WHEN ARE SUMMARY ROC CURVES APPROPRIATE FOR DIAGNOSTIC META-ANALYSES?

F M Chappell\textsuperscript{1} and \textsuperscript{2}, G M Raab\textsuperscript{1}, J M Wardlaw\textsuperscript{2}

\textsuperscript{1}School of Nursing Midwifery and Social Care, Napier University, Edinburgh, UK
\textsuperscript{2}Department of Clinical Neurosciences, University of Edinburgh, Edinburgh, UK

SUMMARY

Diagnostic tests are increasingly evaluated with systematic reviews and this has lead to the recent developments of statistical methods to analyse such data. The most commonly used method is the summary receiver operating (SROC) curve which can be fitted with a non-linear bivariate random effects model. This paper focuses on the practical problems of interpreting and presenting data from such analyses. Firstly, many meta-analyses may be underpowered to obtain reliable estimates of the SROC parameters. Secondly, the SROC model may be inappropriate. In these situations, a summary with two univariate meta-analyses of the true and false positive rates (TPRs and FPRs) may be more appropriate. We characterise the type of problems that can occur in fitting these models and present an algorithm to guide the analyst of such studies, with illustrations from analyses of published data. A set of R functions, freely available to perform these analyses, can be downloaded from (http://www2.napier.ac.uk/depts/fhls/DiagMeta).

\textsuperscript{*} Correspondence to Francesca Chappell, Francesca Chappell, University of Edinburgh, Dept of Clinical Neurosciences, Bramwell Dott Building, Western General Hospital, Crewe Road, Edinburgh EH4 2XU. Email francesca.chappell@ed.ac.uk
1 INTRODUCTION

Diagnostic tests are an important aspect of current medical practice in order to establish a diagnosis, determine appropriate treatment, monitor treatment, screen for disease or monitor substances such as illicit drugs. Once the test is carried out a decision is taken to classify each case as ‘diseased’ or ‘not diseased’ according to some threshold level. The threshold may be an actual measure (such as the level of a metabolite or the measured width of a blood vessel), or an ordered category such as ‘occluded/ partially occluded/ not occluded’. Once the true disease status of each subject is determined the true positive and the false positive rates (TPR and FPR respectively) can be estimated at each level of this threshold and the resulting data can then be plotted as an empirical receiver operator characteristic (ROC) curve. Alternatively the ROC curve can be assumed to have parametric form and methods have been developed to fit these; see [1,2, chapter 4,3, chapter 4,4] for reviews.. Note that the ROC curve does not require the actual threshold values, merely that the ranking of subjects is known. Models used for ROC curve analyses have assumed that TPR and FPR are related by corresponding to cumulative probabilities of two related distributions, usually both assumed to be Normal. The independence of the ROC curve from the location or scale of the original data means that only two parameters are needed to describe such parametric ROC curves.

The need to perform systematic reviews and meta-analyses of diagnostic tests has led to the development of methods for summarizing data from different diagnostic studies. Such data usually report results for a particular threshold, as used in clinical practice, so that the data consist of 2 x 2 tables that cross-tabulate true disease status with test results, with one table per study. Recent methods of synthesis of such studies
use summary ROC curves (SROCs) to describe a set of diagnostic studies, a methodology which is advocated by the Cochrane Diagnostic Test Accuracy Working Group [5]. The simplest method only estimates the SROC curve [6] but more complex methods [7-10] model the variation of individual studies around the SROC curve. The TPR and FPR define random effects for each study, which can alternatively be expressed as a latent threshold parameter and accuracy [10]. The latent threshold parameter increases as TPR and FPR increase, while the accuracy parameter increases with increasing TPR and decreasing FPR. The SROC curve combines the information from all studies, and illustrates how the TPR and FPR vary at the average accuracy found from the contributing studies, when the threshold is allowed to vary. Although one formulation of this model [9] has proposed definitions of threshold and accuracy in terms of TPR and FPR, we will show that the latter are not unique and can lead to different SROC curves, some of which may provide inappropriate summaries.

