PSYCHOMETRIKA—VOL. 77, NO. 1, 106–126
JANUARY 2012
DOI: 10.1007/S11336-011-9236-2

LIKELIHOOD-BASED CLUSTERING OF META-ANALYTIC SROC CURVES

HEINZ HOLLING AND WALAILUCK BÖHNING

UNIVERSITY OF MÜNSTER

DANKMAR BÖHNING

UNIVERSITY OF SOUTHAMPTON

Meta-analysis of diagnostic studies experience 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 rates (1-specificity). In the context of meta-analysis one pair represents one study and the associated diagram is called an SROC curve where the *S* stands for "summary". In meta-analysis of diagnostic studies emphasis has traditionally been placed on modelling this SROC curve with the intention of providing a summary measure of the diagnostic accuracy by means of an estimate of the summary ROC curve. Here, we focus instead on finding sub-groups or components in the data representing different diagnostic accuracies. The paper will consider modelling SROC curves with the Lehmann family which is characterised by one parameter only. Each single study can be represented by a specific value of that parameter. Hence we focus on the distribution of these parameter estimates and suggest modelling a potential heterogeneous or cluster structure by a mixture of specifically parameterised normal densities. We point out that this mixture is completely nonparametric and the associated mixture likelihood is well-defined and globally bounded. We use the theory and algorithms of nonparametric mixture likelihood estimation to identify a potential cluster structure in the diagnostic accuracies of the collection of studies to be analysed. Several meta-analytic applications on diagnostic studies, including AUDIT and AUDIT-C for detection of unhealthy alcohol use, the mini-mental state examination for cognitive disorders, as well as diagnostic accuracy inspection data on metal fatigue of aircraft spare parts, are discussed to illustrate the methodology.

Key words: C.A.MAN, diagnostic testing, meta-analysis, sensitivity, specificity, summary receiver operating characteristic (SROC), summary statistics approach.

1. Introduction

Meta-analysis of diagnostic studies deals with the following situation. A number of studies are available each providing an estimate of the sensitivity of the diagnostic test (the probability that the test is positive given the person has the disease or the condition of interest) and an estimate of specificity (the probability that the test is negative given the person does not have the disease or the condition of interest). Available approaches in this setting focus on the summary receiver operating characteristic (SROC) curve. This is done because a simple summary approach of sensitivity and specificity can be largely misleading. One of the reasons for this potential bias is the cut-off value problem (described further below in detail) which, in essence, means that biased inference occurs if the unobserved cut-off value variation is ignored and not adjusted for. A widely accepted way to proceed is to focus on the SROC curve, and previous approaches have provided various ways of doing so (Sutton,

Requests for reprints should be sent to Dankmar Böhning, School of Mathematics and Southampton Statistical Sciences Research Institute, University of Southampton, B39, R2021, Southampton, SO17 1BJ, UK. E-mail: D.A.Bohning@soton.ac.uk

Abrams, Jones, Sheldon, & Song, 2000, Chapter 14; Egger, Smith, & Altman, 2001, Chapter 14). Littenberg and Moses suggested to find an SROC curve by regressing the log-diagnostic odds ratio onto a measure for cut-off value variation (Moses, Littenberg, & Shapiro, 1993; Midgette, Stukel, & Littenberg, 1993) which has been generalised in the *hierarchical* receiver operating characteristic (HSROC) by Rutter and Gatsonis (2001) by incorporating random effects distributions on the parameters of the logistic regression model. A seemingly different approach has been suggested by Reitsma, Glas, Rutjes, Scholten, Bossuyt, and Zwinderman (2005) adapting a bivariate normal regression model on the true logit-sensitivity and the true logit-specificity. Harbord, Deeks, Egger, Whiting, and Stern (2007) point out the close similarities between the HSROC and the bivariate normal random effects approach. In principle, all these approaches focus on providing an estimate of the SROC curve. Here, we focus instead on finding sub-groups or clusters in the data which are similar in their diagnostic accuracy *within* each cluster or component, but are different in their diagnostic accuracy *across* clusters or components. This will not improve the diagnostic performance of the diagnostic device itself, but will deliver additional insights into sources of potential variation in diagnostic accuracy. For example, the various located clusters of diagnostic accuracy might be matched to a variation of the "gold standard" defined as the way the disease or condition has been confirmed. In another scenario the estimated clusters can be matched to the personnel applying the diagnostic device, and a need for additional training in applying the diagnostic device might be indicated.

More precisely, we are interested in the following situation of meta-analysis of diagnostic studies: a variety of diagnostic studies are available providing estimates of the diagnostic measures of specificity $(1-u) = P(T=0|D=0)$ in the ith study as $\hat{u}_i = x_i/n_i$ (estimate of false-positive rate) and of sensitivity $p = P(T=1|D=1)$ in the ith study as $\hat{p}_i = y_i/m_i$ (estimate of sensitivity), where $D=1$ and $D=0$ denote presence or absence of the condition of interest, respectively, and $T=1$ or $T=0$ denote positivity (indicating presence of the condition) or negativity (indicating absence of the condition) of the diagnostic test, respectively. Also, x_i are the number of false-positives out of n_i individuals without the condition, y_i are the number of true positives out of m_i individuals with the condition, for $i=1,\ldots,k$, k being the number of studies. For a more general introduction to meta-analysis of diagnostic studies see Sutton et al. (2000; Chapter 14), Egger et al. (2001, Chapter 14), or Böhning, Holling, and Böhning (2008). We illustrate the situation with a meta-analysis on the Michigan Alcoholism Screening Test (MAST) for alcohol problems.

Example 1 (Meta-Analysis of Diagnostic Accuracy of the Michigan Alcoholism Screening Test (MAST) for Alcohol Problems). Storgaard, Nielsen, and Gluud (1994) provide a meta-analysis on the diagnostic accuracy of the Michigan Alcoholism Screening Test (MAST) for alcohol problems which we report here in Table 1. The MAST was developed by Selzer (1971) in the U.S.A. and was originally composed of 25 questions. The result of the test is a score, and a cut-off value is used to decide if the test is positive (presence of alcohol problems) or negative (absence of alcohol problems). See also Martin, Liepman, and Young (1990) for further details.

1.1. 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 (as in the case of the MAST), 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 is also called the *cut-off value problem*.

TABLE 1.
Meta-analysis of diagnostic accuracy of the Michigan Alcoholism Screening Test (MAST) for alcohol problems. (TP = true positive; FP = false positive; TN = true negative; FN = false negative; See Storgaard et al. 1994, for reference citations for listed studies.)

