Bivariate Random Effects Meta-Analysis of ROC Curves

L.R. Arends, T.H. Hamza, J.C. van Houwelingen, M.H. Heijenbrok-Kal, M.G.M. Hunink and T. Stijnen

Med Decis Making 2008 28: 621 originally published online 30 June 2008
DOI: 10.1177/0272989X08319957

Bivariate Random Effects Meta-Analysis of ROC Curves

L. R. Arends, PhD, T. H. Hamza, PhD, J. C. van Houwelingen, PhD, M. H. Heijenbrok-Kal, PhD, M. G. M. Hunink, PhD, T. Stijnen, PhD

Meta-analysis of receiver operating characteristic (ROC)-curve data is often done with fixed-effects models, which suffer many shortcomings. Some random-effects models have been proposed to execute a meta-analysis of ROC-curve data, but these models are not often used in practice. Straightforward modeling techniques for multivariate random-effects meta-analysis of ROC-curve data are needed. The 1st aim of this article is to present a practical method that addresses the drawbacks of the fixed-effects summary ROC (SROC) method of Littenberg and Moses. Sensitivities and specificities are analyzed simultaneously using a bivariate random-effects model. The 2nd aim is to show that other SROC curves can also be derived from the bivariate model through different characterizations of the estimated bivariate normal distribution. Thereby the authors show that the bivariate random-effects approach not only extends the SROC approach but also provides a unifying framework for other approaches. The authors bring the statistical meta-analysis of ROC-curve data back into a framework of relatively standard multivariate meta-analysis with random effects. The analyses were carried out using the software package SAS (Proc NLMIXED). **Key words:** meta-analysis; diagnostic tests; multivariate random effects models; sensitivity; specificity; receiver operating characteristic (ROC) analysis.***(Med Decis Making 2008;28:621–638)***

For a thorough assessment of the effectiveness of a specific treatment, it is common to execute a meta-analysis of randomized clinical trials reported in the literature. The same is done for the assessment of the characteristics of a diagnostic test to distinguish patients having a certain disease from patients not having that disease. Meta-analyses to assess the reliability, accuracy, and impact of diagnostic tests are essential to guide optimal test selection and the appropriate interpretation of test results.[1] However, the designs of test accuracy evaluations differ from the designs of studies that evaluate the effectiveness of treatments, which means that different criteria are needed when assessing study quality and potential for bias. Additionally, often each evaluation of diagnostic tests reports a pair of related summary statistics (for example, sensitivity and specificity) rather than a single statistic, requiring alternative statistical methods for pooling study results.[1] Receiver operating characteristic (ROC) curves are used in studies of diagnostic accuracy to depict the pattern of sensitivities and specificities observed when the performance of the test is evaluated at several different diagnostic thresholds.

In the past 15 years, several methods for meta-analysis of diagnostic tests have been developed.[2–11] The proposed methods depend on the type of data available. Some[6,8] are designed to be used with individual patient data of the studies. Others are applicable when each study provides an estimate of the area under the ROC curve.[12] Still others are applicable to the situation where per study only one estimated pair of sensitivity and specificity (corresponding to possibly different diagnostic thresholds) is available. In this article, we focus on this last situation, which is by far the most common in practice. For this situation, the

Received 10 September 2005. Originally submitted 1 December 2005. Version 3 accepted 7 September 2007. From the Department of Epidemiology & Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands (LRA, THH, MHH-K, MGMH, ThS); Institute of Psychology, Erasmus University Rotterdam, Rotterdam, The Netherlands (LRA); Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands (JCvH); Department of Radiology, Erasmus Medical Center, Rotterdam, Netherlands (MHH-K, MGMH); and Department of Health Policy & Management, Harvard School of Public Health, Boston, Massachusetts (MGMH). Revision accepted for publication 7 Sept 2007.

Address correspondence to L. R. Arends, PhD, Department of Epidemiology & Biostatistics, Erasmus Medical Center, PO Box 1738, 3000 DR Rotterdam, The Netherlands; e-mail: l.arends@erasmusmc.nl.

DOI: 10.1177/0272989X08319957

aim of the meta-analysis is to estimate the overall ROC curve of the (continuous) diagnostic marker.

Probably the most well known and most commonly used method in practice is the summary ROC (SROC) method proposed by Littenberg and Moses[2] and Moses and others.[3] They plotted the difference versus the sum of the logit(true positive rate) and logit(false positive rate) from each study. Then they fitted 3 types of regression lines (robust, unweighted, and weighted) to these points. Finally, they transformed the line to ROC space.

Although frequently used, the SROC method has a number of serious shortcomings. The 1st aim of this article is to present an approach that extends the SROC method, addresses its drawbacks, and is still easily carried out in practice using familiar statistical packages like SAS. The method follows the general multivariate approach as described in van Houwelingen and others[13] and Arends and others.[14,15] The 2nd aim of this article is to show how other summary ROC curves (regression lines in the logit space) can be derived from the bivariate model, such as the logit(true positive rate) on logit(false positive rate), logit(false positive rate) on logit(true positive rate), the 1st principal component, or the curve corresponding to the method of Rutter and Gatsonis.[9]

Below, we introduce 2 data sets that are used as examples. Next, we give an overview of the SROC method and briefly discuss its shortcomings, after which we briefly discuss other methods proposed in the literature. The new approach is then presented. In the following section, the methods are applied on the 2 example data sets and the results are presented. We use the SAS procedures Proc Mixed and Proc NLMixed for the analyses and give the syntax in the appendix. We end the article with a discussion.

DATA EXAMPLES

To illustrate the methods discussed in this article, we apply them to 2 meta-analysis data sets, one relatively small (29 studies) and one large (149 studies).

Example 1: FNAC of the Breast[16]

Giard and Hermans[16] present 29 studies evaluating the accuracy of fine-needle aspiration cytologic examination (FNAC) of the breast to assess presence or absence of breast cancer. FNAC provides a nonoperative way of obtaining cells for establishment of the nature of a breast lump and therefore plays a pivotal role in the preoperative diagnostic process.[16–19] The sensitivity and specificity of FNAC were determined for each study. Sensitivity was defined as the probability of a malignant or suspect test result in patients with cancer. Specificity was defined as the probability of absence of abnormal cells in the patients without cancer.[16] Table 1 shows the frequencies of the FNAC outcomes given the final diagnosis of benign or malignant breast disease.

In Table 1, Y_1 is the number of patients with a malignant or suspect test result in the patients with cancer. The total number of patients with cancer is n_1. Y_0 is the number of patients with a malignant or suspect test result in the n_0 patients without cancer. The true positive rate *TPR*, or sensitivity, is estimated for a study by Y_1/n_1, and the false positive rate *FPR*, which is 1 minus the specificity, by Y_0/n_0. See Figure 1a for a plot of the estimated *TPR*s against the estimated *FPR*s and Figure 1c for the estimated *TPR*s and *FPR*s on the logit scale.

The estimated *TPR*s and *FPR*s vary considerably across studies. Also, the proportions of patients with benign or malignant disease according to the final diagnosis differed substantially. At the time of publication (1992), no reasonable methods to summarize diagnostic test data across several studies were available. In this article, we use the data to fit the standard fixed-effects SROC model as well as the proposed random-effects models.

Example 2: Imaging Tests for Coronary Artery Disease[20]

Heijenbrok-Kal[20] searched PubMed from January 1990 through May 2003 for meta-analytic studies on the diagnostic performance of imaging tests for coronary artery disease. In all meta-analyses included in her report, angiography is the reference standard, and the source numbers of true and false positives and true and false negatives are reported. Duplicate source studies are excluded. This resulted in 246 patient series that included 24,761 patients who underwent 1 of 8 different imaging technologies for coronary artery disease. The coronary tests showed little difference in diagnostic performance.

To illustrate our approach, we choose from the 246 source studies only those in which the performance of an exercise or stress echo was investigated. This resulted in 149 studies that included 13,303 patients. In Figure 1b, a plot is given of the estimated *TPR*s against the estimated *FPR*s, and Figure 1d represents the estimated *TPR*s and *FPR*s on the logit scale.

Table 1 Example 1: Data from Clinical Studies on Patients with a Breast Mass Who Underwent a Fine-Needle Aspiration Cytological Examination (FNAC)

	FNAC Results for Patients with Benign Disease			FNAC Results for Patients with Malignant Disease		
Study	False Pos (Y_0)	True Neg	Total (n_0)	True Pos (Y_1)	False Neg	Total (n_1)
1	70	939	1009	979	89	1068
2	3	163	166	51	22	73
3	55	894	949	1569	152	1721
4	25	259	284	35	15	50
5	4	121	125	59	12	71
6	18	216	234	56	4	60
7	602	3117	3719	329	39	368
8	10	213	223	125	17	142
9	88	499	587	211	63	274
10	0	31	31	49	1	50
11	26	643	669	336	178	514
12	147	746	893	210	42	252
13	5	25	30	16	3	19
14	16	356	372	258	53	311
15	9	107	116	56	18	74
16	16	112	128	162	28	190
17	6	112	118	116	13	129
18	99	145	244	65	12	77
19	5	78	83	94	10	104
20	0	70	70	26	4	30
21	28	136	164	1318	249	1567
22	55	539	594	569	120	689
23	1	287	288	46	16	62
24	13	76	89	64	6	70
25	1	104	105	39	4	43
26	16	426	442	132	20	152
27	17	161	178	470	22	492
28	25	200	225	28	4	32
29	43	22	65	42	3	45

Note: Patients are cross-classified according to their final diagnosis (benign or malignant breast disease) and their FNAC result.