The data from a diagnostic meta-analysis differ from those from an individual diagnostic study. The individual TPRs and FPRs from a meta-analysis will not usually produce a monotonic plot. In this paper we review our experience in fitting SROC curves to a wide range of diagnostic meta-analyses and make recommendations as to how to decide whether an SROC is appropriate. In section 2 we outline the methods that have already been proposed for SROC curves and introduce notation. Section 3 discusses the methods that can be used to fit SROC curves and illustrates how certain types of data will cause the methods to fail. In Section 4 we present an algorithm to act as a guide to carrying out diagnostic meta-analyses and illustrate it with real data sets, Section 5 discusses the power of diagnostic meta-analyses to estimate SROCs and summarises our recommendations and Section 6 sets out our conclusions. We
also introduce R functions [11] that can be used for analysis and, in particular, to provide graphical output to aid the analyst in deciding how to summarize the data.

2 SUMMARY ROC CURVES

The simplest and most commonly used method for diagnostic meta-analysis is the Moses-Littenberg fixed-effects method [6]. A straight line is fitted to the logits of \( FPR_i \) and \( TPR_i \) of the \( i \)th study (or more precisely the sum and difference of their logits), and its slope and intercept give the parameters of the SROC curve. Despite the ease of implementation, the statistical basis of the Moses-Littenberg can be criticised. It makes no allowance for the non-linear transformation, for the binomial variance of individual studies or for the fact that both the \( x \) and \( y \) variables are measured with error and it does not provide valid estimates of precision.

Two random effects methods have been developed that overcome the limitations of the Moses-Littenberg method and permit studies to vary in both threshold and accuracy. These are bivariate methods [8,12] and the hierarchical SROC [7,9] method, recently shown both to use the same underlying statistical model [10]. For the \( i \)th study, the data used to estimate \( FPR_i \) and \( TPR_i \) are assumed to be independent binomials with \( x_i = \logit FPR_i \) and \( y_i = \logit TPR_i \). In the bivariate model, the \( x_i \), and \( y_i \) are modelled directly as a bivariate Normal distribution with mean \( (\mu_x, \mu_y)^T \), and variance-covariance matrix \( \Omega \), which may also be defined in terms of the variances \( \sigma_x^2 \), \( \sigma_y^2 \), and correlation \( \rho \). In the hierarchical SROC method \( x_i \) and \( y_i \) are modelled as linear combinations of parameters representing latent diagnostic
threshold $\theta_i$ and accuracy $\alpha_i$ parameters which are independent and Normally distributed.

When all studies have TPRs and FPRs assumed to lie exactly on a single ROC curve, the threshold can be equally well taken as the TPR, FPR, or any measure that increases monotonically with the TPR and FPR. However this is not the case for the random effects models, where each study has a different ROC curve. In particular, we show below that, for the random effects case, we can define the threshold parameter as any linear combination of $x$ and $y$. The particular choice of coefficients will not affect the estimation of the parameters of the underlying model, but each choice will generate a different SROC curve.

We can define a threshold

$$\theta_i \propto k_1 x_i + k_2 y_i$$  \hspace{1cm} (1)

where $k_1$ and $k_2$ are any positive constants. Following Rutter and Gatsonis [9] we assume that $\theta$ and $\alpha$, the accuracy parameter, are bivariate Normal and independent. Thus it follows from the equivalence of the HSROC and bivariate models that the accuracy becomes:

$$\alpha_i \propto -(k_2 \sigma_y^2 + k_1 \rho \sigma_x \sigma_y) x_i + (k_1 \sigma_x^2 + k_2 \rho \sigma_x \sigma_y) y_i$$  \hspace{1cm} (2)

so that all such models are equivalent to the bivariate model but the choice of $k_1$ and $k_2$, or more precisely their ratio, will determine the slope of the SROC curve in logit space which becomes

$$\frac{k_2 \sigma_y^2 + k_1 \rho \sigma_x \sigma_y}{k_1 \sigma_x^2 + k_2 \rho \sigma_x \sigma_y}.$$
Four SROC curves, or linear combinations, have been suggested in the literature. We compare their equations in logit-space. From Rutter and Gatsonis [9] we get the SROC curve

\[ y = \frac{\sigma_y}{\sigma_x}(x - \mu_y) + \mu_y \]  

