SROC Curve

S. D. Walter
McMaster University, Hamilton, Ontario, Canada

Petra Macaskill
University of Sydney, NSW, Australia

INTRODUCTION

The summary receiver operating characteristic (SROC) curve has been recommended to represent the performance of a diagnostic test, based on data from a meta-analysis. Under a fixed-effect, logit-threshold model, the position of the SROC curve can be characterized in terms of the overall diagnostic odds ratio and the magnitude of interstudy heterogeneity. The Area Under the Curve (AUC) and an index Q^* are potentially useful summary measures. It is shown that AUC is maximized when the study odds ratios are homogeneous, and that it is quite robust to heterogeneity. An explicit upper bound is derived for AUC in the homogeneous situation, and a lower bound based on the limit case Q^*, defined by the point where sensitivity equals specificity: Q^* is invariant to heterogeneity. The standard error of AUC is derived for homogeneous studies, and is shown to be a reasonable approximation with heterogeneous studies. AUC and its standard error are easily computed in the homogeneous case, and avoid the need for numerical integration in the more general case. SE(AUC) and SE(Q^*) are numerically close, with SE(Q^*) being larger if the odds ratio is very large. Motivation for the use of the AUC and Q^* measures as summaries of the SROC curve are discussed.

A multilevel mixed model is also described, in which test accuracy and threshold are allowed to vary between studies. This provides a more general framework, within which the fixed effect model is a special case. The mixed model provides for the direct estimation of summary values of sensitivity, specificity, and likelihood ratios. Bayesian Markov Chain Monte Carlo (MCMC) methods are required to fit the mixed model, and appropriate software is becoming more available.

META-ANALYSIS OF DIAGNOSTIC TEST DATA

Systematic reviews of primary studies are becoming increasingly important for summarizing evidence about the accuracy of diagnostic tests. Guidelines for the conduct of such reviews[1–3] include defining the objectives, retrieval of the relevant literature, data extraction, meta-analytic methods for obtaining summary estimates of test accuracy, and investigating reasons for variation in test accuracy across studies.

The receiver operating characteristic (ROC) curve is well established as a method of summarizing the performance of a diagnostic test within a single study.[4–6] It indicates the relationship between the true positive rate (TPR) and the false positive rate (FPR) of the test, as the threshold used to distinguish disease cases from noncases varies. For instance, the threshold might be a defined level of serum cholesterol as a marker of cardiovascular disease, or a particular level of cellular abnormality as a marker of malignancy. The summary receiver operating characteristic (SROC) curve and the area under the curve (AUC) have been proposed as a way to describe diagnostic data in the context of a meta-analysis.[1,2,7–10]

A schematic ROC curve is shown in Fig. 1. All of its data points come from a single study, and are defined by the results arising when one of several alternative thresholds (or cut points) on the test results is used to discriminate between disease cases and noncases. Liberal thresholds that give high values of TPR will tend to incorrectly label a relatively high fraction of the noncases as cases, so the false positive rate (FPR) will be high. Conversely, conservative thresholds yield incorrect labels for noncases rather infrequently, but at the cost of detecting a smaller fraction of the true cases; in this situation, the value of TPR and FPR are both relatively low. Thus one generally expects TPR and FPR to be positively associated. In the ROC space, points near the lower-left corner correspond to conservative thresholds (low TPR and low FPR values), and points near the upper-right corner correspond to liberal thresholds (high TPR and high FPR values).

In a single study, changing the threshold necessarily results in monotonic changes in TPR and FPR. Accordingly, the ROC curve can always be empirically represented, in its simplest form by connecting the data points as shown in Fig. 1. Alternatively, smoothed curves

DOI: 10.1081/E-EBS 120024189

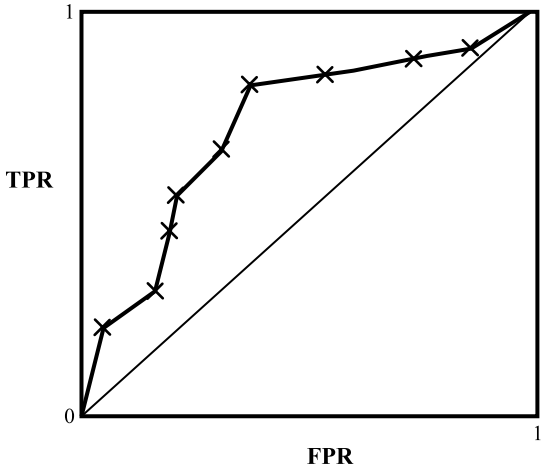

Fig. 1 Schematic ROC curve.

can also be fitted using a latent variable approach, based on a model that the case and noncase test values follow normal or logistic distributions.[11–19]

In a meta-analysis, the units of analysis are separate studies. In the simplest case, each study contributes an estimate of TPR and FPR. The SROC curve represents the relationship between TPR and FPR across studies, recognizing they may have used different thresholds. In contrast to the ROC analysis, the set of (FPR, TPR) points does not necessarily yield a unique, monotonic curve. Several methods have been proposed for fitting an SROC curve;[7,10,20,21] this paper will focus on the properties of the popular method developed by Moses et al.[10] An alternative but more complex method proposed by Rutter and Gatsonis[20,21] is also outlined and its potential advantages are discussed.

