\theta, \hat{p}'_\theta)$. Since it is arbitrary in the ROC diagram which axis is labelled as sensitivity and which one as false positive rate, in other words, which model of the two is chosen for the analysis, the profile likelihood is suitable for the inference since the choice of the nuisance parameter (sensitivity or false positive rate) will ultimately not affect the inference on the parameter of interest.

We have noticed already that $\ell(\theta)$ is almost a Gaussian log-likelihood:

$$\ell(\theta) = \ell(\theta, \hat{u}'_\theta) = -\frac{1}{2}\frac{(z - w\theta)^2}{\underbrace{t^2\theta^2 + s^2}_{\sigma^2(\theta)}}.$$

It differs from $L(\theta) = -\frac{1}{2}\log\sigma^2(\theta) - \frac{1}{2}\frac{(z-w\theta)^2}{\sigma^2(\theta)}$ only by $\frac{1}{2}\log\sigma^2(\theta)$. The main problem of the conventional profile likelihood $\ell(\theta)$ is that it is not a proper likelihood. In particular, first- and second-order properties are not necessarily valid. In addition, it is thought that the curvature of the profile likelihood is *not* correct to give a valid variance estimate. Since the profile likelihood takes the estimated nuisance parameter as a true parameter value, it is thought of *underestimating the variance* of the parameter of interest (Patefield, 1977; Aitkin, 1998; Murphy and Van der Vaart, 2000). In addition, the conventional profile log-likelihood $\ell(\theta)$ breaks down if further variance components are incorporated as this would be necessary if unobserved heterogeneity occurs (see Section 2.2). However, it was shown by Barndorff-Nielsen (1983) that an approximate marginal or conditional likelihood could be found by adjusting the ordinary profile likelihood. Furthermore, it was pointed out by Cox and Reid (1987) that the adjustment term could be simplified to $\hat{I}(\hat{u}_\theta)^{-1/2}$ if parameters are orthogonal (or close to orthogonality). Lee *et al.* (2006, pp. 32–34) provide a discussion on the modified ordinary profile likelihood and call this modified profile likelihood the *adjusted profile likelihood* which turns out in our case to be

$$\hat{I}(\hat{u}_\theta) = -\frac{\partial^2}{\partial u'^2}\ell(\theta, u') = \frac{\partial^2}{\partial u'^2}\left(\frac{1}{2s^2}(z-\theta\hat{u}')^2 + \frac{1}{2t^2}(w-\hat{u}')^2\right), \qquad (2.4)$$

where, for fixed θ, $\hat{I}(\hat{u}_\theta)$ is the *observed Fisher information* $\hat{I}(u)$ evaluated at $\hat{u}_\theta$. As can be seen directly from (2.4)

$$\hat{I}(\hat{u}_\theta) = \frac{\partial^2}{\partial u'^2}\left(\frac{1}{2s^2}(z-\theta\hat{u}')^2 + \frac{1}{2t^2}(w-\hat{u}')^2\right) = \frac{t^2\theta^2+s^2}{s^2t^2},$$

so that

$$-\frac{1}{2}\log[\hat{I}(\theta)] + \ell(\theta) + const. = L(\theta),$$

where the constant is independent of θ, providing an *excellent* justification of the adjusted profile likelihood for our case.

For a sample of *k* studies, we have the *full-sample adjusted profile* log-likelihood as

$$-\sum_i \frac{1}{2}\log\sigma_i^2(\theta) - \sum_i \frac{1}{2}\frac{(z_i - w_i\theta)^2}{\sigma_i^2(\theta)},$$

where $\sigma_i^2(\theta) = t_i^2\theta^2 + s_i^2$. Note that $[\sigma_i^2(\theta)]' = 2t_i^2\theta$. The likelihood above implies that $Z_i \sim N(\theta w_i, \sigma_i^2(\theta))$. However, it is more appealing to formulate the mean structure model without w_i, so that we equivalently formulate the model, conditionally on w_i, as

$$\Theta_i = Z_i/w_i = \theta + \epsilon_i, \qquad (2.5)$$

where $\epsilon_i \sim N(\theta, \sigma_i^2(\theta)/w_i^2)$. The associated log-likelihood is

$$L(\theta) = -\sum_i \frac{1}{2}\log\sigma_i^2(\theta)/w_i^2 - \sum_i \frac{1}{2}\frac{(z_i/w_i - \theta)^2}{\sigma_i^2(\theta)/w_i^2}.$$

We find the following score for the adjusted profile log-likelihood. Using $v_i = \frac{1}{\sigma_i^2(\theta)}$, the *adjusted profile log-likelihood* is

$$\begin{aligned}\frac{\partial L}{\partial \theta} &= \sum_i \left\{ \frac{(z_i - w_i\theta)w_i}{\sigma_i^2(\theta)} + \frac{(z_i - w_i\theta)^2 t_i^2 \theta}{(\sigma_i^2(\theta))^2} - \frac{t_i^2\theta}{\sigma_i^2(\theta)} \right\} \\ &= \sum_i \left\{ (z_i - w_i\theta)w_i v_i + (z_i - w_i\theta)^2 v_i^2 t_i^2 \theta - t_i^2 v_i \theta \right\}. \qquad (2.6)\end{aligned}$$

Note that the expected value of the score $U = \frac{\partial L}{\partial \theta}$ of the adjusted profile log-likelihood meets the conventional first-order property $E(U) = 0$:

$$\begin{aligned}E\left(\frac{\partial L}{\partial \theta}\right) &= \sum_i \left\{ E(z_i - w_i\theta)w_i v_i + [E(z_i - w_i\theta)^2] v_i^2 t_i^2 \theta - t_i^2 v_i \theta \right\} \\ &= \sum_i [0 + \sigma_i(\theta)^2 v_i^2 t_i^2 \theta - t_i^2 v_i \theta] = 0,\end{aligned}$$

whereas this is not the case for the score of the ordinary profile likelihood.

To solve the score equation for the *adjusted profile likelihood*, we note that (2.6) can be written in the form

$$\theta = \frac{\sum_i z_i w_i v_i}{\sum_i \left(w_i^2 v_i - (z_i - w_i\theta)^2 v_i^2 t_i^2 + t_i^2 v_i \right)}. \qquad (2.7)$$

Note that the right-hand side of (2.7) depends on θ, so that a solution needs to be found using the following iterative scheme: given θ_j, compute $\sigma_i^2(\theta_j)$ and $v_i = 1/\sigma_i^2(\theta_j)$. Then use (2.7) to compute a new θ_{j+1} and repeat this process until convergence. This will provide the maximum likelihood estimate for the adjusted profile likelihood at convergence—the *adjusted profile maximum likelihood estimate* (APMLE). There is no theoretical convergence result of this algorithm. However, the algorithm was used in all simulation studies without any failure.