THE STANDARD SROC METHOD

The starting point of a meta-analysis of ROC-curve data is a number of studies providing information on a continuous diagnostic marker or variable M. In the different studies, possibly different thresholds for M are used to obtain a dichotomous diagnostic test. The data provided by each study are the number of patients with a positive test result (y_1) and the total number of patients (n_1) in the group with the disease, and the number of patients with a positive test result (y_0) and the total number of patients (n_0) in the group without the disease. The aim is to estimate the overall ROC curve of the diagnostic marker M based on the available data from the different studies. The standard method used in practice is the SROC method of Littenberg and Moses,[2] which proceeds as follows. The underlying model assumes that there exists a transformation of the continuous diagnostic variable M such that the transformed test, X, follows a logistic distribution both in the population without the disease and in the population with the disease. In other words, it is assumed that the transformation that makes the distribution of M logistic in the nondiseased (which always exists) makes the distribution simultaneously logistic for those with the disease. We assume that the transformation is done such that large values of X correspond with the diseased population.

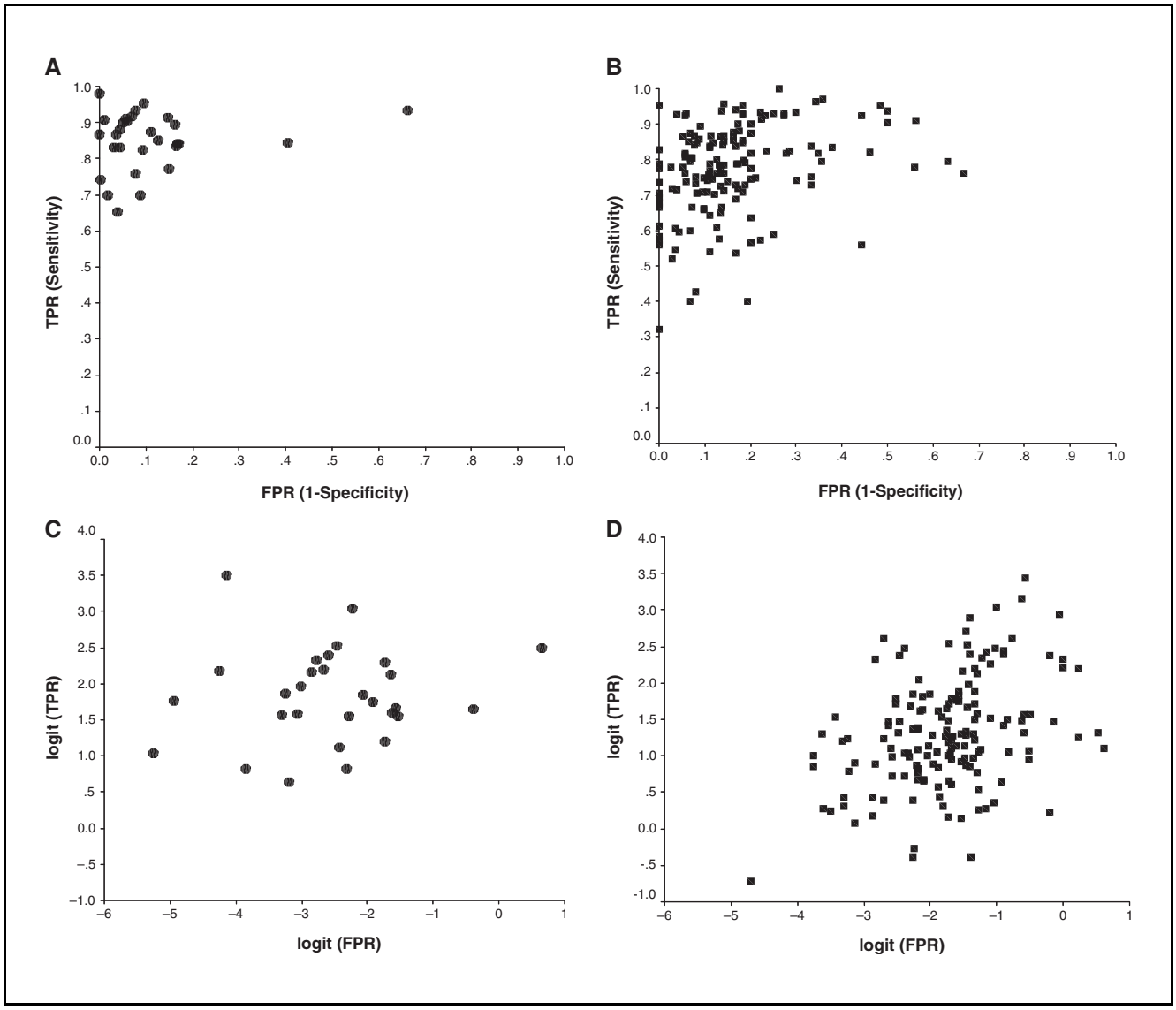

Figure 1 Observed sensitivity against (1-specificity) of data reported across 29 studies that were originally meta-analyzed by Giard and Hermans[16] *(left side of picture) and across 149 studies that were originally meta-analyzed by Heijenbrok-Kal and others*[20] *(right side of picture) on the original scale and on logit transformed scale. TPR = true positive rate; FPR = false positive rate.*

The cumulative distribution of X in the healthy and the diseased populations is given by

$$\Pr(X < x|healthy) = \frac{e^x}{1+e^x} \text{ and}$$
$$\Pr(X < x|disease) = \frac{e^{-\alpha+\beta x}}{1+e^{-\alpha+\beta x}}, \quad (1)$$

for some values of $\alpha \geq 0$ and $\beta > 0$. The difference between the mean value in the population with the disease and without the disease is α/β, and the ratio between the standard deviation of the diseased and the healthy population is $1/\beta$. Thus, $0 < \beta < 1$ corresponds with a higher variance in the population with the disease and $\beta > 1$ with a smaller variance. Figure 2 gives a graphical illustration with the interpretation of α and β, where clearly $0 < \beta < 1$.

If λ denotes the threshold X-value for the test being declared positive, then according to (1) the probability of a false positive result is $1 - e^{\lambda}/(1+e^{\lambda})$ and hence $\text{logit}(FPR) = -\lambda$. Similarly we have $\text{logit}(TPR) = \alpha - \beta\lambda$. In the following, we will use the notation:

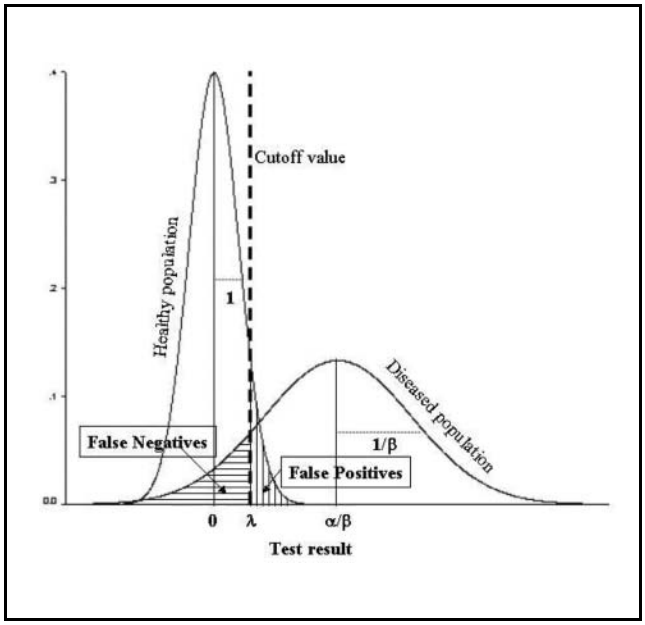

Figure 2 Graphical illustration with interpretation of α and β.

$$\xi = logit(FPR) = -\lambda$$
$$\eta = logit(TPR) = \alpha - \beta\lambda.$$

This implies the linear relationship

$$\eta = \alpha + \beta\xi. \quad (2)$$

Following Rutter and Gatsonis,[9] α can be called the accuracy parameter and β the scale or asymmetry parameter. If $\beta = 1$, the resulting ROC curve is symmetric (with respect to the minus 45° diagonal), otherwise it is asymmetric.

In the SROC approach of Littenberg and Moses, the relation (2) is written as

$$\eta - \xi = \alpha' + \beta'(\eta + \xi),$$

with $\alpha' = 2\alpha/(\beta + 1)$ (with $\alpha' \geq 0$) and $\beta' = (\beta - 1)/(\beta + 1)$ (with $-1 < \beta' < 1$). If D and S are the estimated values of $\eta - \xi$ and $\eta + \xi$ from a study (to avoid division by zero, 0.5 is added to all numbers in the 2×2 table of a study), then approximately

$$D = \alpha' + \beta' S, \quad (3)$$

and the values of α' and β' are estimated by a simple weighted or unweighted linear regression. The weights are chosen proportional to the inverse variance of D. D is interpreted as the log odds ratio of a positive test result for diseased individuals relative to healthy individuals, and is often called the diagnostic odds ratio. Its estimated variance is

$$\frac{1}{y_0 + 0.5} + \frac{1}{n_0 - y_0 + 0.5} + \frac{1}{y_1 + 0.5} + \frac{1}{n_1 - y_1 + 0.5}. \quad (4)$$

The summary ROC curve is obtained by transforming the estimate of (3) to the ROC space. A value of $\beta' \neq 0$ indicates that the curve is asymmetric.