Study i	Alcohol problems		No alcohol problems		$n_i + m_i$
	y_i (TP)	$m_i - y_i$ (FN)	$n_i - x_i$ (TN)	x_i (FP)	
Moore 1972	125	3	192	31	351
McAuley 1978	14	1	35	25	75
Zung 1980	9	1	29	52	91
Zung 1982	21	3	48	48	120
Searles 1990	36	5	9	20	70
Benussi 1982	56	0	45	3	104
Yersin 1989	38	16	197	17	268
Selzer 1971	114	2	98	5	219
Breitenbucher 1976	60	10	138	44	252
Rounsaville 1983	79	60	216	30	385
Magruder–Habib 1983	63	29	222	55	369
Mischke 1987	23	20	43	4	90
Garzotto 1988	72	3	69	8	152
Sokol 1989	15	27	892	37	971
Ross 1990	240	5	146	110	501
Zung 1982	20	4	62	34	120
Zung 1982	20	4	66	30	120
Zung 1982	20	4	72	24	120
Rounsaville 1983	50	10	267	58	385

This situation is illustrated in Figure 1 for a continuous outcome T which is normally distributed in the two populations. Moving the cut-off value c in Figure 1 will clearly change sensitivity and specificity (in opposite ways), whilst the ability of the test to separate the two populations is unchanged. Hence it is often argued that a meta-analysis of diagnostic studies requires the cut-off value to be invariant across studies. This requirement, however, does not let the cut-off problem disappear entirely. This becomes clear when thinking of a situation in which the cut-off value is kept identical but the populations with and without the condition experience some shift. Clearly, this affects sensitivity and specificity in a similar manner as a variation of the cut-off itself would.

1.2. Background of Meta-analysis for Diagnostic Studies

Because of this comparability problem for sensitivity and specificity interest is usually focussed on the *summary receiver operating characteristic* (SROC) 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 of the MAST data 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 concluded that the pairs might stem from one common SROC curve, from two or more different SROC curves. Hence we are interested in exploring the *cluster structure* of the SROC diagram. This is approached in the following way. Consider the pair $(\hat{p}_i, \hat{u}_i)$

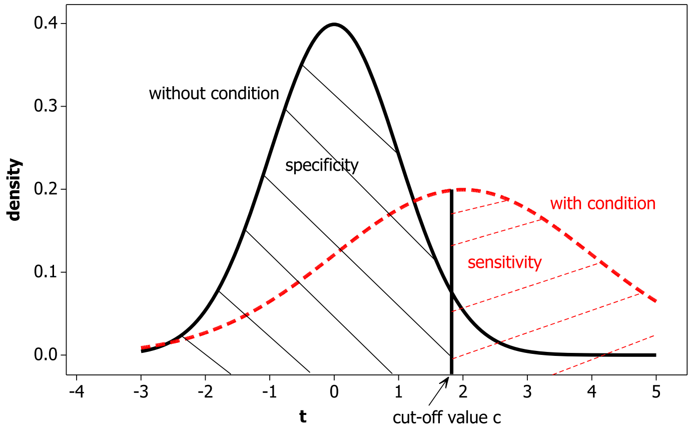

FIGURE 1.
Diagnostic situation illustrated with two normal distributions: one has mean 0 and variance 1 (population without the condition), the other has mean 2 and variance 4 (population with the condition).

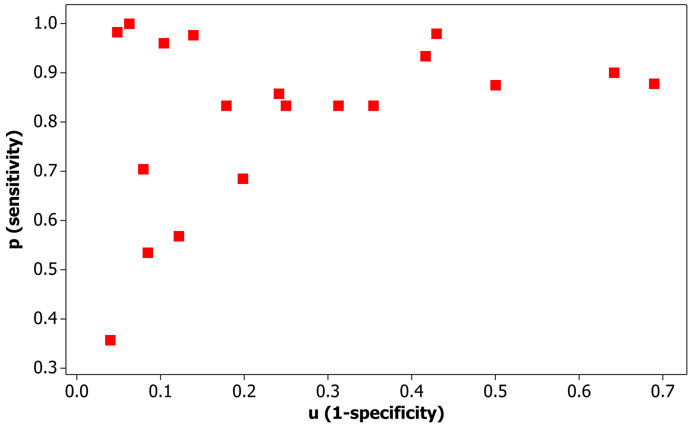

FIGURE 2.
SROC diagram for meta-analysis of the MAST data of Table 1.

of estimated sensitivity and estimated false-positive rate in study i, for $i = 1, \ldots, k$. We will argue in the next section that $\hat{\theta}_i = \log \hat{p}_i / \log \hat{u}_i$ is a reasonable summary measure of the diagnostic accuracy of study i. Note that we are referring to a summary measure of sensitivity and specificity for *one study* not to a summary measure over all studies. Ultimately, we will base our analysis on $\log \hat{\theta}_i = \log(-\log \hat{p}_i) - \log(-\log \hat{u}_i)$ which is interesting for at least two reasons. First, it brings the proposed accuracy measure into the framework of a well-known link function: $\log \hat{\theta}_i$ is the difference of sensitivity and specificity of study i in the complementary log–log link-function space. Note that the complementary log–log transformation is a popular link function in generalised linear models (McCullagh & Nelder, 1989). Second, as we will argue later, $\log \hat{\theta}_i$ is typically closer to a normal distribution than $\hat{\theta}_i$ itself. Now, given diagnostic accuracy measures $\log \hat{\theta}_i$, for $i = 1, \ldots, k$, one could proceed in a conventional meta-analytic way by assuming a normal distributional model for $\log \Theta_i$ (representing the random variable with realisation $\log \hat{\theta}_i$) as follows:

$$\log \Theta_i = \mu + \delta_i + \epsilon_i. \tag{1}$$

Here μ is a fixed parameter, whereas δ_i is a normal random effect with mean zero and unknown variance τ^2, and ϵ_i is a normal random error with mean zero and known study-specific variance σ_i^2, with random effect and random error being independent. This is the meta-analytic *random effects* model which goes back to DerSimonian and Laird (1986) and has received various modifications (primarily with respect to estimation). For an overview see Sidik and Jonkman (2005). If we assume $\tau^2 = 0$, we obtain the simple *fixed effects* model. Whether we are in the random effects or in the fixed effects model, the estimate for μ will provide an overall measure of the diagnostic accuracy of the diagnostic test or device under study. Inference between the fixed effects and the random effects models differs typically with respect to the confidence interval which is larger for the latter. Although this is an interesting approach that allows the reduction of a complex, bivariate problem to a simpler, univariate one, we do not wish to explore this further. Instead, we intend to propose a novel approach in this context which allows the estimation of the random effects distribution itself (and not restricting ourselves to first and second moments as the DerSimonian–Laird method does). We will follow the nonparametric likelihood approach (Böhning, 2000; Lindsay, 1995; Laird, 1978) to find an estimate of the random effects distribution. This estimate will always be discrete and hence has a *cluster* interpretation. These clusters of different diagnostic accuracy might have different interpretations; for example, they might represent different gold standards which have been used in different studies. In another situation, the condition is varied for which the same diagnostic test is used which then might explain the observed cluster structure. For example, the Mini-Mental-State Examination is used to diagnose schizophrenia. However, it is also used to diagnose mild cognitive impairment—with a lower diagnostic accuracy. Hence, if it is possible in practice to match relevant covariates to estimated clusters, the proposed likelihood-based clustering procedure will help practitioners to evaluate a given diagnostic test.