We are easily able to construct a goodness-of-fit statistic. Since $E(Z_i) = \theta \log u_i$ and $E(W_i) = \log u_i$, it follows that $E(Z_i - \theta W_i) = 0$. We have also that $Var(Z_i) = s_i^2$ and $Var(\theta W_i) = \theta^2 t_i^2$, so that $Var(Z_i - \theta W_i) = s_i^2 + \theta^2 t_i^2$. Hence we have that

$$\frac{Z_i - \theta W_i}{\sqrt{s_i^2 + \theta^2 t_i^2}} = \frac{Z_i / W_i - \theta}{\sqrt{s_i^2 + \theta^2 t_i^2} / W_i}$$

is approximately a standard normal variate. Furthermore,

$$\chi_{k-1}^2 = \sum_{i=1}^{k} \frac{(\hat{\theta}_i - \hat{\theta})^2}{(s_i^2 + \hat{\theta}^2 t_i^2) / W_i^2}$$

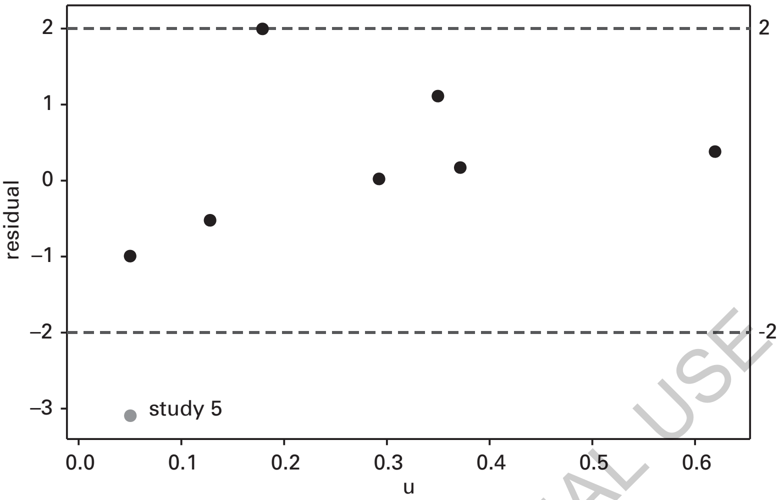

Figure 4 Residual plot for the eight studies in the BNP heart failure meta-analysis

will have an approximate χ^2-distribution with $k - 1$ df if the Lehmann model is correct.

Example 1 (continued): Meta-Analysis of Diagnostic Accuracy of BNP for Heart Failure. We come back to the previously introduced meta-analysis of Doust *et al.* (2004) on the diagnostic accuracy of the BNP as diagnostic test for heart failure. The APMLE is given as $\hat{\theta} = 0.1774$ with 95% CI of 0.1494–0.2054. This corresponds to an AUC of 0.85, a value of moderate diagnostic accuracy although the interpretation will depend on the diagnostic accuracy of alternative diagnostic tests. Furthermore, we find a χ^2-statistic which shows borderline significance with $\chi^2 = 16.23$ (7 df) and a p-value of 0.0231. Note that this test statistics investigates the hypothesis of homogeneity that all eight θ-parameters share the same value. We contemplate this example as a case of homogeneity despite the borderline significance for the following reason. If we consider Figure 4 which represents an index–plot of the residuals $\frac{Z_i - \theta W_i}{\sqrt{s_i^2 + \theta^2 t_i^2}}$, we study that 5 is causing the major contribution to the χ^2. Indeed, if this study is removed, the observed significance disappears.

Example 2: Meta-Analysis of Diagnostic Accuracy of the Alcohol Use Disorder Identification Test (AUDIT) for Alcohol Disorder. One of the most frequently recommended instruments (including a recommendation from the WHO) for screening all forms of unhealthy alcohol use (risky drinking, alcohol abuse, alcohol dependence) is the AUDIT. The full AUDIT consists of 10 items and has been extensively investigated in several settings and countries (Reinert and Allen, 2002). Here we look at a meta-analysis provided by Kriston *et al.* (2008). The data are provided in Table 2

Table 2 Meta-analysis of diagnostic accuracy of the AUDIT for alcohol disorder

	Alcohol disorder		No disorder		
Study i	y_i(TP)	$m_i - y_i$(FN)	$n_i - x_i$(TN)	x_i(FP)	$n_i + m_i$
1	48	7	738	101	894
2	138	39	1506	309	1992
3	24	5	173	31	233
4	37	2	227	127	393
5	137	12	936	234	1319
6	73	13	127	30	243
7	53	14	508	27	602
8	571	180	5707	496	6954
9	54	10	172	19	255
10	148	44	2687	672	3551
11	143	18	334	130	625
12	47	13	464	76	600
13	34	1	65	12	112
14	154	49	261	92	555

and the associated SROC curve in Figure 5. The analysis of the meta-analysis on AUDIT and alcohol disorders provides an APMLE of $\hat{\theta}$ = 0.0980 with 95% CI of 0.0922–0.1038. This corresponds to an AUC of 0.91, a value of good diagnostic accuracy. However, there is strong evidence of heterogeneity as indicated by a highly

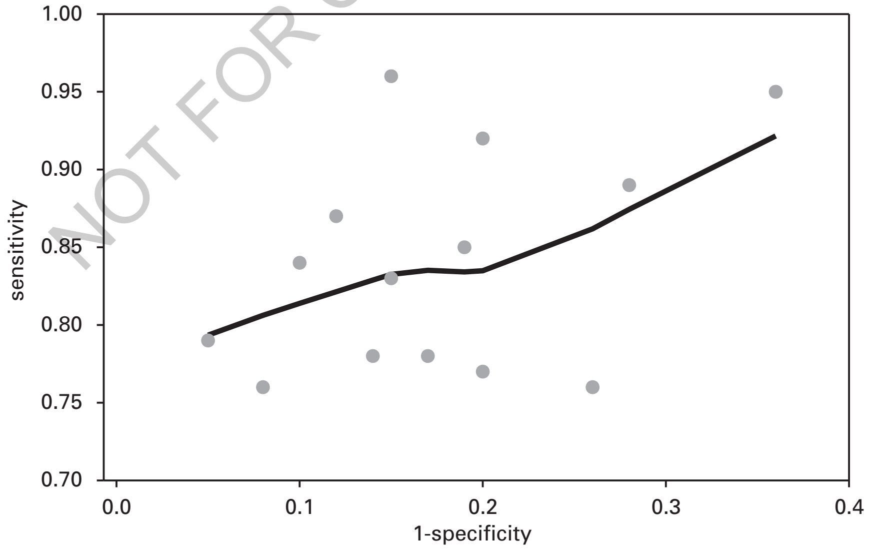

Figure 5 SROC diagram for meta-analysis of AUDIT and alcohol disorder with LOWESS smoother (solid line)

significant χ^2 statistic of 54.60 (13 df). This becomes also evident from Figure 5. Hence, it is concluded that additional heterogeneity is present in this meta-analysis and must be incorporated into the inference to achieve, for example, valid confidence intervals.

