STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:1237–1256 (DOI: 10.1002/sim.1099)

Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data

S. D. Walter*,†

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

SUMMARY

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. However, little is known about the basic properties of the SROC curve or its estimate. In this paper, the position of the SROC curve is characterized in terms of the overall diagnostic odds ratio and the magnitude of inter-study heterogeneity in the odds ratio. The area under the curve (AUC) and an index Q^* are discussed as potentially useful summaries of the curve. It is shown that AUC is maximized when the study odds ratios are homogeneous, and that it is quite robust to heterogeneity. An upper bound is derived for AUC based on an exact analytic expression for 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 shown to be a reasonable approximation with heterogeneous studies. The expressions for 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 found to be numerically close, with SE(Q^*) being larger if the odds ratio is very large. The methods are illustrated using data for the Pap smear screening test for cervical cancer, and for three tests for the diagnosis of metastases in cervical cancer patients. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: ROC curve; meta-analysis; diagnosis; area under the curve (AUC)

1. INTRODUCTION

The receiver operating curve (ROC) is well established as a method of summarizing the performance of a diagnostic test [1–3]. It indicates the relationship between the true positive rate (TPR) and the false positive rate (FPR) of the test at various thresholds used to distinguish disease cases from non-cases. More recently the summary receiver operating characteristic (SROC) curve and the area under the curve (AUC) have been proposed as a way to assess

* Correspondence to: Stephen D. Walter, Department of Clinical Epidemiology & Biostatistics, McMaster University Health Sciences Centre, 2C16, 1200 Main Street West, Hamilton, Ontario L8N 325, Canada
† E-mail: walter@mcmaster.ca

Contract/grant sponsor: Canadian Institutes of Health Research
Contract/grant sponsor: NSERC, Canada

Received March 2000
Accepted July 2001

diagnostic data in the context of a meta-analysis [4–8]. While it is regarded as a potentially useful summary of the data, little is known about the basic properties of the SROC curve and the AUC or their estimates.

In this paper we will examine various properties of the SROC curve, by characterizing its position as a function of the overall diagnostic odds ratio (OR) between the test result and the disease state, and of the degree of inter-study heterogeneity. In Section 2 we consider the general form for the SROC curve and its AUC. In Section 3 we evaluate the empirical behaviour of the SROC curve. In Section 4 we evaluate AUC empirically for the general case, and derive explicit expressions for certain special cases. We also consider an index Q^*, calculated at the point on the SROC curve where sensitivity and specificity are equal. In Section 5, standard errors for AUC and Q^* are evaluated, with a derivation of analytic expressions for homogeneous studies. Two practical examples are described in Section 6, based on a screening test for cervical cancer and three diagnostic tests for metastases in cervical cancer patients. Further points of discussion are considered in Section 7, including consideration of some strengths and weaknesses of AUC, Q^*, and other summary measures of the SROC curve.

2. ROC AND SROC CURVES

An ROC curve is shown schematically in Figure 1. Each data point comes from a single study in which several alternative diagnostic thresholds (or cutpoints) are used to discriminate between disease cases and non-cases. The true positive rate (TPR, or sensitivity) is high if a liberal threshold is adopted, while a conservative threshold gives a low TPR. Higher TPR values are associated with higher false positive rates (FPR), and vice versa. Points near the lower-left corner of the ROC plot correspond to conservative thresholds, and points near the upper-right corner correspond to liberal thresholds.

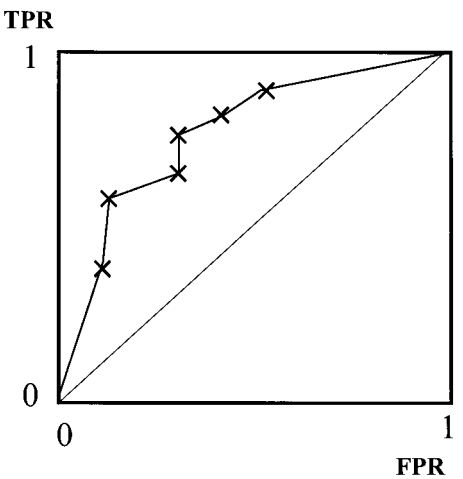

Figure 1. ROC plot (schematic).

In a single study, changing the threshold necessarily results in monotonic changes in TPR and FPR. Accordingly, the ROC curve can always be fitted empirically, by connecting the data points as shown in Figure 1. Alternatively, smoothed curves can also be fitted using a latent variable approach, often incorporating the binormal or bilogistic model [9–15].

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 is intended to represent 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 need not necessarily yield a unique, monotonic curve.

Smoothed fitting of the SROC curve can be achieved by using a regression model proposed by Moses *et al.* [8]. The dependent and independent variables in the regression are

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

and

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

respectively. D is equivalent to the diagnostic log-odds ratio, ln(OR), which conveys the test's accuracy in discriminating cases from non-cases. S can be interpreted as a measure of the diagnostic threshold, with high values corresponding to liberal inclusion criteria for cases. $S = 0$ when $\mathrm{TPR} = 1 - \mathrm{FPR}$, that is, on the anti-diagonal from the top-left to bottom-right corners of the SROC space.

The regression equation

$$D = a + bS \tag{3}$$

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 inter-study heterogeneity with respect to the sample variance of D. The coefficient b represents the dependence of the test accuracy on threshold. If $b \approx 0$, then the studies are homogeneous and can be summarized by an overall OR noting that $a = \ln(\mathrm{OR})$. If $b \neq 0$, then the studies are heterogeneous with respect to OR. In this case, a can be thought of as the value of ln(OR) when $S = 0$.

Once the regression has been fitted, one can reverse the transformations (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 the entire SROC curve. While there may be interest in identifying particular points on the SROC curve, it is often useful to have an overall summary measure of the curve's behaviour. One appropriate measure is the area under the curve (AUC), which can be calculated as

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

In general, AUC must be obtained numerically, because there is no closed form for the integral in (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. AUC is widely used in ROC analysis, where it can be interpreted as the probability that the diagnostic test values for a random pair of diseased and non-diseased individuals would be correctly ranked; it also represents the (unweighted) average of TPR over all possible values of FPR.