The advantage of the SROC method, which explains its popularity, is that it is very simple to understand and can be carried out in any statistical package. Despite this important advantage of simplicity, a number of critical comments must be made.

First, the SROC method is a fixed effects method; that is, it assumes that the values of α and β do not vary across studies. Thus, variation is due only to the threshold effect and within-study sampling variability. However, in many practical cases it is likely that there is between-study variation beyond those sources. Study characteristics such as technical aspects of the diagnostic test, patient selection, study settings, and experience of readers are among the potential contributors to between-study variation in the estimates of diagnostic performance.[9] Modern meta-analytical methods take possible variation across studies into account by introducing random effects.[13,21–24] If there is between-study variation, a fixed effects model can give biased estimates and typically underestimates standard errors.[21,25]

Second, the independent variable S in the regression equation (3) is measured with measurement error, which should be taken into account. As a result, regression to the mean[26] and attenuation due to measurement errors[27] could seriously bias the slope of the regression line.[13] Thus, not taking into account the measurement error in S leads to bias in β' (in general, toward zero) and α' and therefore also in β (in general, toward one) and α.[24]

Third, D and S are correlated within a study, positively or negatively, depending on the study. In the standard fixed-effects SROC model, this correlation is ignored. Although probably the correlation is usually small in practice, the consequence of ignoring it is not obvious.

Fourth, it is reasonable that the different studies should be somehow weighed in the analysis, in particular, if the studies vary substantially in size. If there is both between- and within-study variation, then weighing by the inverse of within-study variances, as is done in the weighted SROC approach, will not be optimal.

Finally, to avoid undefined log odds, log odds ratios, and their variances, quite arbitrarily 0.5 is added to the numbers in the 4-fold tables of the trials. As Moses and others[3] showed, the effect of this adjustment can be surprisingly large. Adding 0.5 to all cells tends to push an estimated ROC curve away from the desirable northwest corner of ROC space. The standard SROC method has to do this because it does not use the true binomial distribution of the number of positive test results within a group. It would be preferable to not use this artificial and arbitrary correction.

In a later section, we present a method that does not have the described disadvantages of the SROC method and can still be carried out easily in standard statistical packages. In addition, we show how other summary ROC curves (regression lines in the logit space) can be derived from our method. Before doing so, we discuss in the next section some other methods proposed in the literature.

OTHER METHODS PROPOSED IN THE LITERATURE

Kardaun and Kardaun[28] also assume model (1) and exploit the linear relationship $\eta_i = \alpha + \beta\xi_i$ where $i = 1, \ldots, k$ denotes the number of the study. Using straightforward approximate likelihood methods, all $k+2$ parameters (including the ξ_i's) are estimated. The estimation method is called *approximate* likelihood, because, instead of the exact likelihood based on the true conditional distribution of the estimated ξ (called $\hat{\xi}$) and η (called $\hat{\eta}$) given ξ and η, an approximate likelihood based on the familiar normal approximations $\hat{\xi} \cong N(\xi, 1/y_0 + 1/(n_0 - y_0))$ and $\hat{\eta} \cong N(\eta, 1/y_1 + 1/(n_1 - y_1))$ is used. The drawbacks of the method of Kardaun and Kardaun[28] are, first, that the number of estimated parameters is proportional to the number of trials; hence standard likelihood theory does not apply.[29] For instance, consistency of the estimates when the number of studies tends to infinity is not guaranteed. Second, their computer-intensive method based on profile likelihood is not very practical. Third, it is a fixed-effects model, and fourth, the arbitrary value of 0.5 is added to avoid undefined log odds and their variances; thus, the first and last drawbacks mentioned in the previous section for the SROC method still apply.

Recently, Rutter and Gatsonis[9] proposed a hierarchical Bayesian regression approach that does not have the drawbacks mentioned in the previous section for the SROC method. They assumed the following model. Let π_{i0} be the true *FPR* in the nondiseased and π_{i1} be the true *TPR* in the diseased populations. Then $Y_0 \sim$ Binomial(n_0, π_{i0}) and $Y_1 \sim$ Binomial(n_1, π_{i1}). Defining $\xi_i = \text{logit}(\pi_{i0})$ and $\eta_i = \text{logit}(\pi_{i1})$, the following relationship is assumed to hold between ξ_i and η_i:

$$\begin{aligned} \xi_i &= (\theta_i + \alpha_i X_0) e^{-\beta X_0} \\ \eta_i &= (\theta_i + \alpha_i X_1) e^{-\beta X_1}, \end{aligned} \tag{5}$$

where X_0 and X_1 are chosen to be $-1/2$ and $+1/2$, respectively. This implies the linear relationship

$$\eta_i = \alpha_i e^{-\beta/2} + e^{-\beta}\xi_i. \tag{6}$$

For equation (6), α is called the accuracy parameter, because it measures the difference between *TPR* and *FPR*, and β is called the scale parameter. With this parameterization, if $\beta \neq 0$, the ROC curve is asymmetric.

The between-study variation is modeled by assuming that α_i and θ_i are independent and normally distributed:

$$\begin{pmatrix} \alpha_i \\ \theta_i \end{pmatrix} = N\left(\begin{pmatrix} \bar{\alpha} \\ \bar{\theta} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\theta^2 \end{pmatrix} \right). \tag{7}$$

To compute a summary ROC curve, Rutter and Gatsonis[9] plug in the estimates for $\bar{\alpha}$ and β into the linear relation (6) and transform it into the ROC space.

The method allows for between-study variation by modeling the accuracy parameter α with a random effect. A practical disadvantage is that Rutter and Gatsonis[9] compute the estimates in a Bayesian way using Markov Chain Monte Carlo (MCMC) simulation with the BUGS software, which is rather complicated. MCMC estimation requires programming, simulation, evaluation of convergence and model adequacy, and synthesis of simulation results. Implementation of MCMC simulation entails nontrivial analysis tasks including evaluation of convergence and the adequacy of prior distributions, and these tasks require statistical expertise. As the authors mention, this is a high price that has to be paid for the advantages of the hierarchical SROC model. Furthermore, Rutter and Gatsonis[9] use a relatively complicated parameterization, which can make it difficult for the meta-analyst to fully understand what she or he is doing. Macaskill[30] shows how the model of Rutter and Gatsonis can be fitted in a non-Bayesian way using the SAS NLMixed program for generalized linear mixed models. This

makes the model of Rutter and Gatsonis much more practical.

Recently a straightforward random effects extension of the method of Littenberg and Moses[2] has been used in some medical applications,[31–33] using the STATA program Metareg.[34] This method is as follows. Let η_i and ξ_i again denote the true logit(*TPR*) and logit(*FPR*) for study *i*. Let $D_i = \eta_i - \xi_i$ be the true log odds ratio and $S_i = \eta_i + \xi_i$. The corresponding estimates are given by $\hat{\xi}_i$, $\hat{\eta}_i$, $\hat{D}_i$, and $\hat{S}_i$, respectively. Then the model is:

$$\hat{D}_i = \alpha_i + \beta \hat{S}_i + \varepsilon_i,$$

with

$$\varepsilon_i \cong N\left(0, \frac{1}{y_0 + 0.5} + \frac{1}{n_0 - y_0 + 0.5} + \frac{1}{y_1 + 0.5} + \frac{1}{n_1 - y_1 + 0.5}\right) \text{ and } \alpha_i \cong N(\bar{\alpha}, \sigma_\alpha^2). \quad (8)$$

In this model, all studies have a common slope β, but the intercepts vary randomly between studies according to a normal distribution. The overall ROC line is $\eta = \bar{\alpha} + \beta\xi$, where the individual study lines vary randomly around this line with between-study standard deviation σ_α. This is the standard random effects meta-regression model, and there are many statistical packages in which this model can be fitted, such as SAS, STATA, and R/S-Plus. Measurement error of $\hat{D}_i$ is correctly accounted for, but the measurement error in $\hat{S}_i$ is still neglected. Another drawback for sparse data sets is that it is not possible to use the underlying binomial distributions for $\hat{D}_i$ and $\hat{S}_i$ instead of the normal approximations.

ALTERNATIVE APPROACH

In numerous medical articles, sensitivities or specificities are meta-analyzed separately by the standard random-effects model of DerSimonian and Laird.[21] The method we propose is a direct extension of this approach. We analyze sensitivities and specificities simultaneously using a 2-dimensional random-effects model. We will show that the model implies a linear relationship between η and ξ, and can be seen as an extension of the SROC method of Littenberg and Moses.[2] In the next section, we introduce our model, after which we discuss several types of summary ROC curves and then discuss the relation with the approach of Rutter and Gatsonis.[9] Throughout, we follow a 2-level hierarchical modeling approach, explicitly modeling the within- and between-study variability.