LOGIT-THRESHOLD FIXED EFFECT MODEL FOR SUMMARY RECEIVER OPERATING CHARACTERISTIC CURVE

The most popular method to obtain a smoothed fit of the SROC curve is to use the regression model proposed by Moses et al.[10] The dependent and independent variables are

$$D = \ln\left(\frac{\text{TPR}}{1-\text{TPR}}\right) - \ln\left(\frac{\text{FPR}}{1-\text{FPR}}\right) \quad (1)$$

and

$$S = \ln\left(\frac{\text{TPR}}{1-\text{TPR}}\right) + \ln\left(\frac{\text{FPR}}{1-\text{FPR}}\right) \quad (2)$$

respectively. D is equivalent to the diagnostic log odds ratio, ln(OR), which conveys the test's accuracy in discriminating cases from noncases. S is a function of the diagnostic threshold in a study, with high values corresponding to liberal inclusion criteria for cases. $S=0$ when $\text{TPR}=1-\text{FPR}$, i.e., on the antidiagonal from the top-left to bottom-right corners of the SROC space. The regression equation

$$D = a + bS \quad (3)$$

can be fitted by standard least squares methods, assuming that D is approximately normally distributed for a given value of S. Optionally, weights can be employed to reflect interstudy heterogeneity with respect to the sample variance of D, or a robust fitting procedure may also be used.[10] The coefficient b in Eq. 3 represents the dependence of the test accuracy on threshold. If $b \approx 0$, then the studies will be referred to as *homogeneous*; they can then be summarized by an overall OR, noting that $a=\ln(\text{OR})$. If $b \neq 0$, then the studies are referred to as *heterogeneous* with respect to OR. In this case, a can be thought of as the value of ln(OR) when $S=0$.

This model assumes logistic distributions for the test values in the cases and noncases, but normal distributions may also constitute an adequate approximation. However, the model is generally fitted without directly appealing to the logistic distribution assumption. If the underlying test results actually have logistic distributions, then S for a given study can be expressed as a function of the cut point that gives rise to the sensitivity and specificity for that study.[10] If the variances of the distributions of test results for the cases and noncases are equal, then the SROC is symmetric about the diagonal line where $\text{TPR}=1-\text{FPR}$. If the variances are unequal, the resulting SROC will be asymmetric. No assumptions are made about the distribution of S. Covariates may be added to the model to assess whether test accuracy varies systematically with other study-related factors.[22,23]

If the estimate of test accuracy for each study is weighted by the inverse of its variance, one is assuming that sampling error is the only source of variability between studies. However, this assumption may be unrealistic, given that studies are likely to vary in a number of respects, including the spectrum of disease, conditions under which the test was administered, and other study and patient characteristics that could affect test accuracy.[24–26]

Although the parameters a and b are both fixed because they are assumed to be constant across studies, applying equal weights to studies (i.e., fitting an unweighted regression) has been empirically shown to give results that are consistent with assuming a random effect.[8,10,20] This is because both the within- and between-study variances are taken into account, thereby giving relatively more weight to smaller studies, as would occur in a random effect model. Irwig et al.[8] recommend the

unweighted analysis, because weights based on sample variances may give too much emphasis to inaccurate studies, and hence give a biased SROC curve. Moses et al.[10] also recommend the unweighted approach.

Once the regression has been fitted, one can reverse the transformations (Eqs. 1 and 2) and hence deduce the relationship between TPR and FPR as

$$\mathrm{TPR} = \frac{\exp\left(\frac{a}{1-b}\right)\left(\frac{\mathrm{FPR}}{1-\mathrm{FPR}}\right)^{(1+b)/(1-b)}}{1+\exp\left(\frac{a}{1-b}\right)\left(\frac{\mathrm{FPR}}{1-\mathrm{FPR}}\right)^{(1+b)/(1-b)}} \tag{4}$$

Expression 4 gives TPR at any given value of FPR, and hence defines the entire SROC curve. While there may be interest in identifying particular points on the curve, it is also useful to have an overall summary measure of the curve's behavior. One appropriate measure is the Area Under the Curve (AUC), which can be calculated as

$$\mathrm{AUC} = \int_0^1 \frac{\exp\left(\frac{a}{1-b}\right)\left(\frac{x}{1-x}\right)^{(1+b)/(1-b)}}{1+\exp\left(\frac{a}{1-b}\right)\left(\frac{x}{1-x}\right)^{(1+b)/(1-b)}}\,dx \tag{5}$$

In general, AUC must be numerically obtained, because there is no closed form for expression 5.

Adoption of a summary index is helpful for succinct reporting of a given data set, especially when limited data preclude the reliable identification of particular points on the curve. The AUC measure is widely used in ROC analysis, where it can be interpreted as the probability that the test values for a random pair of diseased and nondiseased individuals would be correctly ranked; it also represents the (unweighted) average of TPR over all possible values of FPR.

The AUC is also a natural candidate summary for an SROC analysis. We will consider the alternative index Q^*, which will be defined as a point of indifference on the SROC curve, where the probabilities of an incorrect test result are equal for disease cases and noncases. The partial AUC has also been proposed as a summary measure, this being the area under some restricted portion of the curve corresponding to FPR values of clinical interest, or in which study data are located.

EMPIRICAL BEHAVIOR OF THE SUMMARY RECEIVER OPERATING CHARACTERISTIC CURVE

Figure 2 shows a set of three symmetric SROC curves with $b=0$, which occurs when the studies are homogeneous and thus exhibit no relationship between OR and threshold (S). The curves were obtained by numerical evaluation of Eq. 4 for $a=1.5$, 2, and 3, which correspond to OR=4.5, 7.4, and 20.1, respectively. As a increases, the SROC curve moves closer to its ideal position near the upper-left corner. If $a\to\infty$, then $AUC\to 1$, which would indicate a perfect test having 100% sensitivity and specificity, and no errors in distinguishing cases from noncases. In contrast, if $a\approx 0$ (or OR$\approx$1), the curve lies close to the diagonal TPR=FPR in the SROC space; then, AUC=1/2 and the test performs no better than chance.

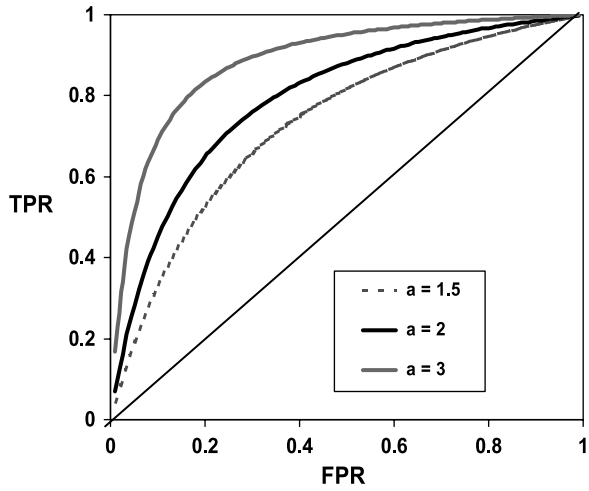

Fig. 2 Summary receiver operating characteristic with various values of a ($b=0$). (*View this art in color at www.dekker.com.*)

For completeness, we mention situations where $a<0$. These correspond to OR<1, when the test discriminates cases and noncases in the ''wrong'' direction, and *worse* than at random. As OR$\to$0, AUC$\to$0, and the curve lies close to the lower-right corner of the SROC space. Such situations are unlikely to occur in practice.