AUC is also a natural candidate summary for an SROC analysis, given that the fitted SROC curve passes through the points $(0,0)$ and $(1,1)$ in the SROC space. We will also consider the 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 non-cases. Other summary measures that have been proposed include the partial AUC, being the area under some restricted portion of the curve corresponding to FPR values of clinical interest, or in which study data are located. Further comments on the strengths and weaknesses of these and other measures are given in the Discussion.

In the next section, we first consider the general empirical behaviour of the SROC curve, together with some special cases for which exact analytical results are derived.

3. EMPIRICAL BEHAVIOUR OF SROC

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. The curves were obtained by numerical evaluation of (4); the values of a are 1.5, 2 and 3, which correspond to $\mathrm{OR}=4.5, 7.4$ 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 $\mathrm{AUC}\to 1$; this would indicate a perfect test, with 100 per cent sensitivity and specificity, and no errors in distinguishing cases from non-cases. In contrast, if $a\approx 0$ (or $\mathrm{OR}\approx 1$), the curve is close to

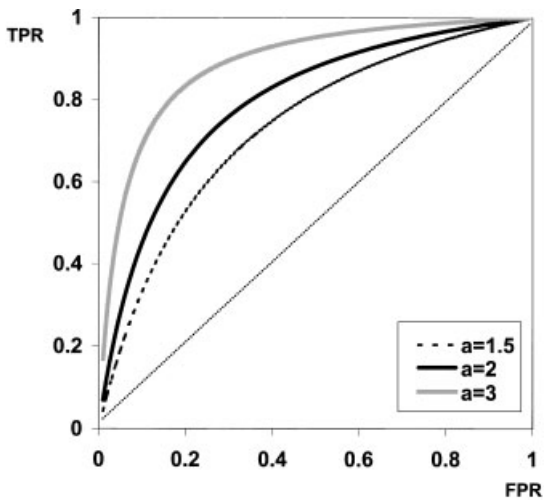

Figure 2. SROC curves with various values of $a, b=0$.

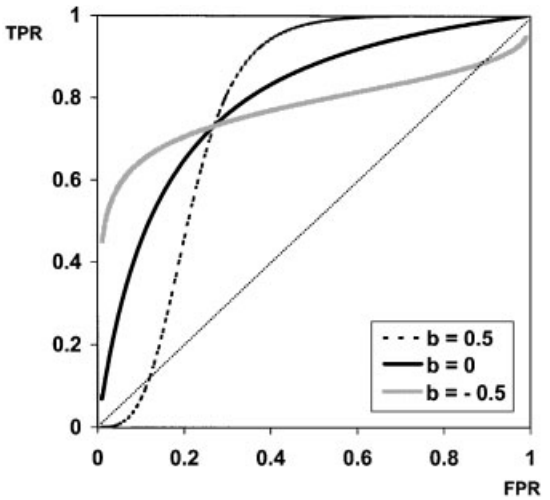

Figure 3. SROC with various values of b ($a = 2$).

the dotted diagonal line TPR = FPR; then AUC = $\frac{1}{2}$ and the test would perform no better than chance.

For completeness, we must mention situations where $a < 0$. These correspond to OR < 1, when the test discriminates cases and non-cases in the 'wrong' direction, and worse than at random. As OR $\to 0$, AUC $\to 0$, with a curve that is 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, where OR is not constant ($b \neq 0$). Figure 3 shows a set of three SROC curves all with $a = 2$ but different values of b. Non-zero values of b give asymmetric SROC 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 behaviour – see, for example, the curve with $b = -0.5$ in Figure 3.

Interestingly, as suggested by Figure 3 and as shown by Moses *et al.* [8], the family of curves defined by a fixed value of a all pass through a common point, located on the anti-diagonal where TPR = 1 − FPR, or sensitivity equals specificity. That point has co-ordinates

$$\text{TPR} = \frac{\exp(a/2)}{1+\exp(a/2)} \quad \text{and} \quad \text{FPR} = \frac{1}{1+\exp(a/2)} \tag{6}$$

Moses *et al.* [8] denote the value of TPR at this point by Q^*, a notation that we now adopt here. In ROC analysis, Q^* has been suggested as a single number summarization of the test performance, being the point closest to the ideal top-left corner of the SROC space for symmetric curves.

The fact that all SROC curves with a given value of a pass through the same Q^* point means that it conveys no additional statistical information beyond the odds ratio. However, use of the Q^* index can be motivated by its easier interpretability for some users, given that Q^* is the point on the SROC where TPR = 1 − FPR; thus it represents the diagnostic threshold at which the probability of a correct diagnosis is constant for all subjects. Furthermore, the

existence of a common value of Q^* permits the derivation of a useful lower bound for AUC, as shown in Section 4.

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, in the curve for $b = 0.5$ in Figure 3, near the lower-left corner. In this region, the test would be predicted to be performing worse than at random. Assuming that $b \geqslant 0$, it can be shown that the point $(\mathrm{FPR}', \mathrm{TPR}')$ where the SROC curve crosses the diagonal has

$$\mathrm{FPR}' = \mathrm{TPR}' = \frac{\exp(-a/2b)}{1 + \exp(-a/2b)} \tag{7}$$

(If $b < 0$, there is a symmetrically opposite point in the top-right corner of the SROC space.) In practice, the anomalous region will usually be rather small. For instance, if $a = 2$ (corresponding to OR = 7.4) and $b = 0.2$ (relatively strong heterogeneity), then from (7) $\mathrm{TPR}' = 0.7$ per cent. Even if $a = 2$ and b takes the more extreme value 0.4, TPR' is still low, at 7.6 per cent. Diagnostic tests would not usually be used at such low values of TPR (or sensitivity), so in practice the 'improper' part of the curve where TPR < FPR is negligible.

3.1. *Extreme values of b*

As $b \to 1$, the 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. Note that the line still passes through the common Q^* point defined by (6). Once $b > 1$, the curve 'inverts', and shows a negative relationship of TPR to FPR. This is an implausible situation in practice except perhaps by chance in small samples.

Similar behaviour 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 were no variation in TPR between studies. If $b < -1$, the curve again becomes inverted and suggests the implausible negative relationship between TPR and FPR. Despite this, all the curves for $|b| > 1$ possess the common value of Q^* given by (6).

In practice, data yielding $|b| > 1$ are unlikely. See the examples in Section 6 for further discussion of this point.