The Bivariate Model

The standard way of meta-analyzing false-positive rates of a diagnostic test in the medical literature is the DerSimonian and Laird[21] random-effects model:

$$\xi_i \cong N(\bar{\xi}, \sigma_\xi^2) \text{ with } \hat{\xi}_i | \xi_i \cong N\left(\xi_i, \frac{1}{x_0} + \frac{1}{n_0 - x_0}\right),$$

where $\hat{\xi}_i$ and ξ_i are the observed and true logit(*FPR*) of study *i*, respectively. Note the well-known formula for the standard error of an estimated log odds. The parameter $\bar{\xi}$ describes the overall mean logit false-positive rate and σ_ξ^2 describes the between-study variance in true logit false-positive rates. Similarly, true positive rates are analyzed using the model:

$$\eta_i \simeq N(\bar{\eta}, \sigma_\eta^2) \text{ with } \hat{\eta}_i | \eta_i \cong N\left(\eta_i, \frac{1}{x_1} + \frac{1}{n_1 - x_1}\right).$$

The straightforward generalization is to assume a bivariate normal model for the pair (ξ_i, η_i):

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \cong N\left(\begin{pmatrix} \bar{\xi} \\ \bar{\eta} \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \sigma_{\xi\eta} \\ \sigma_{\xi\eta} & \sigma_\eta^2 \end{pmatrix}\right). \quad (9)$$

Note that this model implies the standard univariate random-effects meta-analysis model for the ξ_i and η_i separately but now allows that ξ_i and η_i are correlated.

This model fits in the framework of bivariate meta-analysis as originally introduced by van Houwelingen and others.[35] Later on, McIntosh[36] and Arends and others[14] used this model to investigate the relationship between baseline risk and size of treatment effect in clinical trials meta-analysis. In van Houwelingen and others,[13] bivariate meta-analysis was generalized to multivariate meta-analysis, and it was shown how standard general linear mixed-model programs can be used to fit these models. An example of a trivariate meta-analysis is given by Arends and others.[15]

The most simple characterization of the overall accuracy of the diagnostic test would be to take the estimated $\bar{\xi}$ and $\bar{\eta}$ and transform them to the ROC space. A more extensive description would be to characterize the bivariate normal distribution by a line and transform that line to the ROC space. Note that the bivariate normal distribution implies a linear association between ξ_i and η_i. However, as is

discussed in the next section, different lines might be employed, leading to different summary ROC curves. For example, the regression line of η_i on ξ_i could be used. Standard normal distribution theory tells that the regression line of η_i on ξ_i has intercept α and slope β given by

$$\alpha = \bar{\eta} - \frac{\sigma_{\xi\eta}}{\sigma_\xi^2}\bar{\xi} \quad \text{and} \quad \beta = \frac{\sigma_{\xi\eta}}{\sigma_\xi^2}. \tag{10}$$

The residual variance of the regression, given by

$$\sigma_{\eta|\xi}^2 = \sigma_\eta^2 - \frac{\sigma_{\xi\eta}^2}{\sigma_\xi^2},$$

describes the variation in the true sensitivities between studies that have the same specificity. In the following section, we discuss some alternative summary ROC curves.

Similarly, as in the above univariate models for meta-analyzing specificities and sensitivities separately, we model the within-study sampling variability using the fact that the estimated logit transformed *FPR*, $\hat{\xi}_i$, and *TPR*, $\hat{\eta}_i$, are independent and approximately normally distributed:

$$\hat{\xi}_i|\xi_i \cong N\left(\xi_i, \frac{1}{x_0} + \frac{1}{n_0 - x_0}\right) \text{ and}$$
$$\hat{\eta}_i|\eta_i \cong N\left(\eta_i, \frac{1}{x_1} + \frac{1}{n_1 - x_1}\right). \tag{11}$$

If one or more of the denominators are close to zero, 0.5 should be added to the denominators, as in (4). The equations (9) and (11) together specify a general linear mixed model (GLMM), and the parameters can be estimated by (restricted) maximum likelihood using a GLMM program. Subsequently, the intercept α and the slope β of a summary line can be calculated, using, for instance, (10) or one of the formulas given in the next subsection if another type of summary ROC curve is preferred. Standard errors of α and β can be calculated with the delta method. Many statistical packages provide a GLMM program. We used Proc Mixed from the SAS package. The syntax is given in the appendix. Proc Mixed does not give estimates and standard errors of user-defined derived parameters, so we had to calculate the estimates of α and β by hand, though the calculations are very simple. SAS users can avoid these hand calculations, because the model can also be fitted in Proc NLMixed. This program provides estimates and standard errors of user-defined derived parameters. The syntax needed for Proc NLMixed is given in the appendix. Another possibility in Proc NLMixed is to reparameterize the model in such a way that one immediately gets the estimates and standard errors for the parameters of interest.

We call the GLMM approach the *approximate* likelihood approach, because an approximate (normal) model denoted by equation (11) is used for the within-study sampling variability. The practical advantage is that the model remains a GLMM, for which much software is available. The approximate likelihood approach works well for larger data sets.[13] As a rule of thumb, the requirement "all denominators in equation (11) larger than or equal to 5" might be adopted, though this is probably too severe.

In our model, the first drawback of the SROC method as mentioned in an earlier section, that it is a fixed-effects model, no longer applies. Also the model does not suffer from drawbacks 2, 3, and 4. The problem of measurement error (drawback 2) is avoided by assuming a distribution for ξ_i. In general, there are 2 ways of dealing with measurement error, the structural approach and the functional approach.[13] Our approach is in the spirit of the structural approach, similar to Arends and others[14,15] and van Houwelingen and others,[13] which has the important advantage that the parameters can be estimated by straightforward likelihood methods.

Drawback 3 does not apply, because $\hat{\xi}_i$ and $\hat{\eta}_i$ are independent within studies. Even if we would formulate the model in terms of *D* and *S*, as is done in the standard SROC method, then there would be no problem because the correlation can be easily modeled in the GLMM.

Drawback 4 does not apply, because the likelihood method implicitly uses the "correct" weighting based on within- as well as between-study variation. The fifth drawback, that is, arbitrarily adding 0.5 to the numbers in the 4-fold table to avoid undefined log odds, still applies because we assumed an approximate within-study model. If we want to address this drawback as well, the true distribution of $\hat{\xi}_i = Y_{0i}/n_{0i}$ and $\hat{\eta}_i = Y_{1i}/n_{1i}$ should be used. Given the true $FPR_i = (1 + \exp(-\xi_i))^{-1}$ and $TPR_i = (1 + \exp(-\eta_i))^{-1}$ of study *i*, the observed test positive numbers Y_{0i} in the healthy group and Y_{1i} in the diseased group follow binomial distributions:

$$Y_{0i}|n_{0i}, FPR_i \cong \text{Binomial}(n_{0i}, FPR_i);$$
$$Y_{1i}|n_{1i}, TPR_i \cong \text{Binomial}(n_{1i}, TPR_i). \tag{12}$$

The equations (9) and (12) together now specify a general*ized* linear mixed model. This model has the advantage that drawback 5 no longer applies, but a

practical disadvantage is that software for G*ized*LMMs is not available in many packages. We again used Proc NLMixed of SAS. A syntax example is given in the appendix. We call this the *exact* likelihood approach, because the likelihood is based on the exact (i.e., binomial) within-study distribution of the data.

Choice of Summary ROC Curve

Above we have seen that a summary ROC curve can be obtained through a characterization of the estimated bivariate normal distribution given by (9). One possibility is to take the regression line of η_i on ξ_i, as we did above. However, there are other possibilities as well. For example, we could take the regression line of ξ_i on η_i. We now discuss this and other possible choices.

The Regression Line of η_i on ξ_i

$$\eta = \bar{\eta} + \frac{\sigma_{\xi\eta}}{\sigma_\xi^2}(\xi - \bar{\xi}). \tag{13}$$

This summary line estimates the mean logit transformed sensitivity given a specific value for the logit transformed 1-specificity. When transformed to the ROC space, the summary ROC curve estimates the median *TPR* given a specific value for the *FPR*.

The Regression Line of ξ_i on η_i

$$\xi = \bar{\xi} + \frac{\sigma_{\xi\eta}}{\sigma_\eta^2}(\eta - \bar{\eta}),$$

which is equivalent to:

$$\eta = \bar{\eta} + \frac{\sigma_\eta^2}{\sigma_{\xi\eta}}(\xi - \bar{\xi}). \tag{14}$$

There is no a priori reason to regress η on ξ instead of the other way around. Therefore, an alternative summary line is obtained by regressing ξ on η.

This summary line characterizes the mean logit transformed 1-specificity given a specific value for the logit transformed sensitivity. When transformed to the ROC space, the summary ROC curve characterizes the median *FPR* given a specific value for the *TPR*.

The Regression Line of D_i on S_i

Let $D_i = \eta_i - \xi_i$ and $S_i = \eta_i + \xi_i$, as in the classical SROC method. From (9) it follows that the covariance of D and S is equal to $\sigma_\eta^2 - \sigma_\xi^2$ and the variance of S is equal to $\sigma_\eta^2 + \sigma_\xi^2 + 2\sigma_{\xi\eta}$. The regression line therefore is

$$D = \overline{D} + \frac{(\sigma_\eta^2 - \sigma_\xi^2)}{(\sigma_\eta^2 + \sigma_\xi^2 + 2\sigma_{\xi\eta})}(S - \overline{S}).$$

The popularity of this summary line is possibly explained by the fact that it has an appealing interpretation. Given S, which can be interpreted as a proxy for the positivity criterion of the diagnostic test, this regression line estimates D, which can be interpreted as the diagnostic log odds ratio.

In terms of η and ξ, the regression line is

$$\eta = \bar{\eta} + \frac{\sigma_\eta^2 + \sigma_{\xi\eta}}{\sigma_\xi^2 + \sigma_{\xi\eta}}(\xi - \bar{\xi}). \tag{15}$$

This method is a kind of compromise between the vertical way of looking in the 1st method (median *TPR* given a specific value for the *FPR*) and the horizontal way of looking in the 2nd method (median *FPR* given a specific value for the *TPR*), because its slope lies between the slopes in (13) and (14).