(3)
giving \( k_1/k_2 = \sigma_y/\sigma_x \), showing as Harbord et al [10] have noted, that its slope does not depend on \( \rho \), and will be positive, even when \( \rho \) is negative. Reitsma et al [8] discuss SROC curves derived from the expected regressions of \( x \) on \( y \), \( y \) on \( x \), or the orthogonal regression line, giving SROCs

\[ y = \rho \frac{\sigma_y}{\sigma_x}(x - \mu_y) + \mu_y \]  

(4)

\[ y = \frac{1}{\rho} \frac{\sigma_y}{\sigma_x}(x - \mu_y) + \mu_y \]  

(5)

\[ y = \frac{\sigma_y^2 - \sigma_x^2 + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}}{2\sigma_{xy}}(x - \mu_x) + \mu_y. \]  

(6)

We obtain (4) if \( k_1 = 0 \) and \( k_2 \) is any constant, and (5) if \( k_1 \) is any constant and \( k_2 = 0 \), so that the thresholds are taken as the true values of the TPR and FPR respectively. For (6) the SROC is the axis with the positive slope of the fitted bivariate Normal distribution with the attractive property that \( \theta \) and \( \alpha \) remain orthogonal in \( (x, y) \) space as well as being independent, which is not the case for (3), (4), or (5). For (6) \( k_1 \) and \( k_2 \) are proportional to the elements of the eigenvector corresponding to the maximum eigenvalue of \( \Omega \) if \( \rho \) is positive, and this SROC curve summarizes the direction in logit space that lies closest to the estimated bivariate distribution of \((x, y)\). For these reasons we recommend (6) as the most appropriate choice of SROC and we use this choice in the illustrative examples below.
The four SROC curves will coincide when $\rho = 1$, which implies that $\Omega$ is singular and $\sigma^2_\alpha = 0$, implying a fixed effect model where the only source of between-study variation is the threshold. The four choices of SROC curve for the meta-analysis of Mowatt et al [13] are illustrated in Figure 1. In this example, with a positive estimate of $\rho$, the SROCs (3) and (6) are fairly close, but this will not always be the case.

Figure 1: Different choices of SROC curve illustrated for example 1 [13] along with 95% contour of the fitted bivariate density, raw data (dots) and ML estimates of random effects (squares). Dashed line is equation (3), dotted lines equations (4) and (5) and solid line equation (6), (a) in logit space and (b) in ROC space.

3 METHODS OF ESTIMATION

The HSROC and bivariate Normal can each be fitted by maximum likelihood (ML). Since they describe identical models it follows that the corresponding ML estimates should be identical and this has been shown empirically by Harbord et al [10]. However the two different formulations lead to differences in the methods of
maximising the likelihood and the availability of software to implement them. Both are non-linear random effects models for which no explicit expression for the likelihoods exists. Both of these models can be fitted by methods that calculate the likelihood by numerical approximation [14] or by MCMC methods [15], which implement a Bayesian model that requires specification of the priors for the hyperparameters, \( \left( \mu_\alpha, \mu_\beta, \beta, \sigma_\alpha^2, \sigma_\beta^2 \right) \). As the bivariate Normal model is a generalised linear model further methods of estimation based on the linearization of the likelihood at the solution can be used for estimation [16,17].

Interval estimates for the hyperparameters can be obtained directly from the MCMC methods, whereas other methods use Wald methods based on a quadratic approximation to the likelihood at the solution. The Wald method can be a very poor approximation for these models, although an appropriate transformation can improve it [18]. A hybrid solution that involves obtaining estimates by one of the other methods and using MCMC for interval estimates has been implemented in some packages.