We now consider the case of heterogeneous studies under model 3, where diagnostic accuracy depends on threshold ($b\neq 0$). Fig. 3 shows three SROC curves derived from Eq. 4, all with $a=2$ but different values of b. Nonzero values of b give asymmetric curves. If $b>0$, the curve initially rises less steeply than the symmetric curve (with $b=0$), but then it rises more steeply and crosses the symmetric curve to achieve relatively high values of TPR for high values of FPR. The SROC curves with $b<0$ exhibit the opposite behavior—see, for example, the curve with $b=-0.5$ in Fig. 3.

Interestingly, as suggested by Fig. 3 and as shown by Moses et al.,[10] the family of curves defined by a fixed value of a all pass through a common point, located on the antidiagonal, where TPR=1$-$FPR, or sensitivity equals specificity. That point has coordinates

$$\mathrm{TPR} = \frac{\exp(a/2)}{1+\exp(a/2)} \tag{6}$$

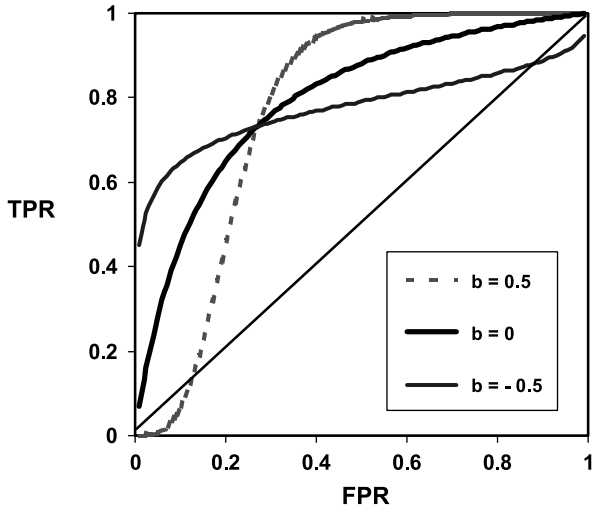

Fig. 3 Summary receiver operating characteristic with various values of b ($a=2$). (*View this art in color at www.dekker.com.*)

and

$$\text{FPR} = \frac{1}{1 + \exp(a/2)} \tag{7}$$

Moses et al.[10] denote the value of TPR at this point by Q^*. In ROC analysis, Q^* has been suggested as a further summary measure. It corresponds to the point closest to the ideal top-left corner of the SROC space in symmetric curves.

The fact that all SROC curves with a given value of a pass through the same Q^* point means that Q^* conveys no additional statistical information beyond the odds ratio. However, use of Q^* can be motivated by its simple interpretability, given that Q^* is the point on the SROC curve where TPR$=1-$FPR; thus it represents the diagnostic threshold at which the probability of a correct diagnosis is constant for all subjects. Also, as will be shown later, the existence of a common value of Q^* permits the derivation of a useful lower bound for AUC.

As a consequence of assuming model 3, when $b \neq 0$, the SROC curve has a region where TPR<FPR, which lies below the main diagonal. This region may be seen, for example, near the lower-left corner in the curve for $b=0.5$ in Fig. 3. In this region, the test would theoretically be performing worse than at random. However, in practice, this region is very small. Even if relatively strong heterogeneity (with $b>0$) is present, the ''improper'' part of the SROC curve (where TPR<FPR) includes very low values of TPR, and these values correspond to test thresholds that are unlikely to be acceptable in clinical practice. On the other hand, if heterogeneity with $b<0$ occurs, the improper part of the curve corresponds to very high values of FPR, which would again have no clinical relevance. See Ref. [27] for numerical details of this effect.

We now briefly discuss the behavior of model 3 for extreme values of b. As $b \to 1$, the fitted SROC curve becomes progressively steeper, and in the limit case at $b=1$ it degenerates to a vertical line. This could occur, for example, if there was no variation in FPR between studies. However, the line still passes through the common Q^* point defined by Eqs. 6 and 7. Once $b>1$, the curve inverts and shows a negative relationship of TPR to FPR. This is implausible in practice, except perhaps by chance in small samples.

Similar behavior is seen if b is near or below -1. Near $b=-1$, the curve is horizontal, indicating no relationship of TPR to FPR. This could occur if there was no variation in TPR between studies. If $b<-1$, the curve again inverts (to a different shape) and suggests an implausible negative relationship between TPR and FPR. Despite this, all the curves for $|b|>1$ possess the common value of Q^* given by Eqs. 6 and 7. In practice, data yielding $|b|>1$ are unlikely. Empirical experience suggests that practical meta-analysis data sets often have b close to 0, and are rarely larger than 0.5 in absolute value.

SUMMARY MEASURES: AREA UNDER THE CURVE AND Q^*

The AUC is a popular index of the overall performance of a test.[4–6,18,19,28] As mentioned earlier, AUC ranges from 1 for a perfect test that always correctly diagnoses, to 0 for a test that never correctly diagnoses. In single studies, AUC can be interpreted as the probability that the test will correctly rank a randomly chosen case/noncase pair with respect to their test values.[29] The AUC is intended to fulfill the same function in meta-analyses, and in effect one assumes that the SROC curve accurately conveys test performance at the individual subject level.