The Rutter and Gatsonis[9] Summary ROC Curve

Their method leads to the summary line (see the following section)

$$\eta = \bar{\eta} + \frac{\sigma_\eta}{\sigma_\xi}(\xi - \bar{\xi}). \tag{16}$$

This line can also be interpreted as a sort of compromise between the regression of η_i on ξ_i and that of ξ_i on η_i, because the slope is equal to the geometric mean of the slopes of the 2 regression lines in (13) and (14).

The Major Axis Method

The last possibility we mention is to characterize the bivariate normal distribution for ξ_i and η_i by the major axis that runs through the extreme points of the contour ellipses (defined by points having the same density) of the estimated bivariate distribution. This results in the summary line[37]

$$\eta = \bar{\eta} + \frac{\sigma_\eta^2 - \sigma_\xi^2 + \sqrt{\left(\sigma_\eta^2 - \sigma_\xi^2\right)^2 + 4\sigma_{\xi\eta}^2}}{2\sigma_{\xi\eta}}(\xi - \bar{\xi}). \tag{17}$$

In fact, taking this line is analogous to summarizing a 2-dimensional distribution by its first principal component.

The summary ROC curves of methods 3 to 5 are symmetric in ξ and η; that is, if the labels of diseased and nondiseased test results and disease status are interchanged, the summary ROC curve does not change. For all of the mentioned summary lines, standard errors for the slope, intercept, and for η at a given value for ξ can be calculated using the delta method. Confidence intervals for the slope and intercept, and a confidence band for the summary line, are calculated using standard methods. A confidence band for the summary ROC curve is obtained by transforming the confidence band of the summary line. No extra programming or hand calculations are needed if a program like SAS Proc NLMixed is used that allows user-defined derived parameters.

Relationship With Model of Rutter and Gatsonis

From (5) and (7), it follows that the model of Rutter and Gatsonis can be written as

$$\begin{pmatrix}\xi_i\\ \eta_i\end{pmatrix} \sim N\left(\begin{pmatrix}(\bar{\theta}+X_0\bar{\alpha})e^{-X_0\beta}\\ (\bar{\theta}+X_1\bar{\alpha})e^{-X_1\beta}\end{pmatrix}, \begin{pmatrix}(\sigma_\theta^2+X_0^2\sigma_\alpha^2)e^{-2X_0\beta} & (\sigma_\theta^2+X_0X_1\sigma_\alpha^2)e^{-(X_0+X_1)\beta}\\ (\sigma_\theta^2+X_0X_1\sigma_\alpha^2)e^{-(X_0+X_1)\beta} & (\sigma_\theta^2+X_1^2\sigma_\alpha^2)e^{-2X_1\beta}\end{pmatrix}\right).$$

This specifies a bivariate normal distribution for (ξ_i, η_i), just as we do in (9). Note that the number of parameters is the same, too. Thus, the 2 models are essentially the same, only the parameterization is different. Rutter and Gatsonis[9] choose $X_0 = -1/2$ and $X_1 = 1/2$ and do not discuss other choices. One can check that their labeling leads to the summary line given by (16), with slope σ_η/σ_ξ. All other choices such that $X_0 = -X_1$ also lead to σ_η/σ_ξ. Alternative choices for X_0 and X_1 lead to other summary lines. For instance, the choice $X_0 = 0$ and $X_1 = 1$ leads to the η on ξ regression line given by (13). The choice $X_0 = 1$ and $X_1 = 0$ leads to the ξ on η regression given by (14). One can show that it is not possible to specify X_0 and X_1 such that it leads to the D on S regression line (15).

We conclude that our bivariate model is in principle identical to that of Rutter and Gatsonis.[9] A minor difference is the different parameterization. Another minor difference is that the slope in the Rutter and Gatsonis model is $e^{-\beta}$, and this is restricted to be positive. We do not restrict the slope in our model, although in practice negative slope estimates will typically not occur. An important practical difference is that Rutter and Gatsonis follow a laborious Bayesian estimation approach, whereas our method can be carried out conveniently using standard statistical packages. Furthermore, we think our method is easier to understand, because it simply assumes a standard random-effects model for the sensitivities and specificities simultaneously.

RESULTS

The Bivariate Model

Example 1: FNAC of the Breast[16]

We fitted the bivariate model as described in the section "The Bivariate Model" on the data of the 29 studies of the meta-analysis of Giard and Hermans.[16] The estimates of the means and variances of η_i and ξ_i resulting from the approximate and exact likelihood approach are presented in the upper part of Table 2. Based on these estimates, the results for the 5 different choices of the summary ROC curve ("Choice of Summary ROC Curve") are presented in the lower part of Table 2.

In Figure 3, the different ROC curves are depicted in the logit-logit space as well as in the ROC space. Also, the 95% coverage regions are given. These regions are based on the fitted bivariate distribution and estimate the area that contains approximately 95% of the true pairs of (logit(*FPR*), logit(*TPR*)), and (*FPR*, *TPR*), respectively.

From Table 2 and Figure 3, it is clear that the results of the exact and approximate approach are similar in this data example. The exact approach results in a somewhat more favorable average sensitivity and specificity. This was to be expected beforehand for 2 reasons. First, as mentioned in the section "The Standard SROC Method," adding 0.5 to the numbers in the 4-fold table, as is done in the approximate approach, results in estimated mean sensitivity and specificity that are biased downward, pushing the ROC curve away from the left upper corner. Second, as shown by Chang and others,[38] even if it is not needed to add 0.5, the estimates of the mean sensitivity and specificity are still somewhat biased toward 0.5. This is due to the fact that the approximate approach does not account for the correlation between the logit(*TPR*) and its variance and between the logit(*FPR*) and its variance.

From Table 2 and Figure 3, it is clear that the ROC curves differ substantially, especially for the first 2 choices "η on ξ" and "ξ on η". As one can see on the basis of the formulas for the slopes given in the section "Choice of Summary ROC Curve," the first 2 types ("η on ξ" and "ξ on η") give a kind of lower and upper bound for the estimated summary

Table 2 First Data Example: Fine-Needle Aspiration Cytological Examination (FNAC) of the Breast[16]

	Approximate Likelihood	Exact Likelihood
Parameter	**Estimate (se)**	**Estimate (se)**
mean logit(*TPR*) ($\bar{\eta}$)	1.774 (0.114)	1.839 (0.119)
mean logit(*FPR*) ($\bar{\xi}$)	−2.384 (0.201)	−2.547 (0.225)
var(logit(*TPR*)) (σ_η^2)	0.286 (0.093)	0.316 (0.104)
var(logit(*FPR*)) (σ_ξ^2)	0.990 (0.313)	1.297 (0.411)
cov(logit(*TPR*),logit(*FPR*)) ($\sigma_{\eta\xi}$)	0.146 (0.132)	0.141 (0.155)

	Approximate Likelihood		Exact Likelihood	
Type of Summary ROC	**α (se)**	**β (se)**	**α (se)**	**β (se)**
1. η on ξ	2.126 (0.32)	0.148 (0.13)	2.115 (0.32)	0.108 (0.12)
2. ξon η	6.431 (3.95)	1.954 (1.65)	7.560 (6.13)	2.246 (2.39)
3. *D* on *S*	2.680 (0.37)	0.380 (0.15)	2.647 (0.37)	0.318 (0.14)
4. Rutter and Gatsonis	3.054 (0.31)	0.537 (0.12)	3.096 (0.32)	0.494 (0.11)
5. Major axis	2.249 (0.42)	0.199 (0.17)	2.196 (0.41)	0.141 (0.15)

Note: In the upper part, estimates are given of the random intercept model (see "The Bivariate Model"), using approximate as well as exact likelihood. In the lower part, the parameter estimates are given for the 5 different choices of the summary receiver operating characteristic (ROC) curves discussed in the section "Choice of Summary ROC Curve."

ROC curves, and types 3 to 5 lie between these 2 curves. In fact, the slopes of choices 3 to 5 could be considered different kinds of "weighted averages" of the slopes of methods 1 and 2. In this example, the curves for approaches 3, 4, and 5 lie closer to the regression of η on ξ, but in general that is not the case. Results depend on the variances of ξ_i and η_i, and the covariance between them. The more similar the variances of ξ_i and η_i are, the more similar will be the results of approaches 3 to 5.

For all summary ROC curves given in Figures 3A and 3B, a 95% confidence band can be calculated. As an example, we have drawn in Figure 3C the "*D* on *S*" summary ROC curve together with its 95% confidence band. From all 5 types of summary ROC curves, this "*D* on *S*" summary ROC curve should be most comparable to the standard summary ROC curve from the Littenberg and Moses[2] (L&M) approach, which also estimates the regression of *D* on *S*. To compare the 2, we have also drawn the L&M summary curve and its confidence band in Figure 3C. The L&M summary ROC curve has a slope that is considerably steeper than our "*D* on *S*" curve, leading to larger estimated sensitivities if the specificity is small, and smaller estimated sensitivities if the specificity is large. This is not a general pattern, as will be seen from the second data example. Furthermore, it is seen that the L&M approach grossly underestimates the variability in the data, leading to a much too narrow confidence band. This is due to the fact that the L&M approach is based on a fixed-effects model, which erroneously assumes that there is no between-study variability.