2. The Lehmann Model

Le (2006) suggests modelling the relationship between sensitivity and false-positive rate using the Lehmann family,

$$p = u^\theta. \tag{2}$$

The Lehmann model has a number of nice properties including that $p \in [0, 1]$ if $u \in [0, 1]$ for $\theta > 0$. Recall that $p = P(T = 1|D = 1)$ is the sensitivity and $u = P(T = 1|D = 0)$ is the true false-positive rate of the diagnostic test. Hence, the Lehmann model represents a feasible reparameterisation 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. Also, two diagnostic 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 $\mathrm{AUC} = \int_0^1 u^\theta\, du = 1/(1+\theta)$. Model (2) is also called a *proportional hazards model* (PHM) since it assumes that the ratio of log-true-positive rate and log-false-positive rate is constant: $\log p(t)/\log u(t) = \theta$.

There are various reasons why model (2) 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 (on log-scale) this pair is represented by one point. Clearly, infinitely many lines pass through this point; in other words, a straight line model is *not identifiable* within study i. However, a straight line that passes through the origin is uniquely characterised 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. However, it is also clear that it is not the only identifiable model in this situation. To mention at least one further example, consider $p = \alpha^* u$ which is in the log space $\log p = \log \alpha^* + \log u = \alpha + \log u$. Here the model has a free intercept parameter and a fixed slope parameter constrained to be one. However, the model is less appealing than the PHM: α^* is restricted to values between 0 and 1, and p will be bounded by a number less than one if α^* is smaller than one which is undesirable.

In the following we are interested in inference for θ in the PHM (2). Note that we are able to fit exactly one model (2) to each of the k pairs $(\hat{p}_i, \hat{u}_i)$ to yield k ROC curves. Under (2) it is possible to write $\theta = \log p / \log u$, so that we are able to construct k estimates of θ and $\log \theta$, namely $\hat{\theta}_i = \log \hat{p}_i / \log \hat{u}_i = z_i / w_i$ and $\log \hat{\theta}_i = \log(-z_i) - \log(-w_i)$, for $i = 1, \dots, k$. Here, 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*. Under the assumption that these $\hat{\theta}_i$ and $\log \hat{\theta}_i$ are realisations of a random variable Θ_i and $\log \Theta_i$, respectively, we are able to base our analysis on these *summary statistics*. Note that these estimates, $\hat{\theta}_i$ and $\log \hat{\theta}_i$, are available for each individual study; and the parameter in model (2) is identifiable in each study. This is in contrast to other approaches which use more than one parameter to model sensitivity and false-positive rate (Rutter & Gatsonis, 2001).

It remains to derive estimates of $\mathrm{Var}(\log \Theta_i)$. For what follows we will use that the associated estimated variances for the log-proportions are provided as

$$\widehat{\mathrm{Var}}(\log \hat{p}_i) = \widehat{\mathrm{Var}}\big(\log(y_i/m_i)\big) = s_i^2 = \frac{1}{y_i} - \frac{1}{m_i}, \tag{3}$$

$$\widehat{\mathrm{Var}}(\log \hat{u}_i) = \widehat{\mathrm{Var}}\big(\log(x_i/n_i)\big) = t_i^2 = \frac{1}{x_i} - \frac{1}{n_i}, \tag{4}$$

assuming that $y_i > 0$ and $x_i > 0$ for $i = 1, \dots, k$. We use the δ-method to achieve

$$\widehat{\mathrm{Var}}(\log \Theta_i) \approx \widehat{\mathrm{Var}} \log(-z_i) + \widehat{\mathrm{Var}} \log(-w_i) \tag{5}$$

$$\approx \left(\frac{s_i^2}{z_i^2} + \frac{t_i^2}{w_i^2}\right) = \frac{s_i^2}{w_i^2} + \frac{z_i^2 t_i^2}{w_i^4} := \sigma_i^2. \tag{6}$$

There are two problems with these estimated variances. Typically, samples sizes per study are large, so that these estimates are expected to be fairly accurate. However, some of the x_i or y_i could be zero. This can be handled by replacing 0 by 0.5 and increasing the associated sample size by one. Similarly, if some of the x_i or y_i are equal to n_i or m_i, respectively, the associated sample size is increased by one as well.

3. Nonparametric Mixture Model

We assume that we have k realisations $\hat{\theta}_1, \dots, \hat{\theta}_k$ of associated k random variables $\Theta_1, \dots, \Theta_k$ available with known variances $\sigma_1^2, \dots, \sigma_k^2$, respectively. To model the cluster structure involved in SROC diagram with the k observations, we assume the following mixture model for the $\log \Theta_i$:

$$\log \Theta_i \sim \sum_{j=1}^{J} q_j \frac{1}{\sigma_i} \phi\left(\frac{\log \theta - \lambda_j}{\sigma_i}\right), \tag{7}$$

where $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ is the standard normal density and q_j are non-negative weights summing to one. We use here the log transformation since we found a better fit under homogeneity ($J = 1$) for the log-normal in comparison to the normal distribution. Note that also $\lambda = \log \theta$ is on the log-scale of diagnostic accuracy.

The mixture kernel in (7) has good properties. Note that the variances are known and study-specific. Hence we have a *bounded* mixture likelihood and a global maximum can be found. This is in contrast to problems where the mixture kernel has unknown variance parameters in each component and the likelihood becomes unbounded (Böhning, 2000; Lindsay, 1995). Also, there is a problem in the case of a common variance parameter since if the number of components J increases the variance parameter estimate becomes arbitrarily small and the problem ill-posed (Böhning, 2000). All these problems do not exist here since the underlying binomial structure induces a variance–mean functional relationship that we exploit here. Hence we can consider the likelihood

$$L = L(Q) = \prod_{i=1}^{k} \sum_{j=1}^{J} q_j \frac{1}{\sigma_i} \phi\left(\frac{\log \hat{\theta}_i - \lambda_j}{\sigma_i} \right) = \prod_{i=1}^{k} \sum_{j=1}^{J} q_j f_i(\log \hat{\theta}_i, \lambda_j), \tag{8}$$

where $f_i(x, \lambda) = \frac{1}{\sigma_i} \phi(\frac{x-\lambda}{\sigma_i})$. Note that the mixture kernel has a study-specific form since a study-specific variance term is included. The likelihood $L(Q)$ is a function of $2J - 1$ parameters, and Q is typically written as a parameter matrix with J columns and 2 rows,

$$Q = \begin{pmatrix} \lambda_1, \ldots, \lambda_J \\ q_1, \ldots, q_J \end{pmatrix},$$

where $\lambda_1, \ldots, \lambda_J$ are the component means and $q_1, \ldots, q_J$ the component weights. Note that there are only $J - 1$ free weights because of the restriction $q_1 + \cdots + q_J = 1$. We can also interpret Q as a discrete probability distribution given weight q_j to component mean θ_j.

As a sideline note that $E(Q) = \sum_{j=1}^{J} q_j \lambda_j = \mu$ and $\text{Var}(Q) = \sum_{j=1}^{J} q_j (\lambda_j - \mu)^2 = \tau^2$ where μ and τ^2 are identical to those given in (1). Having estimated Q this will also supply estimates of μ and σ^2. Hence, we are also able to provide a pooled estimate of diagnostic accuracy with a confidence interval *without* assuming normality for the random effects distribution.