Although numerical integration is required in general to obtain AUC, some special cases are of interest because they yield exact analytic expressions and comparative results, as we now discuss. As expected, AUC increases with a, for fixed b. By examining Eq. 5, one can prove that for a given value of a, AUC is maximized when $b=0$, implying that AUC is optimally large in homogeneous studies.[27] Furthermore, one may show that AUC is symmetric in b, so that negative values of b yield the same value of AUC as the equivalent positive value, this despite the very different shapes of their associated SROC curves. If $a=b=0$, then from Eq. 5

$$\text{AUC} = \int_0^1 x\,dx = \frac{1}{2}$$

Table 1 AUC_{hom}, Q^*, and their difference for various values of the diagnostic odds ratio: homogeneous case

Odds ratio	AUC_{hom}	Q^*	$AUC_{hom}-Q^*$
0.5	0.386	0.414	−0.028
1	0.500	0.500	0.000
1.5	0.567	0.551	0.017
2	0.614	0.586	0.028
3	0.676	0.634	0.042
4	0.717	0.667	0.051
5	0.747	0.691	0.056
10	0.827	0.760	0.067
20	0.887	0.817	0.069
30	0.913	0.846	0.068
40	0.929	0.863	0.065
50	0.939	0.876	0.063

By symmetry, it is evident that AUC=1/2 when $a=0$ even if $b\neq 0$, so AUC=1/2 indicates random overall performance for any set of studies.

In the homogeneous case $b=0$, the general expression 5 becomes

$$\mathrm{AUC} = \int_0^1 \frac{\exp(a)\left(\frac{x}{1-x}\right)}{1+\exp(a)\left(\frac{x}{1-x}\right)}\,dx$$

In this case, we can obtain an exact solution

$$\mathrm{AUC}_{\mathrm{hom}} = \frac{\mathrm{OR}}{(\mathrm{OR}-1)^2}[(\mathrm{OR}-1)-\ln(\mathrm{OR})] \tag{8}$$

where AUC_{hom} indicates the AUC for homogeneous studies, and OR=exp(a). If $a=0$ (or OR=1), then the special value $AUC_{hom}=1/2$ should be used in place of Eq. 8, which is then degenerate. Expression 8 can be used to evaluate AUC for homogeneous studies, without the need for numerical integration. As demonstrated later, expression 8 is also a useful upper bound and close approximation for AUC in heterogeneous studies. For reference, Table 1 shows the value of AUC, Q^*, and their difference for a range of values of OR in the homogeneous case.

By noting that AUC declines with increasing b, and that the limit curve with $b\rightarrow 1$ passes through the common Q^* point, from Eqs. 6 and 7, we may deduce that a lower bound for AUC in curves with a given value of a is

$$Q^* = \frac{\exp(a/2)}{1+\exp(a/2)} = \frac{\sqrt{\mathrm{OR}}}{1+\sqrt{\mathrm{OR}}} \tag{9}$$

which is equivalent to the TPR value given in Eq. 6. Also, Q^* from Eq. 9 and the maximum value AUC_{hom} from Eq. 8 provide easily computable lower and upper bounds, respectively, for AUC with any given value of $a>0$.

Numerical Tabulations

Numerical evaluations of expressions 8 and 9 demonstrate that in the homogeneous case, the difference between AUC_{hom} and Q^* increases for moderate values of OR, but is never more than about 7%.[27] The maximum difference occurs at $a=2.85$ (or OR=17.3), which value is the solution to a transcendental equation. For larger values of a, the difference declines very slowly, with a limit value $AUC_{hom}-Q^*=0$ at $a=\infty$.

The numerical integration of Eq. 5 for the heterogeneous case allows one to assess the impact of heterogeneity on the value of AUC. For a fixed value of OR, AUC declines slowly as b increases from 0 (homogeneous studies) to larger values (increasing heterogeneity). However, the dependence of AUC on b is weak, and the dominant effect on AUC is the value of OR (or a). For $|b|<0.4$, the percentage change in AUC compared to the homogeneous case is less than 2%. Accordingly, AUC_{hom} provides a good approximation to AUC even in heterogeneous studies.[27]

STANDARD ERRORS OF AREA UNDER THE CURVE AND Q^*

We first consider the sample variation in $\widehat{\mathrm{AUC}}$. From Eq. 5, we see that AUC is a function of the regression parameters a and b, and hence the variability in $\widehat{\mathrm{AUC}}$ is a function of the sample variation in $\hat{a}$ and $\hat{b}$. Using the delta method, an approximate variance for $\widehat{\mathrm{AUC}}$ is

$$\begin{aligned}\mathrm{var}(\widehat{\mathrm{AUC}}) = {} & \left(\frac{\partial \mathrm{AUC}}{\partial a}\right)^2 \mathrm{var}(\hat{a}) \\ & + \left(\frac{\partial \mathrm{AUC}}{\partial b}\right)^2 \mathrm{var}(\hat{b}) + 2\left(\frac{\partial \mathrm{AUC}}{\partial a}\right) \\ & \times \left(\frac{\partial \mathrm{AUC}}{\partial b}\right)\mathrm{cov}(\hat{a},\hat{b})\end{aligned} \tag{10}$$

where, from Eq. 5

$$\begin{aligned}\frac{\partial \mathrm{AUC}}{\partial a} = {} & \left(\frac{1}{1-b}\right)\exp\left(\frac{a}{1-b}\right) \\ & \times \int_0^1 \frac{\left(\frac{x}{1-x}\right)^p}{\left[1+\left(\frac{x}{1-x}\right)^p \exp\left(\frac{a}{1-b}\right)\right]^2}\,dx\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial \mathrm{AUC}}{\partial b} = {} & \left(\frac{1}{1-b}\right)^2\exp\left(\frac{a}{1-b}\right) \\ & \times \int_0^1 \frac{\left(\frac{x}{1-x}\right)^p\left[a+2\ln\left(\frac{x}{1-x}\right)\right]}{\left[1+\left(\frac{x}{1-x}\right)^p \exp\left(\frac{a}{1-b}\right)\right]^2}\,dx\end{aligned} \tag{12}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and $p=(1+b)/(1-b)$. The variances and covariance of $\hat{a}$ and $\hat{b}$ can be directly obtained from the standard regression software used to fit model 3 to the data.