4. AREA UNDER THE CURVE (AUC) AND Q^*

The AUC is a useful and popular index of the overall performance of a test [1–3, 16–18]. As mentioned earlier, AUC ranges from 1 for a perfect test that correctly classifies all cases and non-cases of disease, to 0 for a test which never diagnoses correctly. In single studies, AUC can be interpreted as the probability that the test will correctly rank a randomly chosen case/non-case pair with respect to their test values [19]. In meta-analyses, the AUC is intended to fulfil the same function, so effectively one assumes that the SROC curve conveys an appropriate measure of test performance at the individual subject level, and that there is no bias associated with the ecologic aggregation of data to the study level. Further issues around this point are explored further in the Discussion section. 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. Also, by examining equation (5), we can prove that for a given value of a, AUC is maximized when $b=0$ – see the Appendix (result 1). This result implies that AUC is optimally large in homogeneous studies. Furthermore, we 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 (see Appendix, result 2). This is true despite the very different shapes of their associated SROC curves. If $a=b=0$, then from (5)

$$\mathrm{AUC}=\int_0^1 x\,\mathrm{d}x=\frac{1}{2}$$

By symmetry it is evident that $\mathrm{AUC}=\frac{1}{2}$ when $a=0$ even if $b\neq 0$, so $\mathrm{AUC}=\frac{1}{2}$ indicates random overall performance ($\mathrm{OR}=1$) for any set of studies.

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

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

As shown in the Appendix (result 3), we can obtain an exact solution in this case

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

where $\mathrm{AUC}_{\mathrm{hom}}$ indicates the AUC for homogeneous studies, and $\mathrm{OR}=\exp(a)$. If $a=0$ (or $\mathrm{OR}=1$), then the special value $\mathrm{AUC}_{\mathrm{hom}}=\frac{1}{2}$ should be used in place of (8), which is then degenerate. In general, expression (8) can be used to evaluate AUC for homogeneous studies, by adopting one of the usual estimates of OR, and without the need for numerical integration.

Although only valid for homogeneous studies, it turns out that expression (8) is a useful upper bound for AUC in heterogeneous studies, as will now be demonstrated. Furthermore, numerical evaluations to be shown in Section 4.1 indicate that $\mathrm{AUC}_{\mathrm{hom}}$ also provides a good approximation for AUC in heterogeneous studies.

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

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

Note that (9) is equivalent to the TPR value given earlier (expression (6)). By using Q^* from (9) and the maximum value $\mathrm{AUC}_{\mathrm{hom}}$ from (8) we have easily computable lower and upper bounds (respectively) for AUC with a given value of $a>0$. This argument assumes $|b|<1$; as noted earlier, curves with $|b|>1$ are not of practical interest.

4.1. Numerical tabulations

Table I shows $\mathrm{AUC}_{\mathrm{hom}}$, Q^* and $\mathrm{AUC}_{\mathrm{hom}}-Q^*$, for various values of OR in the homogeneous case when $b=0$. The arithmetic difference between $\mathrm{AUC}_{\mathrm{hom}}$ and Q^* increases for moderate values of OR, but is never more than 7 per cent. The maximum difference occurs at a value of a which is the solution to a transcendental equation (A7) described in the Appendix (result 4); at that point, $a=2.85$ or $\mathrm{OR}=17.3$. For larger values of a, the difference declines very slowly, with a limit value $\mathrm{AUC}_{\mathrm{hom}}-Q^*=0$ at $a=\infty$.

Table I. AUC, Q^* and their difference for various values of the diagnostic odds ratio: homogeneous case.

Odds ratio	AUC	Q^*	AUC-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

Table II. AUC for various values of the diagnostic odds ratio (OR)* and the regression slope b: heterogeneous case.

b	OR = 2	OR = 5	OR = 10	OR = 20
0	0.614 (0.0%)†	0.747 (0.0%)	0.827 (0.0%)	0.887 (0.0%)
0.2	0.612 (−0.2%)	0.744 (−0.4%)	0.824 (−0.4%)	0.883 (−0.4%)
0.4	0.608 (−0.9%)	0.736 (−1.4%)	0.814 (−1.6%)	0.874 (−1.5%)
0.6	0.602 (−1.9%)	0.724 (−3.1%)	0.799 (−3.4%)	0.858 (−3.2%)
0.8	0.594 (−3.2%)	0.708 (−5.2%)	0.78 (−5.6%)	0.839 (−5.4%)

*OR refers to diagnostic odds ratio for the homogeneous case ($b = 0$).
†Figures in parentheses give percentage difference in AUC compared to when $b = 0$.

Table II shows the behaviour of AUC for the heterogeneous case when $b \neq 0$. These values were computed by numerical integration of (5), for various values of a and b. For convenience, results are expressed in terms of OR, calculated from $a = \ln(\text{OR})$. Table II also shows the percentage difference in AUC compared to its value in the homogeneous case. For fixed 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 is the value of OR (or a). For $|b| < 0.4$ (as was found in the empirical data discussed later), the percentage change in AUC compared to the homogeneous case is less than 2 per cent. Accordingly, it appears that AUC_{hom} provides a good approximation to AUC even in heterogeneous studies.

5. STANDARD ERRORS OF AUC AND Q^*

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

$\hat{AUC}$ is

$$\mathrm{var}(\hat{\mathrm{AUC}}) = \left(\frac{\partial \mathrm{AUC}}{\partial a}\right)^2 \mathrm{var}(\hat{a}) + \left(\frac{\partial \mathrm{AUC}}{\partial b}\right)^2 \mathrm{var}(\hat{b}) + 2\left(\frac{\partial \mathrm{AUC}}{\partial a}\right)\left(\frac{\partial \mathrm{AUC}}{\partial b}\right) \mathrm{cov}(\hat{a}, \hat{b}) \quad (10)$$

where, from (5)

$$\frac{\partial \mathrm{AUC}}{\partial a} = \left(\frac{1}{1-b}\right) \exp\left(\frac{a}{1-b}\right) \int_0^1 \frac{(\frac{x}{1-x})^p}{[1 + (\frac{x}{1-x})^p \exp(\frac{a}{1-b})]^2}\, \mathrm{d}x \quad (11)$$

and

$$\frac{\partial \mathrm{AUC}}{\partial b} = \left(\frac{1}{1-b}\right)^2 \exp\left(\frac{a}{1-b}\right) \int_0^1 \frac{(\frac{x}{1-x})^p [a + 2\ln(\frac{x}{1-x})]}{[1 + (\frac{x}{1-x})^p \exp(\frac{a}{1-b})]^2}\, \mathrm{d}x \quad (12)$$