Second Example: Imaging Tests for Coronary Artery Disease[20]

We fitted the bivariate model of the section "The Bivariate Model" on the data of the 149 studies included in the meta-analysis of Heijenbrok-Kal.[20] The estimates of the means and variances of η_i and ξ_i based on the approximate and exact likelihood approach are presented in the upper part of Table 3. Based on these estimates, the results for the 5 different choices of the summary ROC curve ("Choice of Summary ROC Curve") are presented in the lower part of Table 3.

In Figures 4A and 4B, the different summary ROC curves are given for the exact and the approximate approach in the logit-logit space as well as in the original ROC space.

In this example, the results of the approximate and exact likelihood approach are also similar. In the approximate likelihood approach, the variances of ξ and η are almost equal, which results in very few differences among methods 3 to 5. For the exact likelihood approach, the difference between the 2 variances is somewhat larger, leading to somewhat larger differences between methods 3 to 5. Notice that in Figure 4 considerably more than 5% of the studies fall outside the 95% coverage region. However, this is expected because the coverage ellipse

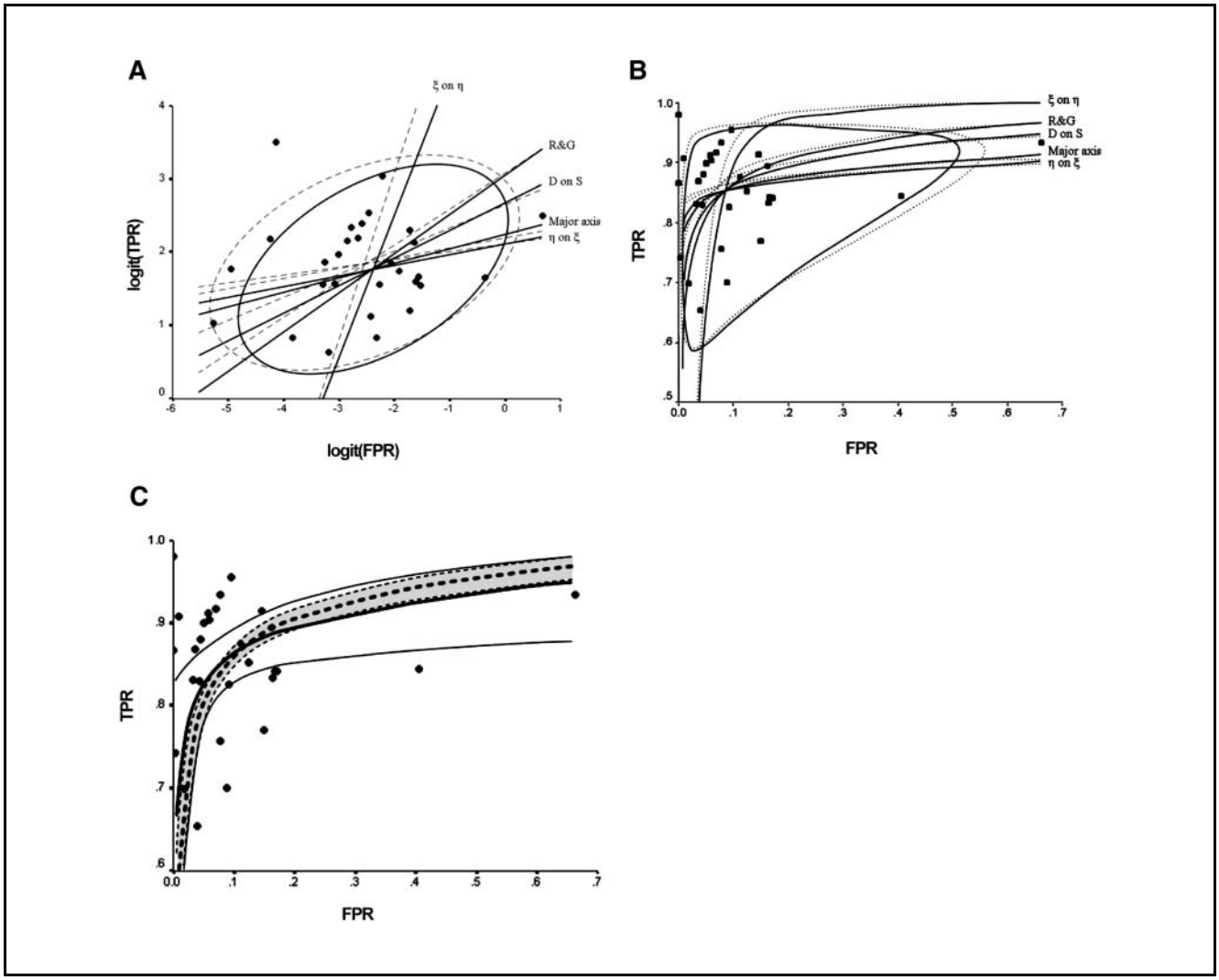

Figure 3 Summary receiver operating characteristic (SROC) curves for the 5 different choices of the SROC curve, as a graphical illustration of Table 2. The curves are presented in logit-logit space (A) as well as in the ROC space (B). Also, the 95% coverage regions are given as an ellipse in A and a triangle in B. The solid lines present the results of the approximate likelihood; the gray dashed lines present the results of the exact likelihood. C presents the SROC curve and 95% confidence band corresponding to the "D on S" model (solid lines) based on the estimated bivariate normal distribution (9) versus the SROC curve and 95% confidence band (dashed lines, with confidence band shaded) corresponding to the estimated fixed-effects Littenberg and Moses model. TPR = true positive rate; FPR = false positive rate; R&G = Rutter & Gatsonis.

describes the variation between the true pairs of sensitivity and specificity, whereas the points in the plot represent the estimates (observed) pairs of sensitivity and specificity. The observed points, of course, should show more variation due to within-study sampling variability. In Figure 4C, we compare again our "*D* on *S*" summary ROC with the standard L&M one. In contrast to the previous example, now the slope of the L&M ROC is smaller than that of our "*D* on *S*" curve. Again, it is clear that the L&M method leads to smaller standard errors.

DISCUSSION

Meta-analysis of diagnostic tests requires statistical techniques that analyze pairs of related summary statistics (e.g., sensitivity and specificity) rather than a single statistic. In the literature, numerous

Table 3 Second Data Example: Imaging Tests for Coronary Artery Disease[20]

	Approximate Likelihood		Exact Likelihood	
Parameter	**Estimate (se)**		**Estimate (se)**	
mean logit(*TPR*) ($\bar{\eta}$)	1.257 (0.057)		1.339 (0.061)	
mean logit(*FPR*) ($\bar{\xi}$)	−1.560 (0.071)		−1.851 (0.085)	
var(logit(*TPR*)) (σ_η^2)	0.333 (0.057)		0.406 (0.066)	
var(logit(*FPR*)) (σ_ξ^2)	0.337 (0.074)		0.585 (0.117)	
cov(logit(*TPR*),logit(*FPR*)) ($\sigma_{\eta\xi}$)	0.182 (0.049)		0.272 (0.065)	
	Approximate Likelihood		**Exact Likelihood**	
Type of Summary ROC	**α (se)**	**β (se)**	**α (se)**	**β (se)**
1. η on ξ	2.098 (0.21)	0.540 (0.13)	2.199 (0.18)	0.465 (0.10)
2. ξ on η	4.106 (0.69)	1.827 (0.45)	4.102 (0.58)	1.493 (0.31)
3. *D* on *S*	2.802 (0.26)	0.991 (0.17)	2.802 (0.23)	0.791 (0.12)
4. Rutter & Gatsonis	2.805 (0.21)	0.993 (0.13)	2.880 (0.19)	0.833 (0.10)
5. Major axis	2.796 (0.37)	0.987 (0.24)	2.677 (0.28)	0.723 (0.15)

Note: In the upper part, estimates are given of the random intercept model (see "The Bivariate Model"), using approximate as well as exact likelihood. In the lower part, the parameter estimates are given for the 5 different choices of the summary receiver operating characteristic (ROC) curves discussed in the section "Choice of Summary ROC Curve."

meta-analyses are published in which one is interested in meta-analyzing only sensitivities or only specificities. For these situations, the standard method of analysis is the DerSimonian-Laird univariate random-effects model. The method we propose in this article is a direct extension of that approach. We analyze sensitivities and specificities simultaneously using a 2-dimensional random-effects model. Our method could also be seen as an extension of the approach of Littenberg and Moses.[2] Their model implies a linear relationship between the logit transformed sensitivity and specificity, which can be transformed into ROC space to obtain a summary ROC curve. Despite its many drawbacks, their approach still seems to be the most popular method for meta-analysis of diagnostic accuracy data where pairs of sensitivity and specificity per study are available. This is probably due to the fact that the method is very easy to carry out in practice. In this article, we have shown that the bivariate meta-analysis model addresses all its shortcomings.

The method of Rutter and Gatsonis[9] is also an appropriate alternative for the Littenberg and Moses method, and recently it was pointed out how this method can be performed in a non-Bayesian way using standard statistical software.[30] We have shown that their model is essentially the same as ours, only the parameterization is different and it leads to a special type of summary ROC. In our opinion, their model is less straightforward and more difficult to understand. The way we present the bivariate model fits into the standard framework of multivariate meta-analysis,[13,25,35,39] thereby bringing the meta-analysis of this kind of diagnostic back into mainstream meta-analysis methods, which can be fitted straightforwardly in standard statistical software. The SAS syntax we used for the examples is given in the appendix to make the method easily accessible to meta-analysts. The approach presented in this article is straightforwardly extended with covariates. In a general linear mixed model program like Proc Mixed, both the mean logit(*FPR*) and mean logit(*TPR*) can be allowed to depend on covariates. If a generalized linear mixed-model program such as SAS Proc NLMixed is used, there are many more possibilities, depending on how the model is parameterized.