Estimates of the within-study random effects \( (\alpha_i, \beta_i) \) or \( (x_i, y_i) \) are usually estimated conditional on the ML estimates of the hyperparameters. These are often called ‘shrunken’ estimates, because they lie between the observed values for the individual study and the overall mean, with ‘shrinkage’ being pronounced when between-study variances are small relative to the corresponding within study variances of the estimated rates. These shrunken estimates for \( (x_i, y_i) \) are illustrated for the fit in Figure 1. We will see below that these estimates are a valuable tool for judging whether an SROC provides a reasonable summary of the data.
When there are sufficient data to give precise estimates and the likelihood at the solution is regular, all of the methods described will give very similar results. When this is not the case, some of the methods will report a failure of the fitting and may provide partial results with no standard errors, or those that are obviously wrong. These practical problems have been noted with the estimation of parameters in both the HSROC and bivariate versions of the model. Riley et al [19] report a simulation of a meta-analyses with 10 studies each where the program failed to converge correctly in 397 out of 1000 simulations. It was noted that this tended to occur when $\hat{\rho}$ was close to 1.

We will use the term ‘threshold effect’ for the situation when some or all of the variation between studies can be explained by differences in $\theta_i$. If present, the threshold effect will cause $TPR_i$ and $FPR_i$ to rise and fall together, and is included in the bivariate model as a positive correlation coefficient $\rho$. However, the estimate of $\rho$ may be close to zero or even negative. This may not invalidate the bivariate model as such, but it does mean that the data should not be summarized by a SROC curve. Conversely, a failure to converge because the maximum of the likelihood lies on the boundary with $\rho = 1$ should not necessarily be regarded as a failure of the model, and it should not be concluded that summarising the data with a SROC curve is invalid. Such a fit is equivalent to a SROC curve fixed effects model. A further type of failure to fit can occur when the ML estimate of either or both of $\sigma_x$ to $\sigma_y$ approaches zero, corresponding to no between-study heterogeneity in the rates. In this case there will be no information in the data to estimate $\rho$. If $\rho$ is poorly estimated due to a very
small estimate of between-study variance, there may be no particular advantage in using the bivariate model over two univariate meta-analyses, one each for $TPR$ and $FPR$. When a diagnostic meta-analysis is summarised by a SROC curve we need to identify when each of these three different situations are present.

4 EXAMPLES FROM THE LITERATURE

We now present an algorithm (Figure 2) of what analyses to perform and how to interpret them, and illustrate its use with examples from the literature. To analyse and interpret data from diagnostic meta-analyses we have developed a set of routines for the R package to guide an inexperienced user in understanding the results. There are three key R functions that perform the following functions:-

- **plotfor** – which displays the data as two forest plots for the TPR and the FPR
- **bivarROC** – which fits the bivariate model and plots the fitted orthogonal regression SROC curve along with shrunken estimates of random effects.
- **twouni** – which carries out two univariate meta-analyses for TPR and FPR

The functions and instructions on how to download the R program and use the functions can be downloaded from http://www2.napier.ac.uk/depts/fhls/diagmeta. The R functions, which make extensive use of Douglas Bates’ lme4 library, have several advantages over other packages (e.g.SAS [20] ) that have been used for diagnostic meta-analyses. In particular R allows graphical diagnostics to be presented with the analysis and the lme4 package obtains interval estimates of the parameters and estimates of the posterior densities (via MCMC) even when a degenerate fit has been obtained. Also the R package [11] is freely available to all and the license allows it to be incorporated freely into other applications. The results of the analyses can be
used to decide on the path through the algorithm but other information, such as prior information that different thresholds exist, can be used to influence the decisions.

**Figure 2:** Algorithm for the meta-analysis of diagnostic test data.

The first step in both the analysis and the reporting of a diagnostic meta-analysis should be to display the data as two forest plots. This allow inclusion of additional
information about individual studies which help readers of the meta-analysis grasp more rapidly the totality of the available data and the likely robustness. For example, on the same line as each box and whisker, the name of the respective study and year of publication, the sample size and number of true and false positives and negatives can all be included, something which it is virtually impossible to do effectively with a SROC curve layout. Ranking the data by the TPR or the FPR can assist interpretation. The forest plots used to illustrate the examples in this paper are shown without a summary line, as is appropriate at the exploratory stage. When the analysis is complete an appropriate fixed-effects or random-effects summary should be added to each of the forest plots.