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

In general, evaluation of $\mathrm{var}(\hat{\mathrm{AUC}})$ requires numerical integration to deal with the partial derivatives (11) and (12). However, for the special case of homogeneous studies where $b = 0$, an approximate, large sample expression is possible. As derived in the Appendix (result 5) by using the delta method, we may obtain

$$\mathrm{SE}(\hat{\mathrm{AUC}}_{\mathrm{hom}}) = \frac{1}{(\mathrm{OR}-1)^3}[(\mathrm{OR}+1)\ln \mathrm{OR} - 2(\mathrm{OR}-1)]\mathrm{SE}(\hat{\mathrm{OR}})$$

or equivalently, recalling that $a = \ln(\mathrm{OR})$

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

Note that (13) implies $\mathrm{SE}(\hat{\mathrm{AUC}}_{\mathrm{hom}})$ is symmetric in $\ln(\mathrm{OR})$, although we would usually be concerned with values $\mathrm{OR} > 1$. In the special case $\mathrm{OR} = 1$, expression (13) is degenerate, but by using L'Hôpital's rule one can show that in the neighbourhood of $\mathrm{OR} = 1$

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

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

$$\mathrm{SE}(\hat{Q}^*) = \frac{1}{2\sqrt{\mathrm{OR}}(\sqrt{\mathrm{OR}}+1)^2}\mathrm{SE}(\hat{\mathrm{OR}})$$

or

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

(Note that Moses *et al.* [8] give this result in an alternative form involving $\cosh(a/4)$.) If $\mathrm{OR} \approx 1$, then

$$\mathrm{SE}(\hat{Q}^*) \approx \mathrm{SE}(\hat{a})/8 \quad (16)$$

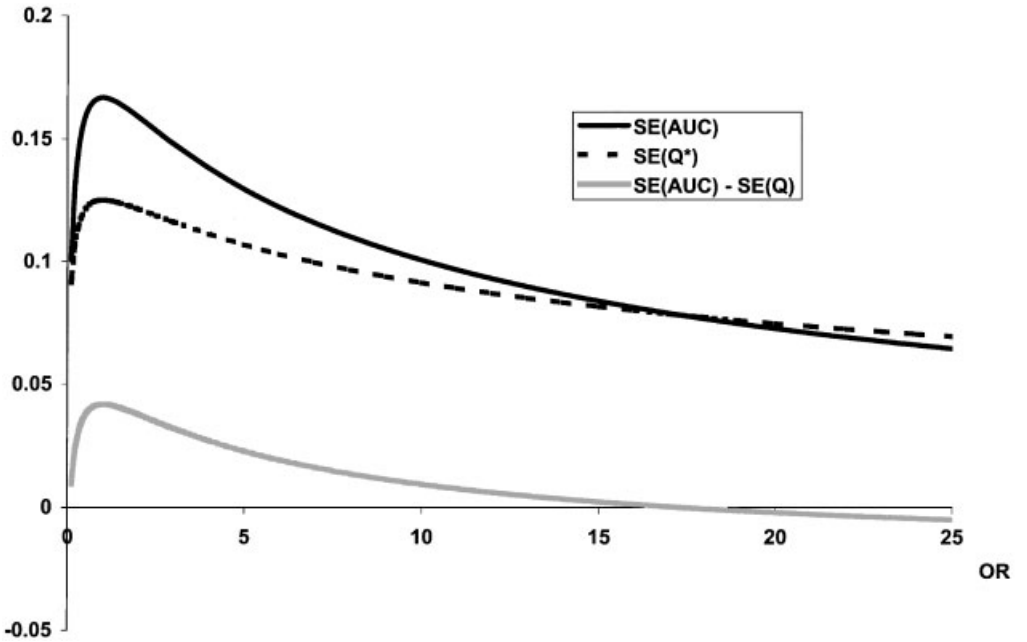

Figure 4. Comparison of SE(AUC) and SE(Q^*).

Figure 4 shows SE($\hat{\text{AUC}}_{\text{hom}}$), SE($\hat{Q}^*$) and SE($\hat{\text{AUC}}_{\text{hom}}$) − SE($\hat{Q}^*$) as a function of OR; for convenience of making the comparison, the value of SE($\hat{a}$) is here taken to be 1. SE($\hat{\text{AUC}}_{\text{hom}}$) and SE($\hat{Q}^*$) are both maximized when OR = 1($a = 0$), so the worst situation for estimating either index is for diagnostic tests which have close to random performance. In that situation, and consistent with equations (14) and (16), AUC has the larger standard error. Note also that in the region of OR = 1, SE($\hat{\text{AUC}}$) ≈ 1/6 and SE($\hat{Q}^*$) ≈ 1/8, consistent with (14) and (16).

SE($\hat{\text{AUC}}$) > SE($\hat{Q}^*$) for OR values between 1 and 17.3, the solution to the same equation (A7) encountered with AUC and Q^* themselves – see Appendix (result 6) for details; for OR > 17.3, AUC has the smaller standard error. Both standard errors approach 0 as $a \to \infty$.

For heterogeneous studies, SE($\hat{\text{AUC}}$) is a function of var($\hat{a}$), var($\hat{b}$) and cov($\hat{a}, \hat{b}$) (see (10)), and it is not necessarily maximized when $a = 0$. In practice, however, it is var($\hat{a}$) which dominates, with var($\hat{b}$) and cov($\hat{a}, \hat{b}$) making relatively small numerical contributions. Thus SE($\hat{\text{AUC}}$) appears to be approximately maximized in practice when OR = 1, even for heterogeneous studies. See Section 6 for further illustration of this point in the context of empirical data.

The approximate standard errors presented here can be used to formulate confidence intervals and significance tests, based on an assumption of normality for the parameter estimates.

6. EXAMPLES

To illustrate the methods, we first consider data from a meta-analysis of the Pap smear screening test for cervical cancer [20]. There were 59 studies with a mean total sample size

Table III. Results of SROC analysis for Pap smear screening test for cervical cancer.

	Unweighted analysis		Weighted analysis	
Intercept (a)	1.58	(0.13)*	1.73	(0.09)
Slope (b)	0.003	(0.06)	−0.04	(0.07)
AUC	0.7432	(0.0176)	0.7625	(0.0115)
AUC_{hom}	0.7432	(0.0176)	0.7626	(0.0116)
Q^*	0.6878	(0.0144)	0.7039	(0.0097)

* Estimate (SE).