We further point out that J is not fixed but is itself estimated. The likelihood $L(Q)$ needs to be maximised in the set Ω of all discrete probability measures:

$$\Omega = \left\{ Q = \begin{pmatrix} \lambda_1, \ldots, \lambda_J \\ q_1, \ldots, q_J \end{pmatrix} \middle| J = 1, 2, 3, \ldots \right\}.$$

The log-likelihood $\log L(Q)$ is concave on Ω and can be globally maximised. The associated mixture distribution $\hat{Q}$ which maximised the likelihood globally is called the *nonparametric maximum likelihood estimator* (NPMLE). Fundamental work on the geometry of the mixture likelihood has been provided by Lindsay (1983) although the consistency result goes back already to Kiefer and Wolfowitz (1956). It has been applied in a variety of areas including nonparametric random effects modelling (Aitkin, 1999a; Skrondal & Rabe-Hesketh, 2004; Rabe-Hesketh, Pickles, & Skrondal, 2003), meta-analysis (Aitkin, 1999b; Böhning, 2000) or disease mapping (Clayton & Kaldor, 1987). An important tool for finding the NPMLE is the *gradient function,* defined as

$$d(\lambda, Q) = \frac{1}{k} \sum_{i=1}^{k} \frac{f_i(\log \hat{\theta}_i, \lambda)}{\sum_{j=1}^{J} q_j f_i(\log \hat{\theta}_i, \lambda_j)}. \tag{9}$$

A fundamental result states that $\hat{Q}$ is NPMLE if and only if the gradient function is bounded by 1: $d(\lambda, \hat{Q}) \leq 1$ (Lindsay, 1983). This results allows us to check in a simple way if the NPMLE has been reached.

Computationally, the NPMLE can be easily computed using the software C.A.MAN (Böhning, Dietz, & Schlattmann, 1998). Note that the NPMLE is always discrete, and usually only a

TABLE 2.
Likelihood-based cluster analysis of the MAST meta-analysis: Cluster analysis starts with the NPMLE ($J = 4$) and provides for each number of components J the estimated mixing distribution with associated log-likelihood, AIC and BIC.

Study	n	P	J	Weight	Mean	log-L	AIC	BIC
MAST	19	7	4	0.1281	−4.2214	−23.3376	60.68	67.29
				0.1367	−4.1948			
				0.3186	−2.0515			
				0.4166	−1.3548			
		5	3	0.2648	−4.2075	−24.3373	58.67	63.40
				0.3186	−2.0515			
				0.4166	−1.3548			
		3	2	0.2675	−4.2073	−26.2915	**58.58**	**61.42**
				0.7325	−1.5370			
		1	1	1.0000	−1.7474	−58.9445	119.89	120.83

small number of components J are required. Hence, for inferential purposes we compute mixture models for all number of components starting with $J = 1$, $J = 2$, and so forth until the NPMLE is reached. For fixed number of components J we use the EM algorithm (Dempster, Laird, & Rubin, 1977) with gradient function update as described in Böhning (2003) to avoid sub-optimal local maxima.

The number of components J needs to be selected. We look at two criteria: the *Akaike Information Criterion* (AIC), defined as

$$\text{AIC} = -2\log L + 2(2J - 1),$$

and the *Bayesian Information Criterion* (BIC), defined as

$$\text{BIC} = -2\log L + (2J - 1)\log k.$$

These are the simplest ones out of a huge diversity of possible choices (Ray & Lindsay, 2008; McLachlan & Peel, 2000). Note that both criteria involve the number of unknown parameters P in the model: $P = 2J - 1$. Empirical evidence based upon simulation work has shown that the BIC often performs best, whereas the AIC suffers under similar singularity problems as the likelihood ratio. Nevertheless we report both criteria for all data analysed here.

Table 2 shows the results achieved by C.A.MAN for all components until the NPMLE is reached with $J = 4$ components. Both criteria select the mixture model with $J = 2$ components. Note that 27% of the studies go to the component with high diagnostic accuracy, whereas the rest are allocated to the component with lower diagnostic accuracy.

The question remains how the individual studies are actually *classified* into the various components. We follow the conventional *maximum posterior probability* rule. If one thinks of the estimate $\hat{Q} = \binom{\hat{\lambda}_1, \ldots, \hat{\lambda}_J}{\hat{q}_1, \ldots, \hat{q}_J}$ as an estimated prior distribution in an empirical Bayesian sense, then

$$\hat{f}_{ij} = \frac{f_i(\log\hat{\theta}_i, \hat{\lambda}_j)}{\sum_{\ell=1}^{J} q_\ell f_i(\log\hat{\theta}_i, \hat{\lambda}_\ell)} \tag{10}$$

is the *estimated* posterior for study i. Note that $\hat{f}_{j|i} = \hat{f}_{ij}$ is a discrete probability distribution for each study i, $i = 1, \ldots, k$. The maximum posterior probability (MAP) rule assigns study i into component j for which

$$\hat{f}_{j|i} = \max_{\ell=1}^{J} \hat{f}_{\ell|i}.$$

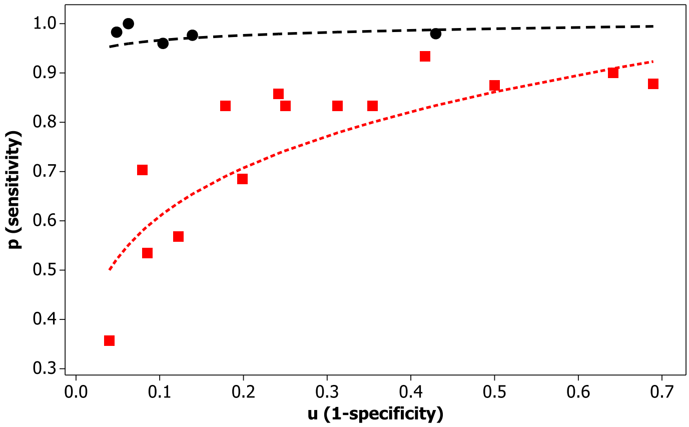

FIGURE 3.
SROC diagram for meta-analysis of the MAST data of Table 1 with classifications of studies into mixture components according to the MAP rule; the dashed, black curve $p = u^{\theta}$ represents the high accuracy component with $\theta = \exp(-4.21)$ whereas the dotted, red curve $p = u^{\theta}$ represents the lower accuracy component with $\theta = \exp(-1.54)$.

Figure 3 shows the classification of the studies of MAST meta-analysis into the two components.

4. Applications

4.1. AUDIT and AUDIT-C for Alcohol Disorders

One of the most frequently recommended instruments (including a recommendation from the WHO) for screening for all forms of unhealthy alcohol use (risky drinking, alcohol abuse, alcohol dependence) is the Alcohol Use Disorder Identification Test (AUDIT). The full AUDIT consists of 10 items and has been extensively investigated in several settings and countries (Reinert & Allen, 2002). Here we look at a meta-analysis provided by Kriston, Hölzel, Weiser, Berner, and Härter (2008). The data are provided in Table 7 in the Appendix, and the associated SROC curves are provided in Figure 4. 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 3-item version is also appropriate to screen for unhealthy alcohol use (Reinert & Allen, 2002). In Table 7 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 5. Both meta-analyses show a similar cluster structure with $J = 3$ components or clusters being identified. Both show a similar component with low diagnostic accuracy. Both have also a similar mean diagnostic accuracy as estimated by the mean of the respective mixing distribution, but evidently the mixing distribution for AUDIT-C has the large variance, hence AUDIT-C shows more heterogeneity.