Example 1 – algorithm path A, B, C, D

Example 1 uses 15 studies of ECG from the meta-analysis of Mowatt et al [13]. As discussed above, the SROC models require heterogeneity in the individual studies’ FPRs and TPRs. A first step is therefore to check for heterogeneity, visually using the standard tool of forest plots of TPR and FPR (Figure 3). The forest plot may also give an indication of whether a threshold effect is present if the FPRs and TPRs appear to increase together. The lack of overlap in the confidence intervals in Figure 3 suggests heterogeneity in the rates, and there is a suggestion of a threshold effect with TPR and FPR increasing together.
Thus we fit the bivariate model and obtain the estimates in Table 1 and kernel plots of the posterior densities shown in Figure 4. The posteriors and interval estimates are based on 10,000 MCMC samples with improper priors as described by Box and Tiao [21, page 460], which implies a prior for $\rho$ that has modes at +1 and -1, but with positive density over all of its range (see Figure 6). The posterior probability that $\rho$ is positive is estimated as 0.95 from the MCMC samples. These results suggest that a bivariate model is an appropriate summary and the data can be summarised by the SROC curve shown in Figure 1b and the estimates in Table 1.

Figure 3: Forest plot for Example 1 [13] with studies sorted by the FPR.

---

† Douglas Bates, personal communication. The formula in [21] implies a Wishart distribution with 2 degrees of freedom. Using 3 degrees of freedom would give a uniform prior for $\rho$, which might be more appropriate here.
Table 1: Estimates of bivariate model for Example 1.

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>MCMC median</th>
<th>lower limit</th>
<th>upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>average TPR%</td>
<td>64.3</td>
<td>64.1</td>
<td>51.9</td>
<td>74.7</td>
</tr>
<tr>
<td>average FPR%</td>
<td>33.3</td>
<td>33.2</td>
<td>25.7</td>
<td>41.5</td>
</tr>
<tr>
<td>SD logit TPR</td>
<td>0.776</td>
<td>0.899</td>
<td>0.601</td>
<td>1.481</td>
</tr>
<tr>
<td>SD logit FPR</td>
<td>0.482</td>
<td>0.575</td>
<td>0.337</td>
<td>1.022</td>
</tr>
<tr>
<td>correlation</td>
<td>0.541</td>
<td>0.557</td>
<td>-0.093</td>
<td>0.886</td>
</tr>
</tbody>
</table>

Figure 4: Kernel density estimates of posterior densities of variance parameters for Example 1.

Example 2 – algorithm path A, E or A, B, C, F, E.

Example 2 uses a meta-analysis of 11 studies of computed tomographic angiography (CTA) for assessing the carotid artery in stroke patients [22]. The forest plot is shown in Figure 5 and immediately suggests that an SROC model is inappropriate because the overlapping confidence intervals suggest no heterogeneity in TPRs or FPRs, so that the algorithm would go straight to two univariate analyses. An attempt to fit the bivariate model to these data results in a solution on the boundary of the parameter space with the ML estimate of $\rho = -1$. Interval estimates can be obtained from the MCMC iterations, though the traces show that convergence is slow and a large number of iterations are required for a stable estimate of the posterior densities (Figure 6).
15

**Figure 5:** Forest plot for Example 2 [22] ordered by FPR.

**Figure 6:** Posterior densities for variance parameters for Example 2 [22] calculated from 100,000 MCMC iterations. The dotted line shows the prior probability for the correlation.

The program reports the posterior probability of a positive $\rho$ as 0.38, but more importantly the posterior density shows that $\rho$ is poorly estimated. This is typical of results that are found when the estimates of one or more of $\left(\sigma_x, \sigma_y\right)$ have substantial
density close to zero. This leads to a posterior for the correlation with modes at +1 and -1, more extreme than the prior, due to the fact that the posteriors for the variances lie close to zero. Fitting two univariate meta-analyses suggests that two fixed effects are appropriate. The conclusion is that there is no evidence of between study variation in either the TPR or FPR for this meta-analysis.