Unfortunately, heterogeneity is prevalent in many of these forms of meta-analysis and needs to be incorporated appropriately. This is the topic of the next section.

2.2 Heterogeneity

Previously, we have assumed that the Lehmann model $p = u^\theta$ holds across studies allowing study-specific false positive rates, but an identical proportionality parameter θ. This is now generalized in the sense that heterogeneity is allowed for θ which may vary from study to study. If heterogeneity with respect to the diagnostic accuracy parameter θ occurs, it seems appropriate to include a further random effect variance component parameter τ^2, so that $\sigma_i^2(\theta)/w_i^2$ is replaced by $\sigma_i^2(\theta)/w_i^2 + \tau^2$. This is accomplished by extending the fixed effect model by a further *random effect* δ_i, independent of ϵ_i, with $E(\delta_i) = 0$ and $Var(\delta_i) = \tau^2$

$$\Theta_i = Z_i/w_i = \theta + \delta_i + \epsilon_i, \tag{2.8}$$

so that $\Theta_i|w_i \sim N(\theta, \sigma_i^2(\theta)/w_i^2 + \tau^2)$. The *full-sample adjusted profile* log-likelihood with *random effect* is then

$$L(\theta, \tau^2) = -\sum_{i=1}^{k} \frac{1}{2} \log[\sigma_i^2(\theta)/w_i^2 + \tau^2] - \sum_{i=1}^{k} \frac{1}{2} \frac{(\hat{\theta}_i - \theta)^2}{\sigma_i^2(\theta)/w_i^2 + \tau^2}, \tag{2.9}$$

where $\sigma_i^2(\theta) = t_i^2\theta^2 + s_i^2$. We will base all inference in the following on this adjusted profile log-likelihood (2.9) which clearly is a true log-likelihood.

We find the following scores for the full-sample adjusted profile log-likelihood (2.9):

$$\begin{aligned} \frac{\partial L}{\partial \theta} &= \sum_i \left\{ \frac{(\hat{\theta}_i - \theta)}{\sigma_i^2(\theta)/w_i^2 + \tau^2} + \frac{(\hat{\theta}_i - \theta)^2 (t_i^2/w_i^2)\theta}{(\sigma_i^2(\theta)/w_i^2 + \tau^2)^2} - \frac{(t_i^2/w_i^2)\theta}{\sigma_i^2(\theta)/w_i^2 + \tau^2} \right\} \\ &= \sum_i \left\{ (\hat{\theta}_i - \theta)v_i + (\hat{\theta}_i - \theta)^2 v_i^2 (t_i^2/w_i^2)\theta - (t_i^2/w_i^2)v_i\theta \right\}, \end{aligned} \tag{2.10}$$

where $v_i = \frac{1}{\sigma_i^2(\theta)/w_i^2+\tau^2}$, and

$$\begin{aligned}\frac{\partial L}{\partial \tau^2} &= \sum_i \left\{ \frac{1}{2}\frac{(\hat{\theta}_i-\theta)^2}{(\sigma_i^2(\theta)/w_i^2+\tau^2)^2} - \frac{1}{2}\frac{1}{\sigma_i^2(\theta)/w_i^2+\tau^2} \right\} \\ &= \sum_i \left\{ \frac{1}{2}(\hat{\theta}_i-\theta)^2 v_i^2 - \frac{1}{2}v_i \right\}, \end{aligned} \tag{2.11}$$

for the partial derivative with respect to τ^2. Note again that the score $U = (\frac{\partial L}{\partial \theta}, \frac{\partial L}{\partial \tau^2})$ has the the first-order property $E(U) = \mathbf{0}$. We write the score equations stemming from (2.10) as

$$\sum_i \left\{ (\hat{\theta}_i-\theta)v_i + (\hat{\theta}_i-\theta)^2 v_i^2 (t_i^2/w_i^2)\theta - (t_i^2/w_i^2)v_i\theta \right\} = 0,$$

or equivalently as

$$\theta = \frac{\sum_i \hat{\theta}_i v_i}{\sum_i v_i - (\hat{\theta}_i-\theta)^2 v_i^2 (t_i^2/w_i^2) + (t_i^2/w_i^2)v_i} \tag{2.12}$$

and (2.11) as

$$\sum_i \left\{ (\hat{\theta}_i-\theta)^2 v_i^2 - (\sigma_i^2(\theta)/w_i^2+\tau^2)v_i^2 \right\} = 0$$

or, equivalently

$$\tau^2 = \frac{\sum_i [(\hat{\theta}_i-\theta)^2 - \sigma_i^2(\theta)/w_i^2]v_i^2}{\sum_i v_i^2}. \tag{2.13}$$

The fixed point equation (2.13) is also of the form of an *iterative weighted least squares* solution and needs to be solved simultaneously with (2.12). Hence, we have the following algorithm for the case of heterogeneity.

Algorithm for APMLE

1. (Initialization). Choose initial values for θ_1 and τ_1^2 such as $\theta_1 = 0.5$ and $\tau_1^2 = 0$. Set $j = 1$.
2. Compute $v_i = 1/[\sigma_i^2(\theta_j)/w_j^2 + \tau_j^2]$ for $i = 1, \ldots, k$.
3. Compute

$$\theta_{j+1} = \frac{\sum_i \hat{\theta}_i v_i}{\sum_i v_i - (\hat{\theta}_i-\theta_j)^2 v_i^2 (t_i^2/w_i^2) + (t_i^2/w_i^2)v_i}.$$

4. Compute

$$\tau^2_{j+1} = \frac{\sum_i [(\hat{\theta}_i - \theta_j)^2 - \sigma_i^2(\theta_j)/w_i^2] v_i^2}{\sum_i v_i^2}.$$

5. Set $j = j + 1$ and go to step 2.

Some appropriate stopping rule needs to be enforced to terminate iteration. We use as stopping rule that $|\theta_{j+1} - \theta_j| < \epsilon$ and $|\tau^2_{j+1} - \tau^2_j| < \epsilon$ must be met.