4.2. Mini-mental State Examination for Dementia

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. The data are reproduced in Table 8 in the Appendix in a form that allows a reanalysis with the methods developed here.

TABLE 3.
Likelihood-based cluster analysis of the AUDIT and AUDIT-C meta-analysis: Cluster analysis starts with the NPMLE ($J = 4$ in both cases) and provides for each number of components J the estimated mixing distribution with associated log-likelihood, AIC and BIC.

Study	n	P	J	Weight	Mean	log-L	AIC	BIC
AUDIT	14	7	4	0.3577	−2.6472	−8.1679	30.3357	34.8091
				0.4187	−2.2287			
				0.2164	−1.7817			
				0.0071	−1.7813			
		5	3	0.3577	−2.6472	−8.7168	**27.4336**	**30.6289**
				0.4187	−2.2290			
				0.2236	−1.7817			
		3	2	0.4053	−2.6365	−12.1819	30.3638	32.2810
				0.5947	−2.0797			
		1	1	1.0000	−2.1523	−14.9285	31.8569	32.4960
AUDIT-C	14	7	4	0.0880	−6.3314	−17.9138	49.8275	64.3009
				0.1648	−3.5694			
				0.4516	−2.4488			
				0.2956	−1.7164			
		5	3	0.1121	−6.1367	−18.3351	**46.6702**	**49.8655**
				0.5872	−2.5170			
				0.3007	−1.7180			
		3	2	0.6913	−2.5676	−20.3757	46.7513	48.6685
				0.3087	−1.7208			
		1	1	1.0000	−1.9724	−30.2533	62.5065	63.1456

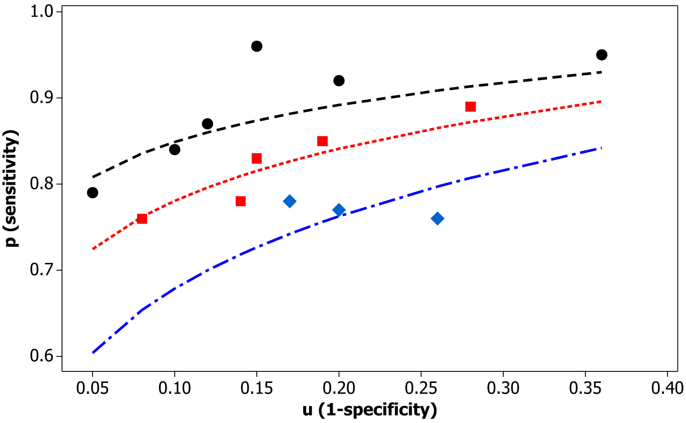

FIGURE 4.
SROC diagram for meta-analysis of the AUDIT data of Table 3 with classifications of studies into mixture components according to the MAP rule; the three curves represent the Lehmann model $p = u^{\theta}$ for the three component means found in Table 3 with $\theta_j = \exp(\lambda_j)$ for $j = 1, 2, 3$.

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. This let $k = 33$ remain in the meta-analysis for further evaluation. The NPMLE is found here with $J = 3$ which is also the best choice according to both AIC and BIC. For details see Table 4. About 37% of the studies are allocated to a low diagnostic accuracy component of $\theta = \exp(-1.63) = 0.20$ which corresponds to an AUC of 0.84, whereas 30% go to a component of moderate good

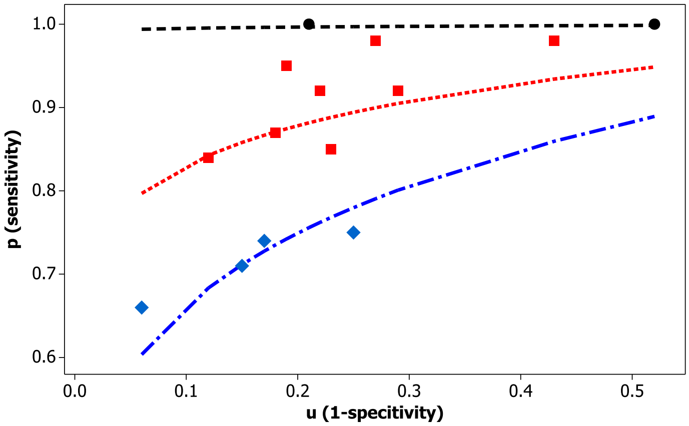

FIGURE 5.
SROC diagram for meta-analysis of the AUDIT-C data of Table 3 with classifications of studies into mixture components according to the MAP rule; the three curves represent the Lehmann model $p = u^{\theta}$ for the three component means found in Table 3 with $\theta_j = \exp(\lambda_j)$ for $j = 1, 2, 3$.

TABLE 4.
Likelihood-based cluster analysis of the MMSE-Meta-Analysis: cluster analysis starts with the NPMLE ($J = 3$) and provides for each number of components J the estimated mixing distribution with associated log-likelihood, AIC and BIC.

Study	n	P	J	Weight	Mean	log-L	AIC	BIC
DEMENTIA	33	5	3	0.3332	−2.9703	−33.2562	**76.5125**	**83.9950**
				0.2928	−2.3585			
				0.3740	−1.6348			
		3	2	0.5239	−2.7469	−39.8383	85.6766	90.1661
				0.4761	−1.6964			
		1	1	1.0000	−2.1400	−91.5335	185.0670	186.5635

accuracy of $\theta = \exp(-2.36) = 0.09$ corresponding to an AUC of 0.91, and 33% are allocated to a component with high accuracy of $\theta = \exp(-2.97) = 0.05$ corresponding to an AUC of 0.95. The associated SROC curves are provided in Figure 6.

4.3. *Meta-analysis on Inspection Data of the Diagnostic Accuracy of Technicians for Inspecting Aircraft Specimens for Metal Fatigue*

Our final example shows a different perspective on how this methodology can be used to identify groups of people with different ability in fulfilling their assigned objectives. This can be crucial in the example we are looking at, since the safety of numerous people will depend on the accurate execution of assigned tasks. Also, it might be helpful in identifying groups of people with a need for additional training and educations. Swets (2009, p. 137) discusses data on inspection of aircraft for metal fatigue. In this study, 148 meta specimens *with and without flaws* were carried to 17 bases of the U.S. AirForce, where they were inspected for defects. Two techniques were used: ultra-sound and eddy-current, the latter showing the higher diagnostic accuracy so that we concentrate here on this technique. We were able to reconstruct the data for 106 technicians from the figures given in Swets (2009, p. 138) and assume that the 27 missing data are missing at random, so that no systematic bias can be expected. In contrast to the previous analysis, here the technician takes the role of the study and each technician produces 148 test data

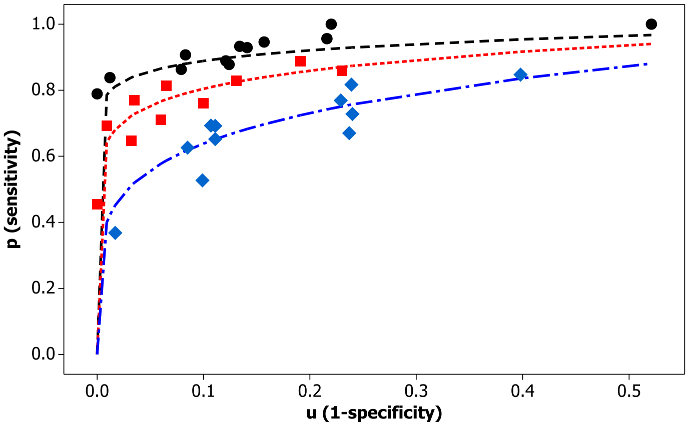

FIGURE 6.
SROC diagram for meta-analysis of the MMSE data of Table 4 with classifications of studies into mixture components according to the MAP rule; the three curves represent the Lehmann model $p = u^{\theta}$ for the three component means found in Table 4 with $\theta_j = \exp(\lambda_j)$ for $j = 1, 2, 3$.