In general, evaluation of var($\widehat{\text{AUC}}$) again requires numerical integration. However, for the special case of homogeneous studies where $b=0$, an approximate, large-sample expression is possible. Using the delta method, we may then obtain

$$\text{SE}(\widehat{\text{AUC}}_{\text{hom}}) = \frac{\text{OR}}{(\text{OR}-1)^3}[(\text{OR}+1)\ln\text{OR} - 2(\text{OR}-1)]\text{SE}(\hat{a}) \quad (13)$$

Eq. 13 implies SE($\widehat{\text{AUC}}_{\text{hom}}$) is symmetric in ln(OR), although we would usually be concerned with values OR>1. When OR=1, expression 13 is degenerate, but by using L'Hôpital's rule, one can show that in the neighborhood of OR=1,

$$\text{SE}(\widehat{\text{AUC}}) \approx \text{SE}(\hat{a})/6 \quad (14)$$

The delta method also yields an approximate standard error for $\hat{Q}^*$ as

$$\text{SE}(\hat{Q}^*) = \frac{\sqrt{\text{OR}}}{2(\sqrt{\text{OR}}+1)^2}\text{SE}(\hat{a}) \quad (15)$$

Moses et al.[10] give this result in an alternative form involving cosh(a/4). Numerical evaluations show that the standard errors of Q^* and AUC are both maximized when OR=1, so the least precise situation for either index is for tests that have close to random performance. In the region of OR=1, SE($\widehat{\text{AUC}}$)≈1/6 and SE($\hat{Q}^*$)≈1/8. Comparisons show SE($\widehat{\text{AUC}}$)>SE($\hat{Q}^*$) for OR values between 1 and 17.3, the same value associated with the values of AUC and Q^*; for OR>17.3, AUC has the smaller standard error. Both standard errors approach 0 as $a\to\infty$.

For heterogeneous studies, SE($\widehat{\text{AUC}}$) is not necessarily maximized when $a=0$, because it involves var($\hat{b}$) and cov($\hat{a}$, $\hat{b}$) as well as var($\hat{a}$). However, it is the last of these terms that dominates, with the other two making relatively small numerical contributions. Thus in practice SE($\widehat{\text{AUC}}$) is maximized approximately when OR=1, even for heterogeneous studies. For most values of OR, the standard error declines by up to about 10% at the most extreme level of b.

OTHER ISSUES WITH THE FIXED-EFFECT MODEL

So far, we have established some basic properties of the SROC curve under model 3, and we found that the value AUC_{hom} associated with homogeneous studies is a reasonable upper-bound approximation even for the general AUC with heterogeneous studies. The corresponding standard error also provides a good approximation for heterogeneous studies, and is conservatively large except in some cases of extreme heterogeneity. In the presence of strong heterogeneity, the AUC would be an inadequate summary of the data anyway, and it would then be preferable to examine the SROC curve in more detail, including specific TPR values for given FPR.

Q^* also provides an easily computed lower bound for AUC, but empirically appears to be not quite as good an approximation as AUC_{hom}. The motivation for Q^* as an index in its own right is that it is located where the SROC curve crosses the antidiagonal from (0,1) to (1,0) of the SROC space. Hence TPR=1−FPR at Q^*, and so the probability of an incorrect result from the test is the same for cases and noncases. Q^* is therefore a point of ''indifference'' between false-positive and false-negative diagnostic errors. In homogeneous studies, Q^* is the point on the SROC curve lying closest to the optimal upper-left corner, but this is not true with heterogeneous studies.

Use of Q^* as the summary measure assumes implicitly that false-negative and false-positive test results are of equal value. In practice, there may be different costs associated with these two types of error: One wishes to minimize false-positive results because of the additional testing required to establish the correct diagnosis (noncase), and because the additional tests tend to be more costly, invasive, or risky. On the other hand, false-negative results lead to disease cases being missed, with possible deterioration in their prognosis. In general, one must weigh the false-positive and false-negative errors to balance the overall performance of the test in a population; the optimal diagnostic threshold does not then correspond in general to the Q^* point.

One can motivate the use of AUC as an index that represents the probability that the test will correctly rank a case/noncase pair of subjects.[4,29] It can also be thought of as the average TPR over the entire range of FPR values. Because it summarizes the whole SROC curve, AUC has a symmetric interpretation with respect to either TPR or FPR. The AUC is affected by the whole SROC curve, including regions with limited or no data, or by sectors corresponding to TPR and FPR values that are unlikely to occur in practice. Accordingly, it has been suggested that partial SROC curves be adopted, by limiting attention to those portions of the SROC curve of clinical interest, or where data are actually observed.[30–33]

There are some unresolved issues on the use of partial SROC curves and the corresponding partial AUC. First, the partial AUC may be thought of as the average TPR within a restricted range of FPR, but not vice versa. Hence the partial AUC has an asymmetric interpretation with respect to TPR and FPR; on the other hand, the complete

AUC enjoys a symmetric interpretation with respect to both types of test error. Second, there may be some arbitrariness in which regions of the curve to select. One might examine the SROC curve within a prespecified range of TPR values, for instance by defining a maximum acceptable level for clinical practice. Another approach would be to choose the region according to the observed range of data in the meta-analysis; the choice itself would then be affected by sampling variation, which might be substantial in meta-analyses involving small studies. Furthermore, a reasonable choice for one test may be unreasonable for another test, thus complicating their comparison.