2.3 Standard errors of estimate and adjusted goodness-of-fit

The partial derivatives w.r.t. θ and τ^2 can be written as

$$\frac{\partial L}{\partial \theta} = \sum_i \left\{ \frac{(\hat{\theta}_i - \theta)}{\sigma_i^2(\theta)/w_i^2 + \tau^2} + \frac{(\hat{\theta}_i - \theta)^2 (t_i^2/w_i^2)\theta}{(\sigma_i^2(\theta)/w_i^2 + \tau^2)^2} - \frac{(t_i^2/w_i^2)\theta}{\sigma_i^2(\theta)/w_i^2 + \tau^2} \right\} = \sum_i u_i^{(1)}$$

and

$$\frac{\partial L}{\partial \tau^2} = \sum_i \left\{ \frac{1}{2} \frac{(\hat{\theta}_i - \theta)^2}{(\sigma_i^2(\theta)/w_i^2 + \tau^2)^2} - \frac{1}{2} \frac{1}{\sigma_i^2(\theta)/w_i^2 + \tau^2} \right\} = \sum_i u_i^{(2)}.$$

Hence, we can find an estimate of the variance-covariance matrix of $(\hat{\theta}, \hat{\tau}^2)^T$ as the inverse of

$$\hat{I}(\theta, \tau^2) = \begin{pmatrix} \sum_i \left(u_i^{(1)}\right)^2 & \sum_i u_i^{(1)} u_i^{(2)} \\ \sum_i u_i^{(1)} u_i^{(2)} & \sum_i \left(u_i^{(2)}\right)^2 \end{pmatrix},$$

so that estimates of $Var(\hat{\theta})$ can be found as

$$\frac{\sum_i \left(u_i^{(2)}\right)^2}{\sum_i \left(u_i^{(1)}\right)^2 \sum_i \left(u_i^{(2)}\right)^2 - \left(\sum_i u_i^{(1)} u_i^{(2)}\right)^2} \tag{2.14}$$

and of $Var(\hat{\tau}^2)$ as

$$\frac{\sum_i \left(u_i^{(1)}\right)^2}{\sum_i \left(u_i^{(1)}\right)^2 \sum_i \left(u_i^{(2)}\right)^2 - \left(\sum_i u_i^{(1)} u_i^{(2)}\right)^2}. \tag{2.15}$$

This first-order method of estimating the variance-covariance matrix has been suggested including McLachlan and Krishnan (1997, p. 122) since it often provides a more reliable way of estimating the variance-covariance matrix than second-order methods. In the simulation study (provided as online supplementary material), it is shown that this approximation is reasonable and slightly conservative.

We note in passing that also an estimate of τ^2 can be constructed following the DerSimonian-Laird approach. Consider again the realizations $\hat{\theta}_i$ of Θ_i for $i = 1, \ldots, k$. Let $\hat{\omega}_i = \hat{\theta}_i^2 \left(\frac{s_i^2}{z_i^2} + \frac{t_i^2}{w_i^2} \right)$ denote the associated variances. Then an estimate of τ^2 can be provided by the DerSimonian–Laird estimator (DerSimonian and Laird, 1986; Malzahn *et al.*, 2000; Böhning *et al.*, 2002)

$$\hat{\tau}^2 = \frac{\chi^2 - (k-1)}{\sum_i \frac{1}{\hat{\omega}_i} - \frac{\sum_i 1/\hat{\omega}_i^2}{\sum_i 1/\hat{\omega}_i}},$$

where $\chi^2 = \sum_{i=1}^k (\hat{\theta}_i - \bar{\theta})^2/\hat{\omega}_i$ and $\bar{\theta} = \frac{\sum_i (\hat{\theta}_i/\hat{\omega}_i)}{\sum_i 1/\hat{\omega}_i}$. With $\hat{\tau}^2$ available, we can define

$$\bar{\theta}_{DL} = \frac{\sum_i (\hat{\theta}_i / [\hat{\omega}_i + \hat{\tau}^2])}{\sum_i 1/(\hat{\omega}_i + \hat{\tau}^2)}.$$

For the inverse-variance weighted estimate, we will use

$$Var(\bar{\theta}_{DL}) = \frac{1}{\sum_i 1/(\hat{\omega}_i + \hat{\tau}^2)}. \tag{2.16}$$

Having fitted the heterogeneity model (2.8) with parameter $(\hat{\theta}, \hat{\tau}^2)^T$, the adjusted χ^2 goodness-of-fit is

$$\chi^2_{het} = \sum_{i=1}^{k} \frac{(\hat{\theta}_i - \hat{\theta})^2}{(\sigma_i^2(\hat{\theta})/w_i^2 + \hat{\tau}^2)},$$

which has now $(k-2)$ degrees of freedom since we loose 2 df for estimating two parameters. For the inverse-variance weighted method, a similar χ^2 goodness-of-fit is obtained.

We have investigated in a simulation study (supplied as supplementary material) the behaviour of these estimators. The results indicate that (2.14) and (2.15) provide excellent approximations to the true variances. The simulation study also shows that the DerSimonian–Laird approach is not as efficient as the APMLE. For more details, see under archives at http://stat.uibk.ac.at/smij/

Hence, we concentrate on the latter in the following applications.

Table 3 Meta-analysis of diagnostic accuracy of the AUDIT-C for alcohol disorder

	Alcohol disorder		No disorder		
study	y(TP)	$m-y$(FN)	$n-x$(TN)	x(FP)	$n+m$
1	47	9	738	101	894
2	126	51	1543	272	1992
3	19	10	192	12	233
4	36	3	276	78	393
5	130	19	959	211	1319
6	84	2	89	68	243
7	67	0	423	112	602
8	751	0	2977	3226	6954
9	59	5	136	55	255
10	142	50	2788	571	3551
11	137	24	358	107	625
12	57	3	437	103	600
13	34	1	56	21	112
14	152	51	264	88	555

3 Applications

In the following, we will discuss some applications from medicine and psychology in more detail.

3.1 AUDIT and AUDIT-C for alcohol disorders

Kriston *et al.* (2008) consider in their meta-analysis, besides the AUDIT itself, also the consumption part of the AUDIT, called the AUDIT-C. The background of this is as follows. Since the diagnostic instrument is designed to be applied to a large number of people, it is beneficial to have a short instrument available. The AUDIT-C uses only the three items of the original AUDIT related to alcohol intake and there is evidence that this three-item version is also appropriate to screen for unhealthy alcohol use (Reinert and Allen, 2002). In Table 3, we reproduce the data in Kriston *et al.* (2008) on 14 studies using the AUDIT-C. Here the question of interest is if the AUDIT-C represents a similar diagnostic accuracy as the original AUDIT. The associated SROC diagrams are provided in Figure 6. The analysis in Table 4 on the basis of the adjusted profile likelihood (2.9) incorporating heterogeneity variance shows a difference in diagnostic accuracy between AUDIT and AUDIT-C (in fact, AUDIT-C having the better accuracy), but this difference is non-significant. However, AUDIT-C shows the larger heterogeneity in terms of the estimated heterogeneity variance τ^2 which leads to a larger confidence interval for the AUDIT-C meta-analysis. Hence, the less complex AUDIT-C questionnaire is on average as accurate as the AUDIT

Table 4 Meta-analysis for AUDIT/AUDIT-C data

Estimator	$\hat{\theta}$	$\widehat{SE}(\hat{\theta})$	95% CI	χ^2 (*p*-val)
			AUDIT-C	
APMLE	0.0894	0.0198	0.0417–0.1191	13.84 (0.3111)
			AUDIT	
APMLE	0.0965	0.0132	0.0707–0.1223	13.89 (0.3076)