**Example 3. Scheidler LAG data, algorithm path A, C, F, E or G.**

Examples 3 and 4 are taken from the meta-analysis of Scheidler et al [23] that have been extensively analysed in other studies [7-10]. Example 3 selects 17 studies using Lymphangiography (LAG) and Example 4 the 19 studies using computed tomography (CT). From the forest plot of example 4 (not shown) evidence for heterogeneity was not clear, so we proceed to fit the bivariate model. This gave a negative estimate of $\rho$ at -0.24, but the posterior density (Figure 7) indicates that it is poorly estimated with a 95% confidence limit covering almost all of its range. Separate meta-analyses of TPR and FPR indicate that a fixed effect model is appropriate for the FPR and a random effects model for the TPR.

![Figure 7: Example 3 posterior densities for variance parameters](image)

An SROC curve for these data would not be appropriate since, since the negative slope means it would not be valid. But we illustrate it in Figure 8 to compare it with...
Draft do not cite or quote without permission

the SROC curve derived from equation (3) presented in the analysis of Harbord et al [10] which lies much further from the shrunken estimates.

With no evidence of a threshold effect and little evidence of heterogeneity in the TPRs we would argue that it is inappropriate to summarise these data with an SROC curve. This leaves the user with the choice of presenting the bivariate results or summarising by two univariate meta-analyses. This choice would depend on the plausibility of a negative correlation from other knowledge of the methods. As we can see from Figure 7, the data themselves provide little evidence for this.

Figure 8: Fitted SROC curve and random effects for Example 3. Dots give raw rates and squares shrunken estimates of random effects. The solid line shows the SROC curve from equation (6) and the dotted line that from equation (3).

Example 4 Scheidler et al[23] CT data– algorithm path A, B, C, D
The forest plots (not shown) suggest significant heterogeneity, and evidence that TPR and FPR increase together. The fit to the bivariate model confirms that the heterogeneity for TPRs and FPRs but the ML estimates converge with a value of $\rho$ of 1.00, on the boundary of the parameter space, but with a 95% interval calculated from the MCMC results of (0.78 to 1.0). This implies that the final fit is equivalent to a fixed effects analysis with all the studies lying on a common SROC curve, as we see in Figure 9. This is an example of a degenerate fit that, nevertheless, is a valid model with an acceptable SROC summary. The interval estimates (Table 2) provide evidence of heterogeneity of both the FPRs and the TPRs and the posterior probability that $\rho$ is positive exceeds 0.99. For a fit on the boundary with $\rho = 1.0$ the same estimates would be obtained by fitting a fixed effect model that constrained the correlation to be 1.0 and the points all to lie on a common SROC curve. However, this would not provide any more information and might give interval estimates that are too narrow when we can see from Figure 9 that the data are compatible with a range of SROC curves with differing slopes. The summary results from the bivariate model with confidence intervals from the MCMC results (Table 2) are appropriate.
Figure 9: Fitted SROC curve and random effects for example 4. Dots give raw rates and squares shrunken estimates of random effects. The solid line shows the SROC curve from equation (6) and the dotted lines are a sample of 50 curves from the posterior density where the estimated slope lies within its 95% limits.

Table 2: Estimates of bivariate model for Example 4.

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>MCMC median</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>average TPR%</td>
<td>49.5</td>
<td>48.8</td>
<td>31.9 - 65.6</td>
</tr>
<tr>
<td>average FPR%</td>
<td>8.6</td>
<td>8.4</td>
<td>4.6 - 13.9</td>
</tr>
<tr>
<td>SD logit TPR</td>
<td>1.068</td>
<td>1.265</td>
<td>0.738 - 2.222</td>
</tr>
<tr>
<td>SD logit FPR</td>
<td>0.898</td>
<td>1.056</td>
<td>0.647 - 1.791</td>
</tr>
<tr>
<td>correlation</td>
<td>1.000</td>
<td>1.000</td>
<td>0.776 - 1.000</td>
</tr>
</tbody>
</table>

5 DISCUSSION

The summary of a particular meta-analysis by an SROC curve may fail either because the model is an inappropriate one or because there are insufficient data available to estimate the model parameters. We have identified two circumstances of the first
reason: there may be no threshold effect, identified by a correlation which is not positive, or there may be no heterogeneity in one or both of the true TPR and FPR values for the studies. Faced with the data from one meta-analysis it may be impossible to tell whether the model is appropriate or not because the information in the data may be insufficient to estimate the parameters of the model with any precision. Therefore there is a risk that the SROC curve analysis might look appropriate without actually being so, in addition to the circumstances where it is possible to see that the model has not worked. This risk can be averted by paying attention to the precision with which the SROC curve is estimated from the available data.