The analytic methods for AUC presented in this paper can be extended to cover the partial AUC and its standard error. We may expect the effect of interstudy heterogeneity to be greater for the partial AUC than the weak effect seen for the full AUC. For instance, if attention is limited to values of FPR<0.2 when $a=2$ (Fig. 3), the corresponding partial AUC will be greater when $b>0$ than when $b<0$. Hence the partial AUC lacks symmetry with respect to b. On the other hand, the complete AUC has some compensating decreases in contributions to the area at higher values of FPR, so there are only modest changes in the total AUC as b varies (Fig. 3). We may also recall that AUC is symmetric with respect to b, so that the same summary value will be obtained for a given strength of dependence of diagnostic accuracy on the test threshold. The partial AUC does not possess these properties, and hence it will show greater dependence on the degree of interstudy heterogeneity. Further work is needed to explore the properties of the partial AUC in more detail. Similar investigation is needed for other summary indices, such as the ASC (Area Swept out by the Curve), PLC (Projected Length of the Curve),[34] and the Gini/Lorenz coefficients.[35]

Sampling variability and dependence between test accuracy and test threshold are unlikely to fully account for the observed heterogeneity in test accuracy between studies. Additional sources of heterogeneity can be explored by examining the association between test accuracy and study level covariate information.[3,8,36] However, given the limited data on patient and study design characteristics that are typically available, it is unlikely that study level variables will fully explain the ''excess'' variability in test accuracy.[26] Inappropriately assuming a fixed effect can lead to spurious precision and result in covariates being incorrectly identified as significantly associated with test accuracy.[26] A mixed model that includes random effects for test accuracy can take into account unexplained variability between studies. This approach is used in the hierarchical (mixed) model outlined briefly below, which, unlike the Moses approach, directly models the TPR and FPR for each study.

HIERARCHICAL (MIXED) EFFECTS MODEL

An alternative approach to fitting SROC curves has been proposed by Rutter and Gatsonis.[20,21] As with the Moses model, each study (i) contributes an estimate of TPR and FPR to the meta-analysis. For each study, the number in the diseased group who have a positive test result (y_{i1}) and the number in the nondiseased group who have a positive test result (y_{i2}) are both assumed to follow a binomial distribution such that $y_{ij} \sim \beta(n_{ij}, \pi_{ij})$, where π_{ij} represents the probability of a positive test for group j ($j=1,2$) in study i, and n_{ij} represents the total number of positive and negative test results in group j. The model is based on an ordinal logistic regression model[37] and takes the form $\text{logit}(\pi_{ij})=(\theta_i+\alpha_i \text{dis}_{ij})\exp(-\beta\ \text{dis}_{ij})$ where dis_{ij} represents the true disease status (coded as -0.5 for the nondiseased and 0.5 for the diseased). The model parameters estimate the implicit threshold (θ_i) for a positive test result in study i and the diagnostic accuracy (α_i) for study i. The parameter β allows for an association between test accuracy and test threshold. The estimated SROC is symmetric when $\beta=0$. This hierarchical (multilevel) model takes into account both the within- and between-study variability.

The within-study variability is modeled at the first level. For the ith study, $\text{logit}(\text{TPR}_i)$ and $\text{logit}(\text{FPR}_i)$ are modeled to estimate the implicit threshold (θ_i) and diagnostic accuracy (α_i) for that study. Hence random effects are assumed for both threshold and accuracy as these may vary across studies. When the SROC is symmetric (i.e., $\beta=0$), the level 1 analysis for each study reduces to a ordinary logistic regression model where α_i is estimated by $\text{logit}(\text{TPR}_i)-\text{logit}(\text{FPR}_i)$ (which is equivalent to the observed value D_i in study i for the dependent variable in model 3) and θ_i is estimated by $(\text{logit}(\text{TPR}_i)+\text{logit}(\text{FPR}_i))/2$ (which is equivalent to $S_i/2$, where S_i is the observed value in study i for the independent variable in model 3). The sampling variability for both TPR_i and FPR_i is taken into account.

The variability between studies is modeled at level 2. The random effects for test threshold and accuracy are assumed to be independent (uncorrelated) and normally distributed with $\theta_i \sim N(\Theta, \tau_\theta^2)$ and $\alpha_i \sim N(\Lambda, \tau_\alpha^2)$. The hyperparameters Θ and Λ represent the expected threshold and accuracy, respectively, across all studies in the analysis. If $\tau_\theta^2=0$ and $\tau_\alpha^2=0$, the model reduces to a fixed-effect model. Because each study contributes only one point in ROC space to the analysis, a single study does not provide information on the shape of the SROC. Hence the shape parameter (β) is assumed to be a fixed effect, which is estimated from the study points jointly considered. The hierarchical model is fitted using Markov Chain Monte Carlo (MCMC) Bayesian methods. This requires a third level to specify prior distributions for all model parameters.

A summary ROC curve can be constructed by choosing a range of values of FPR and using the estimated model parameters to compute the predicted values for sensitivity. The expected TPR at a chosen FPR value is given by $\mathrm{TPR(FPR)} = 1/(1+\exp[-\{\Lambda\exp(-0.5\beta)+\mathrm{logit(FPR)}\exp(-\beta)\}])$.

The expected operating point on the curve is estimated using $E(\mathrm{TPR}) = 1/(1+\exp[-\{(\Theta+\Lambda/2)\exp(-\beta/2)\}])$ and $E(\mathrm{FPR}) = 1/(1+\exp[-\{(\Theta-\Lambda/2)\exp(-\beta/2)\}])$. Covariates can be added to the model to assess whether threshold, accuracy, and/or the shape of the SROC vary with patient or study characteristics. Such terms are generally fitted as fixed effects, but could also be included as random effects.

Advantages of the Hierarchical Model

The Rutter and Gatsonis method provides a general framework for the meta-analysis of diagnostic test performance. The Moses method does not directly model the TPR and FPR values for each study. Because D and S are functions of both TPR and FPR, the parameters for the Moses model cannot be used to obtain the expected operating point on the SROC, the corresponding likelihood ratios at that point, or the standard errors of these estimates. The hierarchical model can be used to model variation in test threshold as well as test accuracy through the inclusion of study level covariates, whereas the Moses approach can only be used to model variation in test accuracy. However, the Moses model does have the advantage that no assumption is made about the distribution of S. Lastly, the hierarchical model incorporates both the within- and between-study variability, and takes account of unexplained heterogeneity between studies through the inclusion of random effects for test threshold and accuracy.