Our contribution in this article has been to compare different types of summary ROC curves by showing how they are all related to our bivariate model. We discussed 5 types of summary ROC curves, each of which has its own interpretation and properties. In the Littenberg and Moses approach, the choice is made explicitly as the regression of *D* on *S*. In the approach of Rutter and Gatsonis,[9] the choice is implicitly made, and we pointed out that it is a kind of geometric mean between the regression line of logit(*TPR*) on logit(*FPR*) and the regression line of logit(*FPR*) on logit(*TPR*). Thus the 2 methods estimate different summary curves and the resulting curves are therefore in principle not the same. If one wants to describe the median sensitivity of studies with a fixed value of the specificity, one can choose the regression

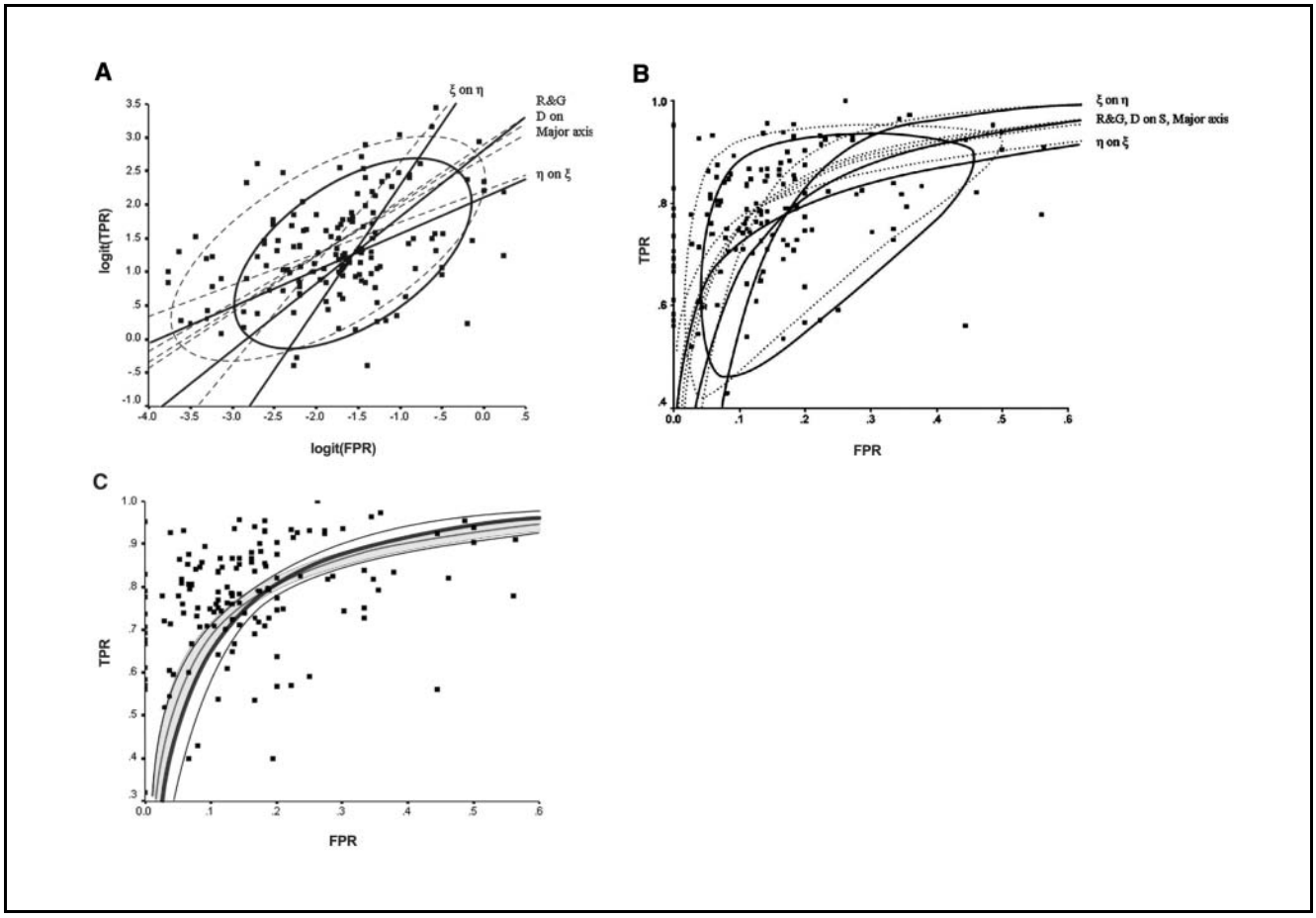

Figure 4 Second data-example of Heijenbrok-Kal and others. Receiver operating characteristic (ROC) curves for the 5 different choices of the summary ROC (SROC) curve, as a graphical illustration of Table 3. The curves are presented in logit-logit space (A) as well as in the ROC space (B). Also the 95% coverage regions are given as an ellipse in A and a triangle in B. The solid lines present the results of the approximate likelihood; the gray dashed lines present the results of the exact likelihood. In C, the SROC curves of our random intercept and slope model (solid lines) versus the fixed Littenberg and Moses model (gray dashed lines) are given together with their 95% confidence bands. TPR = true positive rate; FPR = false positive rate; R&G = Rutter & Gatsonis.

of η_i on ξ_i, but if one prefers to describe the median specificity with a fixed sensitivity, one can choose the regression of ξ_i on η_i. The other 3 summary ROC curves are compromises between these 2 options.

Notice that in our approach no assumptions about individual study curves have to be made. For example, the method does not require an underlying continuous diagnostic test, and hence it can also be applied to intrinsically dichotomous tests. Our bivariate random effects model simply leads to a description of the distribution of the pairs (ξ_i, η_i). The summary ROC curve is just a 1-dimensional representation of this distribution and cannot be interpreted as a kind of average curve or a curve typical for the study-specific ROC curves. It can have a shape that is very different from the study-specific shapes. The different kinds of summary ROC curves discussed in the section "Choice of Summary ROC Curve" are still interpretable, even if the test is intrinsically dichotomous. For example, the regression of η_i on ξ_i simply describes the median sensitivity of studies with a fixed value of the specificity.

We fitted our models with standard software based on straightforward likelihood methods. In our examples, this approach worked well, although sometimes some convergence problems were met. In our 2 clinical data examples, these convergence problems were addressed by specifying better starting values. However, we can imagine that, especially for small meta-analyses, this could be more of a problem. An alternative is to fit the models in a Bayesian way. This can be

done using the free available software program WinBUGS.[40] The Bayesian approach has the advantage of being more flexible; for instance, one can assume nonnormal parametric distributions for logit(TPR) and logit(FPR). Also in applications with a relatively small number of studies, the Bayesian method might perform better, because the standard likelihood is based on large sample theory. A disadvantage is that it is more time consuming and is less easily done by nonstatisticians.

The bivariate model we proposed in this article can be fitted using approximate or exact likelihood. Using approximate likelihood has the advantage that a general linear mixed model program can be used, which is widely available. The exact likelihood method can only be used if one has an appropriate generalized linear mixed model program available. Unfortunately, these programs are still rather scarce. Simulation studies are needed to compare the 2 approaches.

APPENDIX

In this appendix, we provide the SAS syntax needed to reproduce the results given in Tables 2 and 3. First we describe the data format that is needed. We have 2 records per study, 1 for the diseased and 1 for the healthy group, as in the following table.

study	group	n	npos	disease	healthy	y	est
.	.	.	.	.	.	.	0.20
.	.	.	.	.	.	.	0.10
.	.	.	.	.	.	.	0.20
1	1	1009	70	0	1	−2.58974	0.01525
1	2	1068	979	1	0	−2.58974	0.01219
2	3	166	3	0	1	−3.84405	0.29183
2	4	73	51	1	0	−3.84405	0.06386
3	5	949	25	0	1	−2.77988	0.01914
:	:	:	:	:	:	:	:

The meaning of the variables is: study = number of the study; group = unique identifier for the diseased and healthy group; n = number per group; npos = number with positive diagnostic test; disease = 0 for healthy group, = 1 for diseased group; healthy = 1 for healthy group, = 0 for diseased group; y = ln(npos/(n-npos)); est = 1/(npos + 0.5) + 1/(n-npos + 0.5). The first 3 lines have only a nonmissing value for the variable est. These 3 values serve as starting values for the variance of ξ_i.