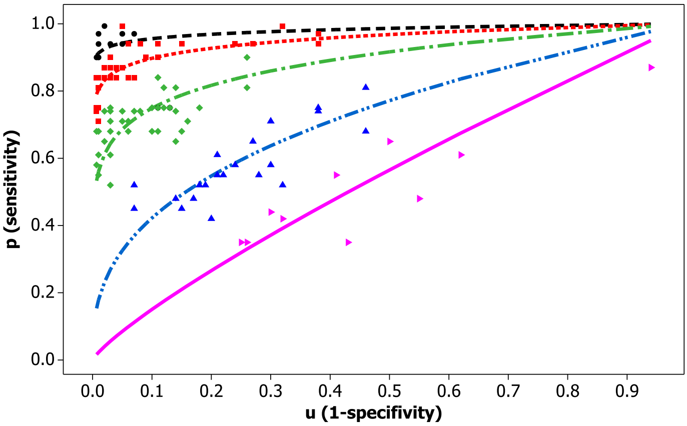

FIGURE 7.
SROC diagram for meta-analysis of the inspection data of Table 5 with classifications of studies into mixture components according to the MAP rule; the five curves represent the Lehmann model $p = u^{\theta}$ for the five component means found in Table 5 with $\theta_j = \exp(\lambda_j)$ for $j = 1, \ldots, 5$.

(item faulty or item not-faulty). Hence we are interested in investigating the performance of the 106 technicians with respect to their heterogeneity in diagnostic accuracy.

The analysis is provided in Table 5. There is clear evidence of a cluster structure consisting of $J = 5$ components. This is also the NPMLE which is clearly confirmed by both AIC and BIC. There are two groups with high and moderately high diagnostic accuracy, $\exp(-3.93) = 0.02$ and $\exp(-3.06) = 0.05$, corresponding to AUCs of 0.98 and 0.96. In contrast there are groups of technicians with very low and low diagnostic accuracy, $\exp(-0.19) = 0.83$ and $\exp(-0.98) = 0.38$, corresponding to AUCs of 0.55 and 0.73. These two groups appear to be well separated from the rest. Note that the smallest attainable value for the AUC is 0.5 which is achieved for $\theta = 1$. This implies that the diagnostic accuracy in Group 5 is indeed very low (see Figure 7) and, if this would occur in a timely application, this group would definitely need additional training.

TABLE 5.
Likelihood-based cluster analysis of the Meta-Analysis on inspection data of the diagnostic accuracy of 106 technicians using the eddy-current technique for inspecting aircraft specimen for metal fatigue: cluster analysis starts with the NPMLE ($J = 3$) and provides for each number of components J the estimated mixing distribution with associated log-likelihood, AIC and BIC.

Study	n	P	J	Weight	Mean	log-L	AIC	BIC
eddy-current	106	9	5	0.0839	−3.9267	−165.72	**349.45**	**373.42**
				0.2715	−3.0576			
				0.3480	−2.0740			
				0.2000	−0.9821			
				0.0966	−0.1947			
		7	4	0.1005	−3.8594	−185.37	384.75	403.39
				0.2756	−2.9870			
				0.3468	−2.0066			
				0.2770	−0.6806			
		5	3	0.3432	−3.2224	−186.94	383.89	397.20
				0.3772	−2.0473			
				0.2796	−0.6853			
		3	3	0.6961	−2.4466	−236.95	479.91	487.90
				0.3039	−0.7392			
		1	1	1.0000	−1.6415	−452.54	907.07	909.74

5. Model Fit and Appropriateness of Approach

The proposed approach is not without assumptions. Two aspects need to be clearly distinguished. The first aspect is that the approach uses the summary measure θ of sensitivity and false-positive rate for *each* of the available studies, so that for k studies we achieve k summary estimates $\hat{\theta}_1, \dots, \hat{\theta}_k$. This is equivalent of assuming the validity of Equation (2), $p = u^\theta$, *within each of the k studies*. The second aspect is that, conditional upon the validity of the summary measure, it is assumed that $\log \hat{\theta}_i$ is normally distributed *within each study i*, but allowing arbitrary heterogeneity between studies. We comment on both aspects in the following.

Appropriateness of the Lehmann Model Within Each Study 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

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

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

TABLE 6.
Performance of the three straight line models for the 10 studies involved in the CAGE meta-analysis data (Aertgeerts et al., 2004): full model, $\log p = \alpha + \theta \log u$; slope-only model, $\log p = \theta \log u$; and intercept-only model, $\log p = \alpha + \log u$; SSE = error sum of squares.

Study	Full model		Slope model		Intercept model	
	SSE	$R^2\%$	*SSE%*	$R^2\%$	*SSE*	$R^2\%$
1	0.165	81.750	0.312	65.464	2.905	0.000
2	0.336	74.575	0.470	64.471	2.980	0.000
3	0.042	93.976	0.138	80.026	1.258	0.000
4	0.009	98.630	0.115	82.940	3.416	0.000
5	0.114	93.463	0.422	75.781	0.543	68.807
6	0.143	89.946	0.449	68.433	1.599	0.000
7	0.005	98.973	0.103	79.374	2.994	0.000
8	0.323	87.194	0.382	84.843	2.473	1.982
9	1.633	48.592	1.959	38.305	2.460	22.539
10	0.029	99.140	0.245	92.776	1.887	44.421

line by just one pair of observations as 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. However, it is not the only identifiable model. For example, the multiplicative model $p = \alpha^* u$ is identifiable as well as an intercept-only model $\log p = \alpha + \log u$. However, both of these models would give a perfect fit 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. One of these rare examples is the CAGE meta-analysis (Aertgeerts, Buntinx, & Kester, 2004) which we will use as a benchmark data set to investigate for the appropriateness of the approach. 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 4 questions. What makes this meta-analysis so unique is the fact that for each of the $k = 10$ studies included sensitivities and specificities are provided. The data are documented in Table 9. Here, a straight line model is identifiable which we consider as a benchmark model. We fitted three models for these data: the straight line model (usually not identifiable), the slope-only model (2) and the intercept-only model. We use standard measures of performance from regression analysis including $R^2 = 1 - \frac{SSE}{SSTOT} \times 100$ where *SSE* and *SSTOT* are the usual sum-of-squares from the ANOVA table. The results are presented in Table 6. Note that often the intercept-only model shows more variation in its fitted values than the observed variation leading to zero-percentage of explained variance. We find the Lehmann model (2) performs remarkably well in comparison to the full straight line model (again the latter not being identifiable in most cases). Clearly, the performance of the intercept-only model, although identifiable, is rather poor.