Fitting the Model

Despite the potential advantages of the hierarchical model, it has not been widely adopted. This is most likely because of the necessity to use Markov Chain Monte Carlo (MCMC) methods to fit it.[20] Rutter and Gatsonis[21] have demonstrated how the model can be fitted using BUGS (Bayesian inference Using Gibbs Sampling) software. However, they comment that the process involves MCMC simulation, and is still relatively complex. The recent availability of software for fitting nonlinear mixed models in SAS[38] provides an alternative and potentially more straightforward approach that does not require specification of prior distributions. PROC NLMIXED in SAS allows for a nonlinear function of the model parameters, and for a non-normal error distribution, but the random effects are restricted to be normally distributed. Summary estimates of TPR, FPR, likelihood ratios can be computed and their asymptotic confidence intervals estimated using the delta method. The distribution of the random effects may be checked by examining a histogram and normal probability plot of their Empirical Bayes (EB) estimates. This follows the same approach adopted for checking distributional assumptions for linear mixed models.[39,40]

CONCLUSION

The mixed model provides a rigorous method of analysis that takes proper account of the sources of variability in test performance between studies. The resulting standard errors (and confidence intervals) reflect the unexplained heterogeneity that is likely to be present in many meta-analyses. The mixed model also allows estimation of clinically relevant indicators of test performance such as the expected TPR, FPR and likelihood ratios. A range of alternative fixed-effect and random effects methods are available for obtaining these summary measures.[22] However, they can lead to results that are potentially inconsistent within the same meta-analysis. For instance, for the same data, summary estimates of TPR and FPR could be computed that are not consistent with summary estimates of likelihood ratios or an SROC estimated using the Moses model. A useful feature of the mixed model is that other approaches can be regarded as special cases.

The complexity of the Bayesian (MCMC) method for fitting the mixed effects model is likely to discourage its use. The NLMIXED procedure is more accessible, but at present this software can only deal with two levels in the response variable. For meta-analyses where studies contribute more than one (FPR, TPR) pair to the analysis, an additional level would be required. A Bayesian approach potentially provides greater flexibility for fitting the mixed effects model in that it allows the meta-analyst to specify alternative distributions for the random effects and also allows prior information to be taken into account. However, in practice, a normal distribution is often assumed for the random effects and noninformative priors are commonly used. Likelihood-based methods for estimating mixed model parameters are also consistent with current approaches used in the meta-analysis of clinical trials.[39,41,42] Model checks are required to assess the adequacy of the distributional assumption for the random effects, but assessing normality will clearly be difficult in small meta-analyses.[39]

The Moses fixed-effect SROC model has the advantage of simplicity, and it makes no assumptions about the distribution of S. It is useful as a preliminary step before

fitting the mixed model. Major differences in the variables found to be associated with test accuracy or the shape of the SROC would warrant investigation as this may be indicative of convergence problems.

An issue that potentially affects all of the methods discussed here is that reference test (or gold standard) has been assumed to be error-free. In fact, the reference standard is often itself subject to measurement error, as illustrated, for example, in the observable differences between pathologists and radiologists in their assessments of the same sample material. Several methods have been proposed to correct for referent errors, but they primarily pertain to data from single studies. Rather little has been done on this problem in the context of combining studies in a meta-analysis, but one proposed approach[43] has been to use a latent class framework to extend the logit-threshold model (Eq. 3), and to recognize the fact that both the candidate and referent tests are potentially subject to error. The true disease state of an individual cannot be directly observed, but an estimate of the probability of an individual having disease can be obtained from the latent class model. This then permits a deattenuation of the errors in the data, and the resulting adjustment to the SROC curve tends to lead to an improved estimate of test performance.

Further development of the SROC methods are required to allow comparison of SROC curves when not all component studies in a meta-analyses are independent of one another. For example, some of the component studies may make direct comparisons of two tests while other studies only evaluate one. Methods to take such dependencies into account have been proposed for therapeutic studies,[44] and extensions to diagnostic test comparisons would be useful.

The techniques discussed here apply to situations where one has only summary measures (TPR and FPR) from each study. Sometimes one has access to the individual level data in each study; one would then be able to carry out a multilevel analysis, taking both inter- and intrastudy variation into account, as well as the effects of subject-specific covariates. However, in practice, current reporting of diagnostic studies and meta-analyses is methodologically poor, and this level of detail in the data would often be difficult to achieve.[1,36,45,46] Accordingly, further work to understand the properties of the SROC curve based on summary data from each study seems warranted.

REFERENCES

1. Irwig, L.; Tosteson, A.N.A.; Gatsonis, G.; Lau, J.; Colditz, G.; Chalmers, T.C.; Mosteller, F. Guidelines for meta-analyses evaluating diagnostic tests. Ann. Intern. Med. **1994**, *120* (8), 667–676.
2. Vamvakas, E.C. Meta-analyses of studies of the diagnostic accuracy of laboratory tests. Arch. Pathol. Lab. Med. **1998**, *122* (8), 675–686.
3. Deville, W.L.; Buntinx, F. Guidelines for Conducting Systematic Reviews of Studies Evaluating the Accuracy of Diagnostic Tests. In *The Evidence Base of Clinical Diagnosis*; Knottnerus, J.A., Ed.; BMJ Books: London, 2002; 145–165.
4. Hanley, J.A. Receiver Operating Characteristic (ROC) Curves. In *Encyclopedia of Biostatistics*; Armitage, P., Colton, T., Eds.; Wiley: Chichester, 1998; 3738–3745.
5. Beck, J.R.; Shultz, E.K. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch. Pathol. Lab. Med. **1986**, *110* (1), 13–19.
6. Metz, C.E.; Herman, B.A.; Shen, J.H. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat. Med. **1998**, *17* (9), 1033–1053.
7. Kardaun, J.W.P.F.; Kardaun, O.J.W.F. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf. Med. **1990**, *29* (1), 12–22.
8. Irwig, L.; Macaskill, P.; Glasziou, P.; Fahey, M. Meta-analytic methods for diagnostic test accuracy. J. Clin. Epidemiol. **1993**, *48* (1), 119–130.
9. Midgette, A.S.; Stukel, T.A.; Littenberg, B. A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic-summary point estimates. Med. Decis. Mak. **1993**, *13* (3), 253–256.
10. Moses, L.E.; Shapiro, D.E.; Littenberg, B. Combining independent studies of a diagnostic tests into a summary ROC curve: Data-analytic approaches and some additional considerations. Stat. Med. **1993**, *12* (14), 1293–1316.
11. Dorfman, D.D.; Alf, E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—Rating method data. J. Math. Psychol. **1969**, *6* (3), 487–496.
12. Hanley, J.A. The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests. Stat. Med. **1996**, *15* (14), 1575–1585.
13. McCullagh, P. Regression models for ordinal data (with discussion). J. R. Stat. Soc., Ser. B **1980**, *42* (2), 109–142.
14. Ma, G.; Hall, W.J. Confidence bands for receiver operating characteristics curves. Med. Decis. Mak. **1993**, *13* (3), 191–197.
15. Green, D.M.; Swets, J. *Signal Detection Theory and Psychophysics*; Wiley: New York, 1966.
16. Hanley, J.A. Receiver operating characteristic (ROC) methodology: The state of the art. Crit. Rev. Diagn. Imaging **1989**, *29* (3), 307–335.
17. Ogilvie, J.C.; Creelman, C.D. Maximum likelihood estimation of ROC curve parameters. J. Math. Psychol. **1968**, *5* (3), 377–391.
18. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **1982**, *143* (1), 29–36.