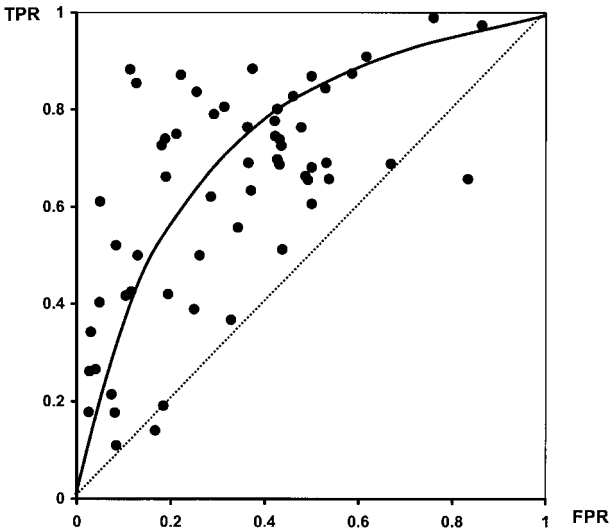

Figure 5. SROC curve for a meta-analysis of Pap smear screening for cervical cancer.

of 295 women. Unweighted and inverse-variance weighted regressions were calculated, using Moses's model, after the addition of $\frac{1}{2}$ to each cell in the 2×2 table for each study [8, 21]. Table III shows the regression results and corresponding summaries of the SROC curve. The AUC values were calculated numerically from (5), AUC_{hom} from (8), and Q^* from (9); corresponding standard errors were obtained from (10), (13) and (15). Figure 5 shows the study data points and the SROC curve from the unweighted analysis. Irwig *et al.* [4] recommend the unweighted analysis, because use of weights based on sample variances may give too much weight to studies with poor accuracy, and hence give a biased SROC curve. Moses *et al.* [8] also recommend the unweighted approach.

The results in this example show that the slope b is very close to zero, so that the SROC curve is essentially symmetric. Consequently the values of AUC and AUC_{hom} are very similar. As expected, Q^* is slightly smaller than AUC. The diagnostic odds ratio is $\exp(1.58) = 4.85$ (unweighted analysis), a value considerably less than the critical value of 17.3 identified in the Appendix, so Q^* has a smaller standard error than does AUC. In this example, the use of weighting slightly increased the intercept a, and the standard errors of the SROC curve

Table IV. Results of SROC analyses for three diagnostic tests for cervical cancer metastases.

Test		Unweighted analysis		Weighted analysis	
LAG test	Intercept (a)	2.09	(0.38)*	1.70	(0.30)
	Slope (b)	−0.35	(0.25)	−0.28	(0.25)
	AUC	0.7948	(0.0494)	0.7537	(0.0435)
	AUC_{hom}	0.8044	(0.0414)	0.7592	(0.0374)
	Q^*	0.7397	(0.0364)	0.7011	(0.0312)
CT test	Intercept (a)	2.79	(0.36)	2.61	(0.33)
	Slope (b)	0.22	(0.12)	0.14	(0.11)
	AUC	0.8667	(0.0258)	0.8537	(0.0267)
	AUC_{hom}	0.8708	(0.0291)	0.8554	(0.0290)
	Q^*	0.8012	(0.0287)	0.7863	(0.0277)
MR test	Intercept (a)	3.51	(0.61)	3.23	(0.68)
	Slope (b)	0.25	(0.19)	0.04	(0.23)
	AUC	0.9140	(0.0282)	0.9023	(0.0422)
	AUC_{hom}	0.9192	(0.0334)	0.9025	(0.0435)
	Q^*	0.8525	(0.0383)	0.8339	(0.0469)

*Estimate (SE).
LAG, lymphangiography; CT, computed tomography; MR, magnetic resonance;
AUC, area under the curve; SE, standard error.

summary measures became somewhat smaller. However, these are not generalizable findings, as will be seen in the second example.

For the second example, we examine data from a meta-analysis by Scheidler *et al.* [22] of three tests for the diagnosis of lymph node metastases in cervical cancer patients. The detection of metastases is helpful in determining the preferred therapy – patients without metastases are better candidates for surgery, whereas patients with metastases would usually be given radiotherapy. Lymphangiography was historically the diagnostic method of choice, and it is based on the detection of nodal filling defects. The computed tomography and magnetic resonance imaging methods are based on detection of nodal enlargements. Scheidler's meta-analysis was intended to compare these tests and assess their clinical utility.

We will use the data on eligible patients as given in Tables I–III of Scheidler *et al.* [22]. Meta-analyses on the three tests – lymphangiography, computed tomography and magnetic resonance imaging – were based on 17, 19 and 10 studies, respectively; mean total study sample sizes were 82, 54 and 84. The percentage of patients who were cases varied by study between 15–51 per cent, 10–60 per cent and 15–43 per cent, and FPR ranged from 1–53 per cent, 0–30 per cent and 0–16 per cent.

Table IV shows the results for each of the three tests. The a values are approximately in the range 2 to 3, and correspond to OR values of about 7 to 20. The most heterogeneous case is when $b = -0.35$ in the unweighted regression for lymphangiography, although even this value was not significantly different from zero. Figure 6 shows the corresponding SROC curves from the unweighted analysis; note that for simplicity of presentation, the actual data points for the three tests have not been plotted. In contrast to the first example, only a limited range of FPR values actually occurred, as noted above. Consequently the reader should be aware that the plotted SROC curves extend beyond the empirical range of the data.

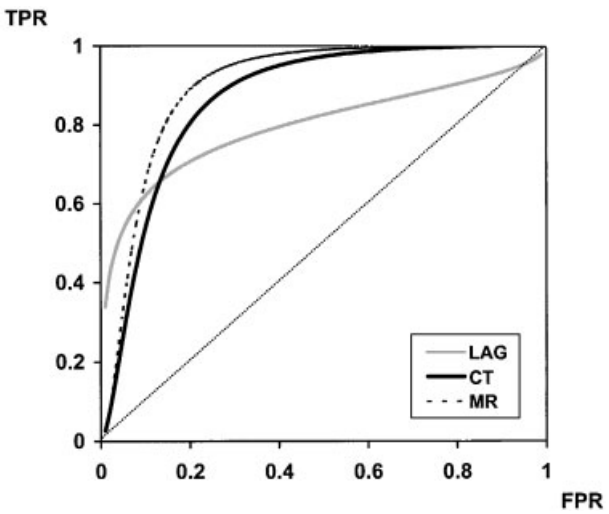

Figure 6. SROC curves for three tests for cervical cancer metastases: unweighted fit.