Appropriateness of the Normal Mixture Kernel The nonparametric mixture model (7) does not make any assumption on the mixing distribution of $\log \theta$. One can think of this mixing distribution as a random effects distribution of the random effect *diagnostic accuracy* as measured by $\log \theta$. In other words, we make no assumption on this random effect distribution; it is left unspecified. The random effects distribution might be normal or it might be something else. In any case, since it is an unobserved distribution it will be difficult to diagnose. However, leaving the random effects distribution unspecified will always guarantee a likelihood at least as large as any specified distribution (for further details see Skrondal & Rabe-Hesketh, 2004, or Böhning, 2000). In addition, it can be shown that the nonparametric maximum likelihood estimator

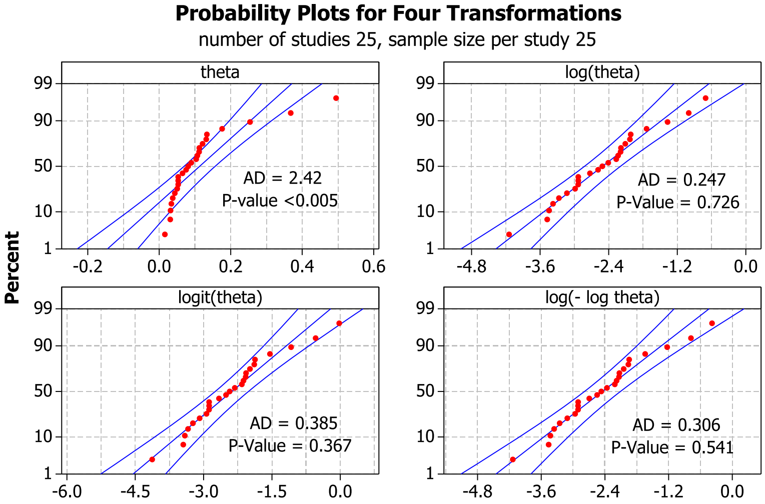

FIGURE 8.

Probability plots for a sample of estimates of the diagnostic accuracy parameters and three transformations; number of studies is 25 and the mean within-study sample size is 25; the plots also include the Anderson–Darling test (measuring goodness-of-fit) with associated p-values.

of this unspecified distribution is always discrete (Lindsay, 1995) and, as such, very much appropriate for a cluster-analytic interpretation. This means also that no other mixing or random effects distribution exists which would be able to provide a better fit. However, the nonparametric mixture model (7) has as another element: the mixture kernel. Whilst we would like to remain in the normal distributional framework, it is a matter of choice whether to work with the untransformed parameter θ itself or use the log-transformed value which we did in our analysis, or something else. Clearly, assuming within-study validity of the Lehmann model, 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 $\theta \in (0, 1)$. A simulation study was designed to mimic the reality of meta-analysis of diagnostic studies. The number k of studies was selected to be $k = 25, 50, 100$. A number of sample sizes where 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. A baseline heterogeneity was assumed for the false-positive rate in that u_i was sampled from a uniform distribution with interval ends 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), 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. Differences in the resulting distributions, in particular in their closeness to the normal distribution, did occur only for small within-study sample sizes, whereas the results were fairly stable when the number of studies was varied. In addition, for the larger sample sizes such as 50 and 100, all transformations given samples with a distribution close to normality, even the untransformed data values $\hat{\theta}_i$ were distributionally well behaved. Differences mainly occurred for the small sample sizes. Figure 8 shows the probability plots for $k = 25$ and mean within-study sample size 25. There is evidence that the untransformed sample is not normal since the Anderson–Darling test is highly significant, whereas all other transformations achieve transformed data closer to normality with best results for the log transformation.

6. Discussion

The approach is attractive since it is based on a simple measure of diagnostic accuracy per study—the ratio of log-sensitivity to log-false-positive rate. It was pointed out by one reviewer that the estimate of the ROC curve based on only one pair is very sensitive to the location of the point (p, u) in the ROC space, in particular, if this point falls into the vicinity of $(0, 0)$ or $(1, 1)$, the lower left corner and the upper right corner, respectively, of the ROC square. This is certainly a weakness of the approach. To diminish this sensitivity to extreme points, a smoothing constant of 0.5 has been used whenever the binomial event count was zero which helps, jointly with transforming proportions into the log space, to stabilise the proportion estimator as well as its variance (Sweeting, Sutton, & Lambert, 2004; Böhning & Viwatwongkasem, 2005).

Heterogeneity is of primary interest in general meta-analysis and typically approached by adjusting summary estimates in terms of their weights and increased variance by an estimate of the variance of a random effect distribution which is thought of modelling the heterogeneity present in the meta-analysis (Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Cooper & Hedges, 1994; Schulze, Holling, & Böhning, 2003; Sutton et al., 2000; Whitehead, 2002). An early contribution in this direction is the work by Dersimonian and Laird (1986). Other papers followed in this direction including Hardy and Thompson (1996, 1998), Biggerstaff and Tweedie (1997), Brockwell and Gordon (2001), and Böhning et al. (2002). However, approaches exploring the random effects distribution in more detail are less frequent. A likelihood-based cluster analysis has been suggested by Aitkin (1999a, 1999b), Böhning et al. (1998), and Kuhnert and Böhning (2007) for exploring the structure in a meta-analysis at hand. Although this approach also provides estimates of the overall mean as well as the heterogeneity variance using the mean and variance of the estimated mixing distribution, respectively, the approach is much more powerful in terms of providing a full modelling of the random effects distribution, including an allocation of studies into the various components (clusters).

Meta-analysis of diagnostic studies is a relatively young discipline with pioneering papers going back to Hasselblad and Hedges (1995) and Irwig, Tosteson, Gatsonis, Lau, Colditz, Chalmers, and Mosteller (1994, 1995). A state-of-the-art review on meta-analysis of diagnostic testing is provided by Gatsonis and Paliwal (2006) and Macaskill, Glasziou, and Irwig (2005), both pointing out the importance of the SROC concept for the area of interest. Random effects modelling in the SROC context has been approached by various authors including van Houwelingen, Zwinderman, and Stijnen (1993), Reitsma et al. (2005), Rutter and Gatsonis (2001) and Harbord, Deeks, Egger, Whiting, and Sterne (2007). However, approaches exploring heterogeneity in terms of its inherent cluster structure are not yet available. Hence, we feel that the present paper presents a contribution to fill this gap.

Appendix: Meta-analytic Data

TABLE 7.
Meta-analysis of diagnostic accuracy of the Alcohol Use Disorder Identification Test (AUDIT) and Alcohol Use Disorder Identification Test Consumption (AUDIT-C) for alcohol disorders.