19. Hilden, J. The area under the ROC curve and its competitors. Med. Decis. Mak. **1991**, *11* (2), 95–101.
20. Rutter, C.M.; Gatsonis, G. Regression methods for meta-analysis of diagnostic test data. Acad. Radiol. **1995**, *2* (Suppl. 1), S48–S56.
21. Rutter, C.M.; Gatsonis, G. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat. Med. **2001**, *20* (19), 2865–2884.
22. Deeks, J.J. Systematic Reviews of Evaluations of Diagnostic and Screening Tests. In *Systematic Reviews in Health Care: Meta-Analysis in Context*; Egger, M., Davey-Smith, G., Altman, D.G., Eds.; BMJ Publishing Group: London, 2001.
23. Fahey, M.T.; Irwig, L.; Macaskill, P. Meta-analysis of Pap smear accuracy. Am. J. Epidemiol. **1995**, *141* (7), 680–689.
24. Irwig, L.; Bossuyt, P.; Glasziou, P.; Gatsonis, G.; Lijmer, J. Designing Studies to Ensure that Estimates of Test Accuracy Will Travel. In *Designing Studies on Diagnostic Tests*; Knottnerus, A., Ed.; BMJ Books: London, 2001.
25. Ransohoff, D.F.; Feinstein, A.R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N. Engl. J. Med. **1978**, *299* (17), 926–930.
26. Lijmer, J.G.; Bossuyt, P.M.M.; Heisterkamp, S.H. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat. Med. **2002**, *21* (11), 1525–1537.
27. Walter, S.D. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat. Med. **2002**, *21* (9), 1237–1256.
28. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver-operating characteristic curves: A non-parametric approach. Biometrics **1988**, *44* (3), 837–845.
29. Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating graph. J. Math. Psychol. **1975**, *12* (4), 387–415.
30. McClish, D.K. Analyzing a portion of the ROC curve. Med. Decis. Mak. **1989**, *9* (3), 190–195.
31. Thompson, M.L.; Zucchini, W. On the statistical analysis of ROC curves. Stat. Med. **1989**, *8* (10), 1277–1290.
32. Wieand, S.; Gail, M.H.; James, B.R. A family of nonparametric statistics for comparing diagnostic tests with paired or unpaired data. Biometrika **1988**, *76* (3), 585–592.
33. Jiang, Y.; Metz, C.E.; Nishikawa, R.M. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology **1996**, *201* (3), 745–750.
34. Lee, W.C.; Hsiao, C.K. Alternate summary indices for the receiver operating characteristic curve. Epidemiology **1996**, *7* (6), 605–611.
35. Lee, W.C. Probabilistic analysis of global performances of diagnostic tests: Interpreting the Lorenz curve-based summary measures. Stat. Med. **1999**, *18* (4), 455–471.
36. Lijmer, J.G.; Mol, B.W.; Heisterkamp, S. Empirical evidence of design related bias in studies of diagnostic tests. JAMA **1999**, *282* (11), 1061–1066.
37. McCullagh, P.; Nelder, J.A. *Generalized Linear Models,* 2nd Ed.; Chapman and Hall: London, 1989.
38. SAS Institute. *SAS/STAT User's Guide, Version 8*; SAS Institute: Cary, 1999.
39. Turner, R.M.; Omar, R.Z.; Yang, M.; Goldstein, H.; Thompson, S.G. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat. Med. **2000**, *19* (24), 3417–3432.
40. Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data*; Springer: New York, 2000.
41. Normand, S.T. Meta-analysis: Formulating, evaluating, combining, and reporting. Stat. Med. **1999**, *18* (3), 321–359.
42. Van Houwelingen, H.C.; Arends, L.R.; Stijnen, T. Advanced methods in meta-analysis: Multivariate approach and meta-regression. Stat. Med. **2002**, *21* (4), 589–624.
43. Walter, S.D.; Irwig, L.; Glasziou, P.P. Meta-analysis of diagnostic tests with imperfect reference standards. J. Clin. Epidemiol. **1999**, *52* (10), 943–951.
44. Bucher, H.C.; Guyatt, G.H.; Walter, S.D.; Griffith, L. The results of direct and indirect treatment comparisons in meta-analysis of randomized clinical trials. J. Clin. Epidemiol. **1997**, *50* (6), 683–691.
45. Walter, S.D.; Jadad, A.R. Meta-analysis of screening data: A survey of the literature. Stat. Med. **1999**, *18* (24), 3409–3424.
46. Reid, M.C.; Lachs, S.; Feinstein, A.R. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA **1995**, *274* (8), 645–651.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016