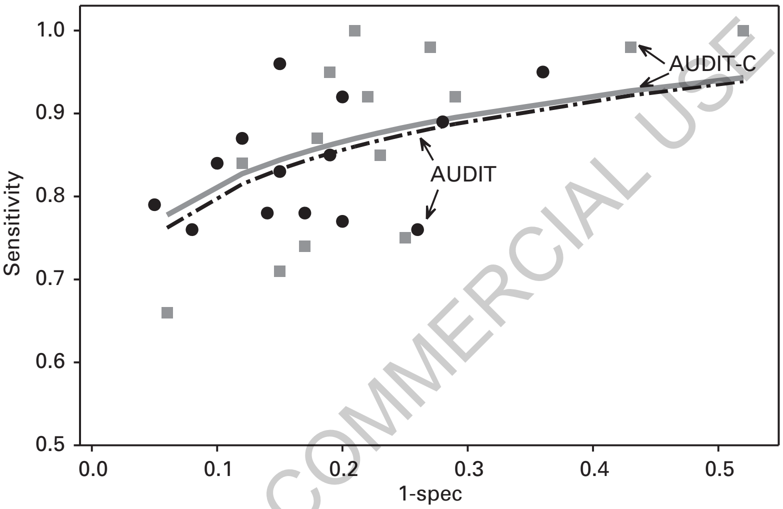

Figure 6 SROC diagram for 14 studies in the AUDIT/AUDIT-C—alcohol disorder meta-analysis

questionnaire, but the variation across studies is larger for the AUDIT-C. Hence, a price of less precision seems to be paid if using the less complex AUDIT-C.

3.2 Mini-mental state examination for dementia and cognitive impairment

In the following, we consider a meta-analysis by Mitchell (2009) on the mini-mental state examination (MMSE) as a diagnostic test for the detection of dementia and mild cognitive impairment (MCI). The data are reproduced in Table 5 in a form that they allow a reanalysis with the methods developed here. Note that one dementia study had to be excluded from the analysis since it was impossible to calculate the frequencies of true positives, false positives, true negatives and false negatives.

Figure 7 shows the SROC diagram for the dementia as well as for the studies with MCI. In this case, we are not comparing two tests but the diagnostic accuracy of the MMSE for the two conditions, namely dementia and MCI. There is a clear difference in diagnostic accuracy between the two conditions with a clear indication of higher

Table 5 Meta-analysis of diagnostic accuracy of the MMSE of dementia and MCI

	Condition		No condition	
Condition	y(TP)	$m-y$(FN)	x(FP)	$n-x$(TN)
Dementia	65	3	240	870
Dementia	117	12	10	110
Dementia	48	19	63	989
Dementia	134	8	28	152
Dementia	24	5	44	292
Dementia	67	15	48	153
Dementia	64	17	0	71
Dementia	281	64	20	286
Dementia	13	1	44	286
Dementia	262	20	29	177
Dementia	143	18	29	123
Dementia	183	33	33	51
Dementia	22	0	152	140
Dementia	112	0	590	2091
Dementia	152	81	126	1009
Dementia	29	26	26	236
Dementia	31	6	3	247
Dementia	10	3	12	333
Dementia	707	88	1438	10 447
Dementia	181	108	17	184
Dementia	59	29	23	74
Dementia	74	23	16	143
Dementia	27	12	26	209
Dementia	40	6	75	528
Dementia	317	52	173	578
Dementia	387	116	16	54
Dementia	118	65	1	44
Dementia	44	7	34	396
Dementia	123	46	98	309
Dementia	25	43	3	171
Dementia	73	32	2	225
Dementia	37	45	0	440
Dementia	78	34	45	376
MCI	72	12	53	214
MCI	106	23	410	379
MCI	37	36	22	118
MCI	67	30	22	75
MCI	17	77	0	90

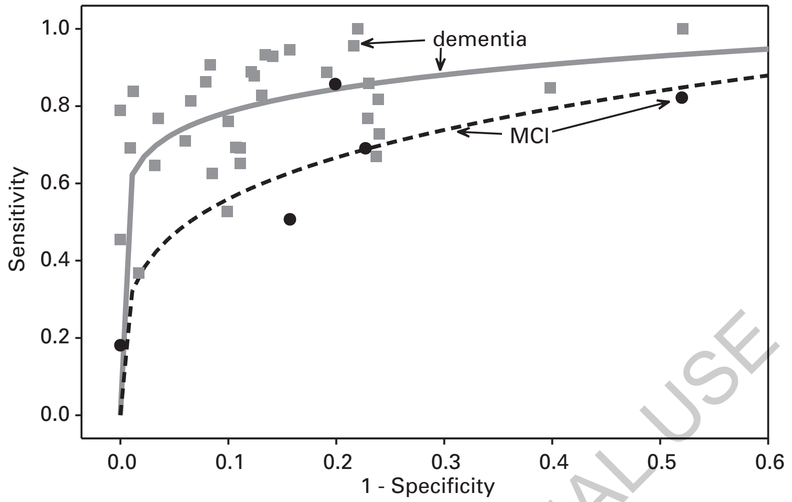

Figure 7 SROC diagram for 38 studies in the MMSE dementia/MCI meta-analysis

accuracy for the dementia condition. Table 6 provides the details on the modelling. For both conditions, there is empirical evidence for heterogeneity as expressed in the χ^2-value in column 5 of Table 6. Hence, the analysis is based on the heterogeneity model (2.8) and parameter estimates, standard errors and confidence intervals are then computed under the adjusted profile maximum likelihood (2.9). The diagnostic accuracy is higher for dementia in comparison to MCI. In addition, the standard error of $\hat{\theta}$ is a lot larger for MCI indicating also less precision of the MMSE for this condition in comparison to dementia. It appears that the MMSE works better in terms of its diagnostic accuracy for dementia than for MCI.

3.3 A database of diagnostic meta-analytic applications

Since there is limited space, we were only able to present a small number of meta-analyses from a wide range of possible choices. As supporting empirical background research of this paper, a database of meta-analytic datasets was formed. For more details, see under archives at http://stat.uibk.ac.at/smij/

Table 6 Meta-analysis for MMSE dementia/MCI data

Estimator	$\hat{\theta}$	$\widehat{SE}(\hat{\theta})$	95% CI	χ^2 (p-val)	χ^2_{het} (p-val)
			dementia		
APMLE	0.1052	0.0132	0.0793–0.1311	459.79 (< 0.0000)	34.92 (0.2870)
			MCI		
APMLE	0.2521	0.0794	0.0965–0.4077	23.92 (0.0001)	4.42 (0.2199)

This database is updated continuously and contains currently about 50 meta-analytic datasets and is available as supplementary information on the journal's website. These datasets were collected on the basis of *published* literature (in contrast to compiling an independent meta-analysis to a given topic). All areas of relevance in medicine and psychology (including neighbouring areas) were considered. The *only* criterion for becoming part of the database was that the published evidence allowed to identify the following minimum information from each study involved in the meta-analysis of interest: true positives, false negatives for the diseased arm (group with the condition) and true negatives, false positives for the non-diseased arm (group without the condition). This criterion was chosen since we wanted to provide datasets that would allow a secondary analysis with the methods provided here and elsewhere in the literature. These four frequencies turned out to be essential in doing so, in contrast to situations where only sensitivities and specificities were given which would not allow any reanalysis of the meta-analytic datasets.