Study	Alcohol disorder		No disorder		$n+m$
	y (TP)	$m-y$ (FN)	$n-x$ (TN)	x (FP)	
AUDIT					
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
AUDIT-C					
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

TABLE 8.
Meta-analysis of diagnostic accuracy of the MMSE of dementia.

Study	Condition		No condition	
	y (TP)	$m - y$ (FN)	x (FP)	$n - x$ (TN)
1	65	3	240	870
2	117	12	10	110
3	48	19	63	989
4	134	8	28	152
5	24	5	44	292
6	67	15	48	153
7	64	17	0	71
8	281	64	20	286
9	13	1	44	286
10	262	20	29	177
11	143	18	29	123
12	183	33	33	51
13	22	0	152	140
14	112	0	590	2091
15	152	81	126	1009
16	29	26	26	236
17	31	6	3	247
18	10	3	12	333
19	707	88	1438	10447
20	181	108	17	184
21	59	29	23	74
22	74	23	16	143
23	27	12	26	209
24	40	6	75	528
25	317	52	173	578
26	387	116	16	54
27	118	65	1	44
28	44	7	34	396
29	123	46	98	309
30	25	43	3	171
31	73	32	2	225
32	37	45	0	440
33	78	34	45	376

TABLE 9.
Meta-analysis of the diagnostic accuracy of the CAGE questionnaire for alcohol abuse and dependency.

Study	CAGE-score	Se[a]	Sp	Study	CAGE-score	Se	Sp
1–Saitz	1	0.92	0.73	2–McQuade	1	0.87	0.80
	2	0.80	0.93		2	0.66	0.92
	3	0.55	0.98		3	0.43	0.99
	4	0.27	0.99		4	0.19	0.99
3–Brown	1	0.79	0.77	4–Chan	1	0.96	0.68
	2	0.70	0.85		2	0.87	0.84
	3	0.52	0.95		3	0.56	0.96
	4	0.27	0.98		4	0.34	0.99
5–Aergeerts	1	0.61	0.87	6–Buchsbaum	1	0.89	0.81
	2	0.46	0.95		2	0.73	0.91
	3	0.24	0.98		3	0.44	0.98
	4	0.11	0.99		4	0.19	0.99
7–Joseph	1	0.98	0.75	8–Bradley	1	0.71	0.59
	2	0.82	0.90		2	0.53	0.87
	3	0.53	0.97		3	0.27	0.98
	4	0.40	0.99		4	0.09	0.99
9–Jones	1	0.88	0.88	10–Indran	1	0.99	0.37
	2	0.48	0.99		2	0.92	0.62
	3	0.24	0.99		3	0.46	0.88
	4	0.08	0.99		4	0.10	0.99

[a] Se = Sensitivity, Sp = Specificity.

References

Aergeerts, B., Buntinx, F., & Kester, A. (2004). 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*, 30–39.

Aitkin, M. (1999a). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 117–128.

Aitkin, M. (1999b). Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, *18*, 2343–2351.

Biggerstaff, B.J., & Tweedie, R.L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, *16*, 753–768.

Böhning, D., Dietz, E., & Schlattmann, P. (1998). Recent developments in computer-assisted analysis of mixtures (C.A.MAN). *Biometrics*, *54*, 367–377.

Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others*. Boca Raton: Chapman & Hall/CRC.

Böhning, D., Malzahn, U., Dietz, E., Schlattmann, P., Viwatwongkasem, C., & Biggeri, A. (2002). Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. *Biostatistics*, *3*, 445–457.

Böhning, D. (2003). The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Statistics and Computing*, *13*, 257–265.

Böhning, D., & Viwatwongkasem, C. (2005). Revisiting proportion estimators. *Statistical Methods in Medical Research*, *14*, 1–23.

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*, 543–554.

Brockwell, S.E., & Gordon, I.R. (2001). A comparison of statistical methods for meta-analysis. *Statistical Methods in Medical Research*, *20*, 825–840.

Clayton, D.G., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disese mapping. *Biometrics*, *43*, 671–681.

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

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B*, *39*, 1–38.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

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

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

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.

Hardy, R.J., & Thompson, S.G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, *15*, 619–629.

Hardy, R.J., & Thompson, S.G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*, 841–856.

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

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hunter, J.E., Schmidt, F.L., & Jackson, B.G. (1982). *Meta-analysis: Cumulative research findings across studies*. Beverly Hills: Sage Publications.

Irwig, L., Tosteson, A.N., Gatsonis, C., Lau, J., Colditz, G., Chalmers, T.C., & Mosteller, F. (1994). Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal Medicine*, *120*, 667–676.

Irwig, L., Macaskill, P., Glasziou, P., & Fahey, M. (1995). Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology*, *48*, 119–130.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, *27*, 886–906.

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

Kuhnert, R., & Böhning, D. (2007). A comparison of three different models for estimating relative risk in meta-analysis. *Statistics in Medicine*, *28*, 2277–2296.

Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*, 805–811.

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.

Lindsay, B.G. (1983). The geometry of mixture likelihoods: A general theory. *Annals of Statistics*, *11*, 86–94.

Lindsay, B.G. (1995). *Mixture models: Theory, geometry, and applications*. Hayward: Institute of Statistical Mathematics.

Macaskill, P., Glasziou, P., & Irwig, L. (2005). Meta-analysis of diagnostic tests. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics*. doi:10.1002/0470011815.b2a04026.

Martin, C.S., Liepman, M.R., & Young, C.M. (1990). The Michigan alcoholism screening test: False positives in a college student sample. *Alcoholism, Clinical and Experimental Research*, *14*, 853–855.

McCullagh, P., & Nelder, J.A. (1989). *Generalized linear models*. London: Chapman & Hall.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Midgette, A.S., Stukel, T.A., & Littenberg, B. (1993). A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic-summary point estimates. *Medical Decision Making*, *13*, 253–257.

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.

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

Ray, S., & Lindsay, B.G. (2008). Model selection in high dimensions: A quadratic-risk-based approach. *Journal of the Royal Statistical Society. Series B*, *70*, 95–118.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, *3*, 215–232.

Reinert, D.F., & Allen, J.P. (2002). The alcohol use identification test (AUDIT): A review of recent research. *Alcoholism, Clinical and Experimental Research*, *26*, 272–279.

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 measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*, 982–990.

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

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.

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

Selzer, M.L. (1971). The Michigan alcoholism screening test: The quest for a new diagnostic instrument. *American Journal of Psychiatry*, *127*, 1653–1658.

Sidik, K., & Jonkman, J.N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society. Series C*, *54*, 367–384.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Storgaard, H., Nielsen, S.D., & Gluud, C. (1994). The validity of the Michigan alcoholism screening test (MAST). *Alcohol and Alcoholism*, *29*, 493–502.

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

Sweeting, M.J., Sutton, A.J., & Lambert, P.C. (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, *23*, 1351–1375.

Swets, J.A. (2009). *Signal detection theory and ROC analysis in psychology and diagnostics*. New York: Psychology Press.

van Houwelingen, H.C., Zwinderman, K.H., & Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine*, *12*, 2273–2284.

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials*. New York: Wiley.

Manuscript Received: 12 NOV 2010
Final Version Received: 6 JUN 2011
Published Online Date: 12 NOV 2011