Weighting had rather little effect in these data, and similar values of AUC, Q^* and their standard errors pertain with or without weighting. Note that AUC_{hom} provides a closer approximation to AUC than does Q^*, with less than 1 per cent error in all cases. Also, $SE(Q^*) < SE(AUC_{hom})$ for lymphangiography and computed tomography which have $OR = 8.1$ and 16.1, respectively. However, for the magnetic resonance test, $SE(AUC_{hom})$ is smaller, because its OR is large; from the results in Table IV, $a = 3.51$ in the unweighted analysis, corresponding to $OR = 33.4$, a particularly large value which exceeds the threshold value 17.3 (Appendix equation (A7)).

In order to explore the effects of the OR value and heterogeneity on precision, we used expression (10) with various values of a and b, but keeping $\text{var}(\hat{a})$, $\text{var}(\hat{b})$ and $\text{cov}(\hat{a}, \hat{b})$ fixed at their observed values, based on the results from the unweighted analysis of the lymphangiography test as an example. We can thus examine the effects of different diagnostic ORs and different degrees of heterogeneity, but with a constant level of variation and covariation in the regression parameters. The values of OR range from 1 to 20, and b from 0 to 0.5; these cover most of the situations likely to be encountered in practice.

Table V shows $SE(\hat{AUC})$ and its percentage change compared to the case of homogeneous studies with a given value of a. As anticipated from the theoretical results, it is evident that SE(AUC) depends primarily on the value of OR (or a), and only weakly on b. For most values of OR, the standard error declines by up to about 20 per cent at the most extreme level of b. However, with the highest value $OR = 20$, SE(AUC) is not monotonic with b; it declines but then rises again slightly when the studies are very heterogeneous.

The original analysis by Scheidler *et al.* [22] focused on the values of Q^*, with approximate confidence intervals and tests of significance based on assuming normality for the estimates. The SROC curves were fitted using Moses's method, and plotted in a range of FPR values from 0 to approximately 0.6. No significant differences for Q^* between tests were claimed.

The findings in these data using the SROC curve and its AUC are essentially compatible with Scheidler's results using Q^*. Both AUC and Q^* rank the tests in the same

Table V. Standard error of AUC for various values of odds ratio and regression slope b: based on data for the lymphangiography test for cervical cancer metastases, unweighted fit.

Odds ratio	Regression slope b					
	0	0.1	0.2	0.3	0.4	0.5
	(i) $SE(AUC)$					
1	0.0630	0.0628	0.0622	0.0613	0.0599	0.0583
2	0.0601	0.0589	0.0575	0.0558	0.0539	0.0519
5	0.0489	0.0470	0.0450	0.0432	0.0415	0.0399
10	0.0380	0.0358	0.0340	0.0325	0.0315	0.0309
20	0.0275	0.0254	0.0239	0.0232	0.0230	0.0234
	(ii) *Percentage change in* $SE(AUC)$ *relative to homogeneous case* $(b=0)$					
1	0.0%	−0.3%	−1.2%	−2.8%	−4.9%	−7.5%
2	0.0%	−1.9%	−4.3%	−7.1%	−10.2%	−13.7%
5	0.0%	−4.0%	−8.0%	−11.8%	−15.3%	−18.5%
10	0.0%	−5.8%	−10.6%	−14.4%	−17.1%	−18.8%
20	0.0%	−7.6%	−12.9%	−15.7%	−16.1%	−14.6%

order – magnetic resonance, computed tomography and lymphangiography. Although Scheidler claimed no significant differences between tests, z-tests based on approximate normality of the estimates of AUC and Q^* show that magnetic resonance is in fact significantly ($p<0.05$) better than lymphangiography (with no adjustment for multiple comparisons). Visual inspection of the SROC plots suggests that lymphangiography performs less well than the other two tests in terms of TPR when FPR is relatively high. However this comparison must be made with caution because while the lymphangiography meta-analysis includes some studies with FPR over 50 per cent, the studies for magnetic resonance and computed tomography all have FPR <30 per cent. Although AUC is intended to represent the overall performance of the test, we must recognize that the range of available data is limited, and that it is even more restricted for the magnetic resonance and computed tomography tests than for lymphangiography.

7. DISCUSSION

This paper has established some basic properties of the SROC curve, concerning its position and the values and standard errors for its summary measures AUC and Q^*. It was found that the value AUC_{hom} associated with homogeneous studies (constant OR) is a reasonable upper bound approximation even for the general AUC with heterogeneous studies (non-constant OR); AUC_{hom} and its standard error are also relatively simple to compute, without the need for numerical integration.

The standard error for the homogeneous case appears to provide a good approximation for heterogeneous studies, and in most cases it is conservatively large. However, the example of the metastases did illustrate one extreme situation with a large odds ratio combined with strong heterogeneity, where the heterogeneous standard error was larger than the homogeneous estimate. In such cases, the AUC is 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^* 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^*, an index in its own right, is that it is the point where the SROC curve crosses the anti-diagonal 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 non-cases. 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 (non-case), and because the additional tests tend to be more costly, invasive or risky. On the other hand, false negative results lead to true disease cases being missed, with possible deterioration in their subsequent 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 need not then correspond to the Q^* point.

One can motivate the use of AUC as an index representing the probability that the test will correctly rank a case/non-case pair of subjects [1, 19]. 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 index 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 used, by limiting attention to those portions of the SROC curve of clinical interest, or where data are actually observed [23–26]. In the Pap smear screening data, the full range of FPR values was observed; however in the cervical cancer metastasis test data, only relatively small values of FPR occurred, so in this case there would be a stronger argument for the use of a partial SROC curve.