4 Model diagnostics and comparison to other approaches

The question arises how appropriate the suggested Lehmann model is and it compares to other existing approaches. We emphasize that in our situation we have assumed that there is only *one* pair of sensitivity and false positive rate $(\hat{p}_i, \hat{u}_i)$ per study i observed. Situations where several pairs per study are observed (such as in Aertgeerts *et al.*, 2004) are rare and not typical. Hence, we are not able to identify any straight line model *within a study* with *more than one* parameter, since this would require at least two pairs of sensitivity and specificity per study. For this point, see also Rücker and Schumacher (2009, 2010). However, any one-parameter straight line model within each study is estimable including the proposed Lehmann model, although within-model diagnostics is limited since we are fitting the full within-study model. Given that sample sizes within each diagnostic study are typically at least moderately large, it seems reasonable to assume a bivariate normal distribution for $\log \hat{p}$ and $\log \hat{u}$ with means $\log p$ and $\log u$ as well as variances σ_p^2 and σ_u^2, respectively, and covariance σ with $\rho = \sigma/(\sigma_p \sigma_u)$. This is very similar to the assumptions in the approach taken by Reitsma *et al.* (2005) (see also Harbord *et al.* 2007) with the difference that we are using the log-transformation whereas in Reitsma *et al.* (2005) logit transformations are applied. Then, it is a well-known result (Ross, 1985, p. 127) that the conditional mean of the random variable $\log \hat{p}$ (having unconditional mean $\log p$) conditional upon the value of the random variable $\log \hat{u}$ (having unconditional mean $\log u$) is provided as

$$E(\log \hat{p} | \log \hat{u}) = \log p + \rho \frac{\sigma_p}{\sigma_u}(\log(\hat{u}) - \log(u)),$$

which can be written as $\alpha + \theta \log(\hat{u})$, where $\alpha = \log(p) - \theta \log(u)$ and $\theta = \rho \frac{\sigma_p}{\sigma_u}$. This is an *important* result since it means that, in the log-space, sensitivity and false-positive rate are linearly related. Furthermore, if α is zero, the Lehmann model arises.

The question then arises why not work with a straight line model $\log p_{|\log u} = \alpha + \theta \log u$. The answer is that such a model is *not identifiable* since we have only one pair of sensitivity and specificity observed in each study and it is not possible to uniquely determine a straight line by just one pair of observations since there are infinitely many possible lines passing through a given point in the $\log p$–$\log u$ space. However, the Lehmann model as a slope-only model *is* identifiable and it is more plausible than other identifiable models such as the intercept-only model. Clearly, a logistic-transformation would be more consistent with the existing literature (Rutter and Gatsonis, 2001; Walter and Macaskill, 2004) than the log-transformation. However, both models would give a perfect fit (within each study) since there are no degrees of freedom left for testing the model fit. The situation changes when there are repeated observations of sensitivity and specificity *per study* available. These meta-analyses with repeated observations of sensitivity and specificity according to cut-off value variation are very rare, but they exist.

A meta-analysis with repeated observations. One of these rare examples is the CAGE meta-analysis (Aertgeerts *et al.*, 2004) which we will use as a benchmark dataset to investigate for the within-study appropriateness of each model. CAGE is a further instrument for screening the general population for alcohol abuse and dependence. It is a simple instrument consisting of a questionnaire with four questions. What makes this meta-analysis so unique is the fact that for each of the $k = 10$ studies, repeated sensitivities and specificities are provided. The data are documented in Table 7. Here, a straight line model is identifiable on the log-scale as well as on the logistic-scale. We fitted fours models (two for each of the two transformations) for these data: the straight line model (usually not identifiable) and the slope-only model. We use as the standard measure of performance the percentage of explained variance: $R^2 = 1 - \frac{SSE}{SSTOT} \times 100$, where SSE and $SSTOT$ are the usual sum-of-squares from the ANOVA table. The results are presented in Table 8. Note that we used *study* as a categorical covariate, so that an overall performance measure can be presented. We find the performance of the Lehmann model (2.1) remarkably well in comparison to the logistic regression model in the case of an additional intercept parameter (again the latter not being identifiable in most cases). Clearly, the performance of the Lehmann model is superior in the slope-only case which is typically the identifiable case. Here, the logistic model is performing rather poor. As an alternative one could also consider using the complementary log-log transformation on sensitivity and false positive rate, or ultimately, a log-transformation on the θ-parameter of the Lehmann family. We will consider this transformation among others in a simulation study given further below.

Simulation on the choice of transformation. All in all, there are close relationships between the bivariate normal random effects model, the Rutter-Gatsonis model and the Lehmann family with the remaining difference that the Lehmann model works

Table 7 CAGE meta-analysis data (Aertgeerts *et al.*, 2004)

Study	Sensitivity	Specificity
1	0.92	0.73
	0.80	0.93
	0.55	0.98
	0.27	0.99
2	0.87	0.80
	0.66	0.92
	0.43	0.99
	0.19	0.99
3	0.79	0.77
	0.70	0.85
	0.52	0.95
	0.27	0.98
4	0.96	0.68
	0.87	0.84
	0.56	0.96
	0.34	0.99
5	0.61	0.87
	0.46	0.95
	0.24	0.98
	0.11	0.99
6	0.89	0.81
	0.73	0.91
	0.44	0.98
	0.19	0.99
7	0.98	0.75
	0.82	0.9
	0.53	0.97
	0.40	0.99
8	0.71	0.59
	0.53	0.87
	0.27	0.98
	0.09	0.99
9	0.88	0.88
	0.48	0.99
	0.24	0.99
	0.08	0.99
10	0.99	0.37
	0.92	0.62
	0.46	0.88
	0.1	0.99

Table 8 Model performance for CAGE meta-analysis data (Aertgeerts *et al.*, 2004)