The precision of estimate variances and correlations from a meta-analysis depends on the number of studies and on the number of subjects contributing to the TPR and FPR estimates in each study. The sample size requirements for estimating variances and correlations when the true values are known are generally more exacting than those for means. A sample size of 30 is required to be 95% certain that a standard deviation of a normal distribution will be within 25% of its true value[24]. The sample size requirements for a correlation depend on its true value. For correlations of 0.9, 0.7, 0.5 and 0.3 the sample sizes required to be 95% certain that the one sided confidence interval for $\rho$ will be above zero are 8, 17, 39 and 116. But these sample sizes would only apply here if the TPRs and FPRs were estimated from very large samples, so they may be regarded as lower limits of the numbers required. In the SROC model it is the precision on a logit scale of the mean TPR and FPR which is important, and on this scale the standard error of the rates (for a fixed number of patients) increase when the rates are close to zero or one. An effective diagnostic test will often have a high TPR or a low FPR so that the uncertainty in this rate can
contribute to the uncertainty in the parameters of the SROC curve, even when the sample size for individual studies is quite large. Leeflang et al [25] found a median of 14 studies and a maximum of 40 in 30 reviews of diagnostic studies. The total number of studies in these 30 reviews fell from 476 to 70 or 72 when restricted to those satisfying the QUADAS quality criterion[26]. Thus it is likely that, at present, most diagnostic meta-analyses will be underpowered to estimate $\rho$ unless it is close to 1.0 and the individual TPRs and FPRs are well estimated.

Faced with data from a single meta-analysis a decision needs to be made as to how to present the results. It is inappropriate to use the bivariate model if there is insufficient information in the data to estimate it. But how is this to be judged? We suggest using the posterior probability that $\rho$ is positive, with a value of 0.95 as a decision criterion. However this might be lowered if we have any prior reason, such as knowledge of differing policies between centres, to expect that the test might be affected by a threshold. The generation of a set of SROC curves from their posterior density, as shown in Figure 9 can give a visual representation of the range of SROC curves that are compatible with the data. If there are insufficient data to support an SROC analysis then two univariate analyses meta-analyses should be presented. The analyst must choose either fixed or random effects method and guidance on this is available from interval estimates of the between study variances. A fixed effects meta-analysis is appropriate where the density plot is skewed towards zero.

Another situation that might occur is when the data provide evidence of a negative correlation between TPR and FPR (reaching point G on the algorithm). It is difficult to envisage what would lead to such a result and we have not found an example where
it applies. Negative estimates of the correlation are often obtained, but like Example 3 their interval estimates generally include zero.

The examples we have presented here have used meta-analyses where all the data are analysed as a single group. Other authors [7,9,23] have considered subgroups of the data defined by various characteristics of the studies in a meta-regression framework. The conclusions presented here should be equally valid in this context, and the R functions allow grouped data to be analysed where the grouping influences both the mean FPR and the mean TPR, and thus also the mean threshold and accuracy. It should be noted, however, that this analysis assumes that all subgroups share a common variance covariance and this may not always be the case (e.g. for the frequently analysed data of Scheidler et al [23], from which Examples 3 and 4 were taken). Thus the R functions also provide the option of analysing with individual variance covariance matrices.

6 CONCLUSIONS

In this paper we have explored the problems of interpreting the results of a diagnostic meta-analysis. The existing literature [7-10,18,19] has focussed mainly on the technical problems of obtaining fitted SROC curves. We focus on the interpretation of the results and the decision on how they should be presented. In some circumstances the data may show that the SROC model is inappropriate but, more commonly, the information in the data is insufficient to estimate it. As many diagnostic meta-analyses will be underpowered to estimate an SROC curve, it is important that the analyst pays attention to the uncertainty in the estimates and
understands that tools to summarise this in an accessible manner are available. The R functions we have developed are intended to make this possible.

REFERENCES