The following syntax produces the approximate likelihood results given in the upper part of Table 2.

```
proc mixed cl method=ml data=giardcol;                  # call procedure; cl asks for confidence intervals of covariance parameters;
class study d group;                                    # study, d and group are classification variables;
model y=disease healthy/noint s cl                      # model with indicator variables disease and healthy as
  covb ddf=1000,1000;                                     explanatory variables for log odds. Covb asks for
                                                          covariance matrix of fixed effects parameters;
random disease healthy / subject=study type=un s;       # indicators of diseased and healthy group are random
                                                          effects, possibly correlated within a study and
                                                          independent between studies; covariance matrix is
                                                          unstructured. Print empirical Bayes estimates 's';
parms/parmsdata=giardcol eqcons=4 to 61;                # datafile giardcol.sd2 contains the variable 'est' with
                                                          starting values for the three covariance parameters of
                                                          the random effects together with the 58 within study-group
                                                          variances. The latter are assumed to be known and should
                                                          be kept fixed;
repeated/group=group;                                   # each group in a study (diseased and healthy) has its own
                                                          within study-arm variance; within study estimation errors
                                                          are independent;
run;
```

(continued)

APPENDIX (continued)

The different summary ROC curves have to be calculated by hand based on the output of the program. The same results can also be obtained with SAS Proc NLmixed. The advantage is that the parameters of the summary ROC curves can be specified as derived parameters. The syntax is as follows.

```
proc nlmixed data=giardcol;                                  # call procedure
parms meaneta=1.8 meanksi=-2.4 vareta=0.3
   varksi=1 covksieta=0.15 ;                                 # choose starting values for means, variances, and covariance
model y~normal(eta*disease+ksi*healthy,est);                 # log odds y are alternately normally distributed around eta
                                                               and ksi; disease and healthy are indicator variables; 'est'
                                                               contains the within-study variances;
random ksi eta~normal([meanksi,meaneta],[varksi,             # the shrunk parameters ksi and eta are normally distributed
   covksieta,vareta]) subject=study;                           around their common means, with between-study variances
                                                               varksi and vareta and covariance covksieta;
estimate 'eta on ksi: beta' covksieta/varksi;                # estimate slope of ksi on eta regression line
estimate 'eta on ksi: alpha' meaneta-covksieta/              # estimate intercept of eta on ksi regression line
   varksi*meanksi;
estimate 'ksi on eta: beta' vareta/covksieta;                # estimate slope of ksi on eta regression line
estimate 'ksi on eta: alpha' meaneta-vareta/covksieta*       # estimate intercept of ksi on eta regression line
   meanksi;
estimate 'D on S: beta' (vareta+covksieta)/(varksi+          # estimate slope of D on S regression line
   covksieta);
estimate 'D on S: alpha' meaneta-(vareta+covksieta)/         # estimate intercept of D on S regression line
   (varksi+covksieta)*meanksi;
estimate 'R&G: beta' (vareta**0.5)/(varksi**0.5);            # estimate slope of R&G regression line
estimate 'R&G: alpha' meaneta-(vareta**0.5)/                 # estimate intercept of R&G regression line
   (varksi**0.5)*meanksi;
estimate 'major axis: beta' (vareta-varksi+
   ((vareta-varksi)**2+4*covksieta**2)**0.5)/
(2*covksieta);                                               # estimate slope of major axis regression line
estimate 'major axis: alpha' meaneta-(vareta-varksi+
   ((vareta-varksi)**2+4*covksieta**2)**0.5)/
(2*covksieta)*meanksi;                                       # estimate intercept of major axis regression line
run;
```

The following syntax reproduces the right half (exact likelihood) of Table 2.

```
proc nlmixed data=giardcol;                                  # call procedure
parms meaneta=1.8 meanksi=-2.4 vareta=0.3
   varksi=1 covksieta=0.15 ;                                 # choose starting values for means, variances, and covariance
pi=1/(1+exp(-(eta*disease+ksi*healthy)));                    # calculating the 'true' TPR and FPR (pi)
model npos~binomial(n,pi);                                   # the positive numbers in both groups follow binomial
                                                               distributions
random ksi eta ~ normal([meanksi,meaneta],                   # the shrunk parameters ksi and eta are normally distributed
   [varksi,covksieta,vareta]) subject=study;                   around their common means, with between-study variances
                                                               varksi and vareta and covariance covksieta;
estimate 'eta on ksi: beta' covksieta/varksi;                # estimate slope of eta on ksi regression line
estimate 'eta on ksi: alpha' meaneta-covksieta/varksi*       # estimate intercept of eta on ksi regression line
   meanksi;
estimate 'ksi on eta: beta' vareta/covksieta;                # estimate slope of ksi on eta regression line
estimate 'ksi on eta: alpha' meaneta-vareta/                 # estimate intercept of ksi on eta regression line
   covksieta*meanksi;
```

(continued)

APPENDIX (continued)

```
proc nlmixed data=giardcol;                                   # call procedure
estimate 'D on S: beta' (vareta+covksieta)/                   # estimate slope of D on S regression line
  (varksi+covksieta);
estimate 'D on S: alpha' meaneta-(vareta+covksieta)/          # estimate intercept of D on S regression line
  (varksi+covksieta)*meanksi;
estimate 'R&G: beta' (vareta**0.5)/(varksi**0.5);             # estimate slope of R&G regression line
estimate 'R&G: alpha' meaneta-(vareta**0.5)/                  # estimate intercept of R&G regression line
  (varksi**0.5)*meanksi;
estimate 'major axis: beta' (vareta-varksi+                   # estimate slope of major axis regression line
  ((vareta-varksi)**2+4*covksieta**2)**0.5)/(2*covksieta);
estimate 'major axis: alpha' meaneta-(vareta-varksi+
  ((vareta-varksi)**2+4*covksieta**2)**0.5)/
(2*covksieta)*meanksi;                                        # estimate intercept of major axis regression line
run;
```

ACKNOWLEDGMENT

The work of the second author was financially supported by a program grant from The Netherlands Organization for Health Research and Development (ZonMW 912-04-016).

REFERENCES

1. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Smith DG, Altman DG, eds. Systematic Reviews in Health Care: Meta-Analysis in Context. BMJ. 2001;323(7305):157–62.

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

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

4. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull. 1995;117:167–78.

5. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol. 1995;48: 119–30; discussion 131–2.

6. Hellmich M, Abrams KR, Sutton AJ. Bayesian approaches to meta-analysis of ROC curves. Med Decis Making. 1999;19:252–64.

7. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol. 1999; 52:943–51.

8. Kester AD, Buntinx F. Meta-analysis of ROC curves. Med Decis Making. 2000;20:430–9.

9. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865–84.

10. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med. 2002;21: 1237–56.

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

12. McClish DK. Combining and comparing area estimates across studies or strata. Med Decis Making. 1992;12:274–9.

13. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002;21:589–624.

14. Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. Baseline risk as predictor of treatment benefit: three clinical meta-reanalyses. Stat Med. 2000;19:3497–518.

15. Arends LR, Voko Z, Stijnen T. Combining multiple outcome measures in a meta-analysis: an application. Stat Med. 2003;22: 1335–53.

16. Giard RWM, Hermans J. The value of aspiration cytologic examination of the breast—a statistical review of the medical literature. Cancer. 1992;69:2104–10.

17. Mushlin AI. Diagnostic tests in breast cancer: clinical strategies based on diagnostic probabilities. Ann Intern Med. 1985;103:79–85.

18. Committee. HaPP. The use of diagnostic tests for screening and evaluating breast lesions. Ann Intern Med. 1985;103:147–51.

19. Dixon JM, Clarke PJ, Crucioli V, Dehn TCB, Lee ECG, Greenal MJ. Reduction of the surgical excision rate in benign breast disease using fine needle aspiration cytology with immediate reporting. Br J Surg. 1987;74:1014–16.

20. Heijenbrok MH. Assessment of Diagnostic Imaging Technologies for Cardiovascular Disease. Rotterdam, Netherlands: Epidemiology and Biostatistics, Erasmus MC; 2004.

21. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177–88.

22. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med. 1999;18:321–59.

23. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med. 1998;17:841–56.

24. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. Br Med J. 1994;309:1351–5.

25. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. Stat Med. 1998;17:2537–50.

26. Senn S. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials (letter). Stat Med. 1994;13:293–6.

27. Carroll RJ, Ruppert D, Stefanski LA. Measurement Error in Nonlinear Models. London: Chapman & Hall; 1995.

28. Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf Med. 1990;29:12–22.

29. Rothman K, Greenland S. Modern Epidemiology. Philadelphia: Lippincott Williams & Wilkins; 1998.

30. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agrees closely with those of a full Bayesian analysis. J Clin Epidemiol. 2004;57:925–32.

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

32. Nederkoorn PJ, van der Graaf Y, Hunink MG. Duplex ultrasound and magnetic resonance angiography compared with digital subtraction angiography in carotid artery stenosis: a systematic review. Stroke. 2003;34:1324–32.

33. Oei EH, Nikken JJ, Verstijnen AC, Ginai AZ, Myriam Hunink MG. MR imaging of the menisci and cruciate ligaments: a systematic review. Radiology. 2003;226:837–48.

34. Harbord R, Steichen T. Metareg. Stata Module to Perform Meta-Analysis Regression. Boston, MA: Statistical Software Components S4446201; 2004.

35. van Houwelingen HC, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. Stat Med. 1993;12:2272–84.

36. McIntosh MW. The population risk as an explanatory variable in research syntheses of clinical trials. Stat Med. 1996;15:1713–28.

37. Johnson R, Wichern D. Applied Multivariate Statistical Analysis. London: Prentice Hall; 2002.

38. Chang B, Waternaux C, Lipsitz S. Meta-analysis of binary data: which within-study variance estimate to use? Stat Med. 2001;20:1947–56.

39. Kalaian HA, Raudenbush SW. A multivariate mixed linear model for meta-analysis. Psychol Methods. 1996;1:227–35.

40. Spiegelhalter DJ, Thomas A, Best NG, Lunn D. WinBUGS User Manual, version 1.4. Cambridge, UK: MRC Biostatistics Unit; 2003.