Model	$R^2 \times 100\%$
$\log p = \beta \log u$	89.8
$\log[p/(1-p)] = \beta \log[(u/(1-u)]$	16.1
$\log p = \alpha + \beta \log u$	85.8
$\log[p/(1-p)] = \alpha + \beta \log[(u/(1-u)]$	93.6

with log-transformation, whereas the other two use the logit-transformation. What remains is to investigate which transformation provides the best fit. Given the results of the CAGE meta-analysis, it seems reasonable to assume within-study validity of the Lehmann model. Hence, it is desirable to have the arising estimate of the diagnostic accuracy close to normality in distribution. To provide some answer to the question which transformation to use, we looked at the following four cases: the untransformed θ, the log-transformation $\log\theta$, the logit-transformation $\log(\theta/(1-\theta))$ and the complementary log-log transformation $\log(-\log(1-\theta))$, the latter assuming $\theta \in (0, 1)$. In addition, there is the previously mentioned complementary log-log transformation seeing the benefit of bringing the Lehmann model into the framework of a complementary log-log link ($\log\theta = \log(-\log p) - \log(-\log u)$) and, hence, ensuring feasible estimates. A simulation study was designed to mimic the reality of meta-analysis of diagnostic studies. The number of studies k was selected as $k = 25$. Then, sample sizes were generated n_i, m_i arising from a Poisson with mean 25, 50, 100 to mimic sample size variation of the studies involved in the meta-analysis (we only present the case of mean sample size 100 here). A baseline heterogeneity was assumed for the false positive rate in that u_i was sampled from a uniform with interval end 0.05 and 0.5: $u_i \sim U[0.05, 0.5]$. From here the sensitivity p_i was calculated according to the Lehmann model (2.1) and finally y_i was sampled from a binomial with size parameter n_i and event parameter p_i, whereas x_i was sampled from a binomial with size parameter m_i and event parameter u_i. From here the sample of diagnostic accuracy parameters $\hat{\theta}_1, \ldots, \hat{\theta}_k$ as well as the transformations of interest could be determined. We present here the results for $E(n_i) = E(m_i) = 100$ and $\theta = 0.1$ in form of the probability plot. Figure 8 shows the details including the Anderson-Darling test statistic for normality with p-value. For more details on the Anderson-Darling test see Stephens (2006). It can be seen that the results for the untransformed estimates of the diagnostic accuracy parameter are quite satisfactory. The situation changes if the sample sizes *per study* become small. This is illustrated in Figure 9 which is the identical scenario as before with the only difference that we have now on average a sample size of 25 per study. Here the approximation to the normal is less satisfactory. However, in practice of diagnostic studies, sample sizes per study are usually large, with values above 100 not being uncommon.

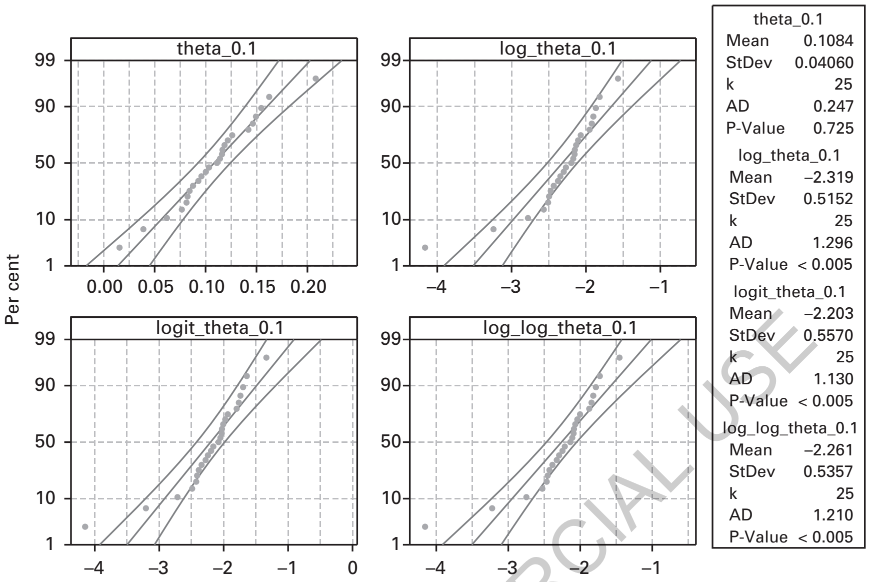

Figure 8 Probability plots for four transformation for simulated data under the Lehmann model with average sample size per study of 100

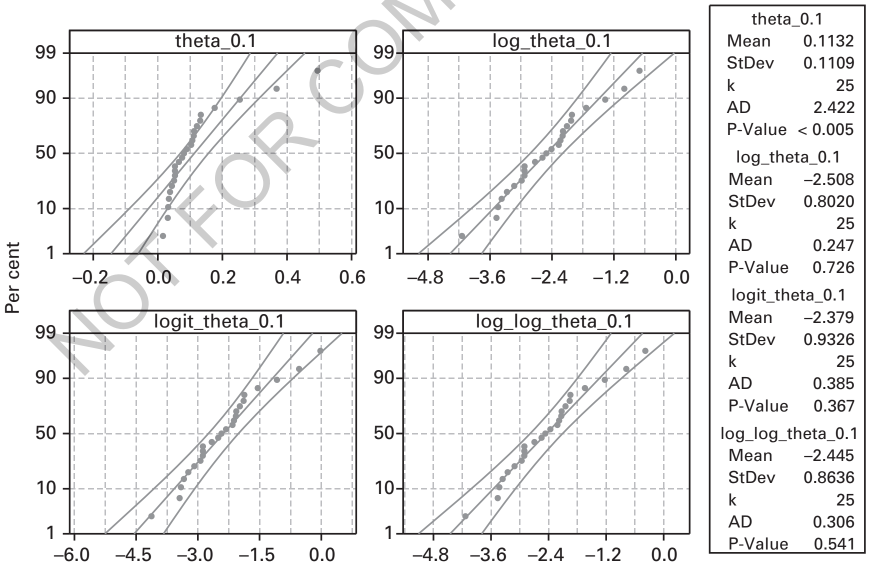

Figure 9 Probability plots for four transformation for simulated data under the Lehmann model with average sample size per study of 25

5 Discussion

Meta-analysis of diagnostic studies is an important subfield requiring special statistical attention. A state-of-the-art analysis requires modelling the SROC curve. Here a simple model, the Lehmann model, was suggested having the beneficial property of being identifiable within each study. In the modelling, study-specific false-positive rates were allowed as nuisance parameters for which then a profile likelihood was derived containing only the parameter of diagnostic accuracy—the parameter of interest. The derivation used initially the normal approximation of the binomial which appears to be justifiable since in most cases of meta-analysis of diagnostic studies, study specific sample sizes are large (often larger than 100). Clearly, this approximation becomes critical if involved studies become sparse.

We come back to the two-level approach by Rutter and Gatsonis (2001) which has been discussed already in the introduction. The model is given as $\log \frac{\pi_{ij}}{1-\pi_{ij}} = (\theta_i + \alpha_i DS_{ij}) \exp(-\beta DS_{ij})$, with the notations as before. It is interesting to compare this model with the Lehmann model. For easiness of comparison, let us consider p and u in the log-space instead of the logit-space. Also, we will use a dummy coding for the disease status variable DS_i. Then, the Rutter-Gatsonis model becomes (in our notation)

$$\log p_i = (\theta_i + \alpha_i) \exp(\beta), \text{ diseased}$$

$$\log u_i = \theta_i, \text{ non-diseased}$$

so that the SROC model (sensitivity as a function of the false positive rate) is

$$\log p_i = (\log u_i + \alpha_i) \exp(\beta).$$

The difference between the Rutter-Gatsonis (RG) model and the Lehmann model becomes clear, if we look at the special case $\beta = 0$. Then, the RG model assumes that the diagnostic accuracy can be represented by differences of log-sensitives and log-false positive rates, whereas the Lehmann model assumes that the diagnostic accuracy can be represented by ratios of the latter. A further analysis for the data on the CAGE meta-analysis (for the sake of brevity not presented here) shows that there is more evidence for a slope-only model than for an intercept-only model.