However, 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. Hence the partial AUC has an asymmetric interpretation with respect to the true and false positive rates; 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 choose to examine the SROC curve within a prespecified range of TPR values, for instance by defining a maximum acceptable level of TPR for clinical practice. Another approach would be to choose the region according to the observed range of data in the meta-analysis; the chosen region 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. For instance, in the metastases example data, the range of observed FPR values was much greater for the LAG test than the other two tests, so it is not immediately clear how a 'fair' comparison between the tests might be achieved using the partial AUC measure. While there are also problems in interpretation of the full AUC in data with only a limited range, it is at least based on

the same common metric of (0,1) for the values of FPR. In the other example (Pap smear), the data were relatively extensive and covered the full range of FPR values; in such situations, the full AUC is a reasonable summary measure for the whole data set. Alternatively, one might focus attention on specific points on the curve, corresponding to desirable levels of TPR in the context of population screening or clinical practice.

The analytic methods for AUC presented in this paper can be extended to cover the partial AUC and its standard error. Intuitively we may expect that the effect of inter-study heterogeneity would be greater for the partial AUC than the weak effect seen in Table II for the full AUC. For instance, if attention is limited to values of FPR < 0.2 when $a = 2$ (see Figure 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 (see Figure 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, regardless of its sign. The partial AUC does not possess these properties, and hence it will show far greater dependence on the degree of inter-study heterogeneity.

Further work is needed to explore the properties of the partial AUC in more detail. Similar investigation is needed for other summary indices proposed recently, such as the ASC (area swept out by the curve), PLC (projected length of the curve) [27], and the Gini/Lorenz coefficients [28].

Another topic worthy of further study in the comparison of SROC curves is the fact that not all the component studies in the meta-analyses may be independent of one another. For example, in the metastases data, some of the component studies made direct comparisons of two tests while other studies only evaluated one test. (None of the studies made direct comparisons of all three tests.) This feature of the data was not recognized in Scheidler's analysis or the results given here. Methods to take such dependencies into account have been proposed for therapeutic studies [29], and extensions or alternative approaches for diagnostic test comparisons would be useful.

All the results presented here are predicated on the validity of Moses's original regression model. As noted, it is an approximate model because it relies on asymptotic normality in the dependent variable D and it also ignores errors in the independent variable S. More research is needed to explore the effects of departures from the standard regression assumptions, and to elaborate the analysis, for example by examining the fit and influence of specific studies in the regression. Also, we need a better understanding of the small sample properties of the methods, both in terms of limited numbers of studies and small numbers of subjects per study; in particular, the adequacy of the normal approximation for the estimates of AUC and Q^* should be studied. Because these parameters are functions of the regression slope and intercept, broadly speaking we may speculate that the approximation will be reasonable in situations where the regression parameters can be taken as normally distributed. In the examples, there were regression sample sizes of 10–59 studies, with no obvious outliers; hence the normal approximations are probably reasonable.

The Moses model, based on the logit difference and sum (D and S) for TPR and FPR, originates from an assumption of logistic distributions for the test values in the disease cases and non-cases; normal distributions for the test values may also constitute an adequate ap-

proximation. Note, however, that the regression model (3) is fitted by ordinary least squares, without appealing directly to the logistic distribution assumption. Emphasis in the Moses approach is on characterizing the inter-study variation, and the unweighted regression fit effectively assumes that inter-study variation is much larger than intra-study variation [8]. In more generality, one might consider other functional relationships between D and S that reflect different distributional assumptions concerning the test values in cases and non-cases. Also, the use of variance components to reflect inter- and intra-study variation has been suggested [8].

The techniques discussed here apply to situations where one has only summary measures (TPR and FPR) from each study. Ideally one would have access to the individual level data in each study; one would then be able to carry out a multi-level analysis, taking both inter- and intra-study variation into account, as well as the effects of subject-specific covariates. In practice, however, methodologic reporting of diagnostic studies and meta-analyses is currently poor, and this level of detail in the data would often be difficult to achieve [5, 30–32]. Accordingly, further work to understand the properties of the SROC curve based on summary data from each study seems warranted.

APPENDIX: FORMAL PROOFS OF SELECTED RESULTS

A1. Result 1: AUC *is maximized when* $b = 0$

From (5), we have that

$$\mathrm{AUC} = 1 - \int_0^1 \frac{1}{1 + \exp(\frac{a}{1-b})(\frac{x}{1-x})^{(1+b)/(1-b)}}\,\mathrm{d}x \tag{A1}$$

We substitute $z = \ln[x/(1-x)] + a/2$, so that

$$\mathrm{d}x = \frac{\exp(z - a/2)}{[1 + \exp(x - a/2)]^2}\,\mathrm{d}z$$

When $x = 0$, $z = -\infty$, and when $x = 1$, $z = \infty$, so

$$\mathrm{AUC} = 1 - \int_{-\infty}^{\infty} \frac{\exp(z - a/2)}{[1 + \exp(z - \frac{a}{2})]^2[1 + \exp[(\frac{a}{1-b}) + (z - \frac{a}{2})(\frac{1+b}{1-b})]}\,\mathrm{d}z$$

After some simplification we can show that

$$\frac{\partial(\mathrm{AUC})}{\partial b} = \frac{2}{(1-b)^2}\int_{-\infty}^{\infty} \frac{z\exp(2z/[1-b])}{[1 + \exp(z - \frac{a}{2})]^2[1 + \exp(\frac{a}{2} + z(\frac{1+b}{1-b}))]^2}\,\mathrm{d}z$$

If $b = 0$, we have

$$\frac{\partial(\mathrm{AUC})}{\partial b} = 2\int_{-\infty}^{\infty} \frac{z\exp(2z)}{[1 + \exp(z - \frac{a}{2})]^2[1 + \exp(\frac{a}{2} + z)]^2}\,\mathrm{d}z = 2\int_{-\infty}^{\infty} z f(z)\,\mathrm{d}z \tag{A2}$$

with an obvious definition of the function $f(z)$. It is straightforward to show that $f(z) = f(-z)$; thus the integrand in (A2) is odd, and the slope of AUC with respect to b is therefore 0 when $b = 0$. Further inspection of expression (A1) reveals that AUC for values of b in the interval $(0, 1]$ (for instance, the particular value Q^* in Equation (9)) is smaller than when

$b=0$. These results together prove that AUC is maximized when $b=0$, for a given value of a.