Finally, we would like to point out that the Lehmann model and the associated inference can be extended in various ways. Here it is crucial that the adjusted profile likelihood is a true normal likelihood. This allows easily to incorporate covariate information such as a variation of the condition of interest (e.g., dementia or MCI), a diagnostic test variation or study-specific properties. These would occur as further fixed effects in the model and associated profile likelihood. In the rare case that repeated observation per study were available this would allow not only validation of the Lehmann model but also the repeated effect to be included in the model. All in all, the Lehmann model with its associated profile likelihood appears to be a flexible approach for coping with the problems of SROC modelling in meta-analysis of diagnostic studies.

Acknowledgements

This work was funded by the *German Research Foundation* (GZ: Ho1286/7-1). We are grateful to the Associate Editor and two referees for their helpful comments.

References

Aertgeerts FB, Buntinx F and Kester A (2003) The value of the CAGE in screening for alcohol abuse and alcohol dependence in general clinical populations: a diagnostic meta-analysis. *Journal of Clinical Epidemiology*, **57**, 30–9.

Aitkin M (1998) Profile likelihood. In Peter Armitage and Theodore Colton (eds), *Encyclopedia of Biostatistics*, volume 5, pp. 3534–36, New York: Wiley.

Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MGM and Stijnen T (2008) Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making*, **28**, 621–38.

Barndorff-Nielsen OE (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–65.

Böhning D, Holling H, Böhning W (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**, 543–54.

Böhning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C and Biggeri A (2002) Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics*, **3**, 445–57.

Cooper H and Hedges L (1994) *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cox DR and Reid N (1987) Parameter orthogonality and approximatic conditional inference. *Journal of the Royal Statistical Society, Series B (Methodological)*, **49**, 1–39.

Deeks JJ (2007) Systematic reviews of evaluations of diagnostic and screening tests. In Matthias Egger, George Davy Smith and Douglas G Altman (eds), *Systematic reviews in health care: meta-analysis in context*, **14**, pp. 248–82, London: BMJ Books.

DerSimonian R and Laird N (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177–88.

Doust JA, Glasziou PP, Pietrzak E and Dobson AJ (2004) A system reviews of diagnostic accuracy of natriyretic peptides for heart failure. *Archive of Internal Medicine*, **2**, 9.

Edwards JH (1966) Some taxonomic implications of a curious feature of the bivariate normal surface. *British Journal of Prevention and Social Medicine*, **164**, 1978—84.

Egger M, Smith GD and Altman DG (2001) *Systematic reviews in health care: Meta-analysis in context*. London: BMJ Publishing Group.

Glas AS, Lijmer JG, Prins MH, Bonsel GJ and Bossuyt PMM (2003) The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, **56**, 1129–35.

Gönen M and Heller G (2010) Lehmann family of ROC curves. *Medical Decision Making*, **30**, 509–17.

Hamza TH, Arends LR, Houwelingen H and Stijnen T (2009) Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. *BMC Medical Research Methodology*, **9**, 73.

Harbord RM, Deeks JJ, Egger M, Whiting P and Sterne JAC (2007) A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, **1**, 1–21.

Hasselblad V and Hedges LV (1995) Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, **117**, 167–78.

Hedges LV and Olkin I (1985) *Statistical methods for meta-analysis*. New York: Academic Press.

Irwig L, Macaskill P, Glasziou P and Fahey M (1995) Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology*, **48**, 119–30.

Jones CM and Athanasiou T (2005) Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. *Annals of Thoracic Surgery*, **79**, 16–20.

Kriston L, Hölzel L, Weiser A, Berner MB and Härter M (2008) Meta-analysis: are 3 questions enough to detect unhealthy alcohol use? *Annals Internal Medicine*, **149**, 879–88.

Krzanowski WJ and Hand DJ (2009) *ROC curves for continuous data*. Boca Raton (FL): CRC Press Taylor & Francis Group, Chapman & Hall.

Le CT (2006) A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research*, **15**, 571–84.

Lee Y, Nelder JA and Pawitan Y (2006) *Generalized linear models with random effects unified analysis via H-likelihood*. Boca Raton (FL): Chapman & Hall/CRC Taylor & Francis Group.

Littenberg B and Moses LE (1993) Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Medical Decision Making*, **13**, 313–21.

Malzahn U, Böhning D and Holling H (2000) Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*, **87**, 619–32.

Mclachlan G and Krishnan T (1997) *The EM algorithm and extensions*. New York: Wiley.

Midgette AS, Stukel TA and Littenberg B (1993) A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. *Medical Decision Making*, **13**, 253–57.

Mitchell AJ (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**, 411–31.

Moses LE, Shapiro D and 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**, 1293–316.

Murphy SA and Van der Vaart AW (2000) On profile likelihood. *Journal of the American Statistical Association*, **95**, 449–85.

Patefield WM (1977) On the maximized likelihood function. *Sankhya, Series B*, **39**, 92–96.

Pepe MS (2000) Receiver operating characteristic methodology. *Journal of the American Statistical Association*, **95**, 308–11.

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press.

Reinert DF and Allen JP (2002) The alcohol use disorders identification test (AUDIT): a review or recent research. *Alcohol: Clinical and Experimental Research*, **26**, 272–79.

Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM and Zwinderman AH (2005) Bivariate analysis of sensitivity and specificity produces informative measures in diagnostic reviews. *Journal of Clinical Epidemiology*, **58**, 982–90.

Ross S (1985) *Introduction to probability models*. Orlando: Academic Press.

Rücker G and Schumacher M (2009) Letter to the editor. *Biostatistics*, **10**, 806–7.

Rücker G and Schumacher M (2010) Summary ROC curve based on a weighted Youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy. *Statistics in Medicine*, **29**, 3069–78.

Rutter CM and Gatsonis CA (2001) A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, **20**, 2865–84.

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

Stephens MA (2006) Anderson-Darling Test of Goodness of Fit. In *Encyclopedia of Statistical Sciences*. New York: Wiley.

Sutton AJ, Abrams KR, Jones DR, Sheldon TA and Song F (2000) *Methods for meta-analysis in medical research*. New York: Wiley.

Swets JA (1996) Signal detection theory and ROC analysis in psychology and diagnostics collected papers. New York: Psychology Press, Taylor & Francis Group.

Van Houwelingen HC, Zwinderman KH and Stijnen T (1993) A bivariate approach to meta-analysis. *Statistics in Medicine*, **12**, 2273–84.

Walter SD and Macaskill P (2004) SROC curve. In Shein-Chung Chow (ed.), *Encyclopedia of biopharmaceutical statistics*, New York: Marcel Dekker.

Youden D (1950) Index for rating diagnostic tests. *Cancer*, **3**, 32–5.