A2. Result 2: AUC *is symmetric in b*

The SROC curve is defined by the equation

$$y(x,a,b)=\frac{\exp(\frac{a}{1-b})(\frac{x}{1-x})^{(1+b)/(1-b)}}{1+\exp(\frac{a}{1-b})(\frac{x}{1-x})^{(1+b)/(1-b)}} \tag{A3}$$

where $y(x,a,b)=\mathrm{TPR}$ and $x=\mathrm{FPR}$, and with $0<x<1$. If we define $u(v,a,b)=1-x$ and $v=1-y$, then we can show that

$$u(v,a,b)=\frac{\exp(\frac{a}{1+b})(\frac{v}{1-v})^{(1-b)/(1+b)}}{1+\exp(\frac{a}{1+b})(\frac{v}{1-v})^{(1-b)/(1+b)}} \tag{A4}$$

If we substitute $b'=-b$ in (A4), we obtain the same relationship of u and v as between y and x in (A3). Hence

$$\int_0^1 y(x,a,b)\,\mathrm{d}x=\int_0^1 u(v,a,-b)\,\mathrm{d}v$$

showing that AUC is unaltered by a change of sign in b.

A3. Result 3: derivation of AUC *for homogeneous studies*

When $b=0$, we have from (A1)

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

Using the standard integrals

$$\int \frac{1}{1+cx}\,\mathrm{d}x=\frac{1}{c}\ln(1+cx)$$

and

$$\int \frac{x}{1+cx}\,\mathrm{d}x=\frac{1}{c^2}[(1+cx)-\ln(1+cx)]$$

we obtain result (8) after noting that $a=\ln(\mathrm{OR})$.

A4. Result 4: comparison of $\mathrm{AUC}_{\mathrm{hom}}$ *and* Q^*

From equations (8) and (9) we may derive that

$$\frac{\mathrm{d}(\mathrm{AUC}_{\mathrm{hom}})}{\mathrm{d}(\mathrm{OR})}=\frac{(\mathrm{OR}+1)\ln(\mathrm{OR})-2(\mathrm{OR}-1)}{(\mathrm{OR}-1)^3} \tag{A5}$$

and

$$\frac{\mathrm{d}(Q^*)}{\mathrm{d}(\mathrm{OR})}=\frac{1}{2\sqrt{\mathrm{OR}}(\sqrt{\mathrm{OR}}+1)^2}$$

and hence

$$\frac{\mathrm{d}(\mathrm{AUC}_{\mathrm{hom}} - Q^*)}{\mathrm{d}(\mathrm{OR})} = \frac{2\sqrt{\mathrm{OR}}(\mathrm{OR}+1)\ln(\mathrm{OR}) - (\mathrm{OR}-1)(\sqrt{\mathrm{OR}}+1)^2}{2\sqrt{\mathrm{OR}}(\sqrt{\mathrm{OR}}+1)^2(\mathrm{OR}-1)(\sqrt{\mathrm{OR}}-1)^2} \tag{A6}$$

Thus the difference $\mathrm{AUC}_{\mathrm{hom}} - Q^*$ is maximized when (A6) equals 0; if $\mathrm{OR} > 1$, this occurs at the solution to the transcendental equation

$$2\sqrt{\mathrm{OR}}(\mathrm{OR}+1)\ln(\mathrm{OR}) = (\mathrm{OR}-1)(\sqrt{\mathrm{OR}}+1)^2 \tag{A7}$$

The numerical root of this equation is approximately $\mathrm{OR} = 17.3$.

A5. Result 5: derivation of $\mathrm{SE}(\mathrm{AUC}_{\mathrm{hom}})$

The delta method gives an approximate variance for $\mathrm{AUC}_{\mathrm{hom}}$ as

$$\mathrm{var}(\hat{\mathrm{AUC}}) \approx \left(\frac{\mathrm{dAUC}_{\mathrm{hom}}}{\mathrm{dOR}}\right)^2 \mathrm{var}(\hat{\mathrm{OR}})$$

Substituting the derivative from (A5) and noting that $a = \ln(\mathrm{OR})$ gives result (13).

A6. Result 6: comparison of $\mathrm{SE}(\mathrm{AUC}_{\mathrm{hom}})$ *and* $\mathrm{SE}(Q^*)$

Using results (13) and (14), we may show after some simplification that

$$\begin{aligned}\mathrm{SE}(\hat{\mathrm{AUC}}) - \mathrm{SE}(\hat{Q}^*) = \frac{\sqrt{\mathrm{OR}}}{2(\mathrm{OR}-1)^3}[&2\sqrt{\mathrm{OR}}(\mathrm{OR}+1)\ln(\mathrm{OR}) \\ &- (\mathrm{OR}-1)(\sqrt{\mathrm{OR}}+1)^2]\mathrm{SE}(\hat{a})\end{aligned}$$

Thus for $\mathrm{OR} > 1$, $\mathrm{SE}(\mathrm{AUC}) = \mathrm{SE}(Q^*)$ at the point where

$$2\sqrt{\mathrm{OR}}(\mathrm{OR}+1)\ln(\mathrm{OR}) = (\mathrm{OR}-1)(\sqrt{\mathrm{OR}}+1)^2$$

which is the same as equation (A7) which arose in the comparison of $\mathrm{AUC}_{\mathrm{hom}}$ and Q^*. Combining these results, we find that AUC is always larger than Q^*, with a maximum difference at $\mathrm{OR} = 17.3$; for $\mathrm{OR} > 17.3$, AUC has the smaller standard error, but if $\mathrm{OR} < 17.3$, Q^* has the smaller standard error.

ACKNOWLEDGEMENTS

The author thanks Dr C. Macaskill of the University of Sydney, Australia, for assistance in proving Appendix Result 1. Dr Les Irwig, also of the University of Sydney, gave helpful comments on early versions of this paper. The research was funded partly from a Senior Investigator award from the Canadian Institutes of Health Research, and a grant from NSERC, Canada.

REFERENCES

1. Hanley JA. Receiver operating characteristic (ROC) curves. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: Chichester, 1998; 3738–3745.
2. Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of Pathology Laboratory Medicine* 1986; **110**:13–19.

3. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine* 1998; **17**:1033–1053.
4. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology* 1993; **48**(1):119–130.
5. Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal Medicine* 1994; **120**:667–676.
6. Vamvakas EC. Meta-analyses of studies of the diagnostic accuracy of laboratory tests. *Archives of Pathology Laboratory Medicine* 1998; **122**:675–686.
7. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic-summary point estimates. *Medical Decision Making* 1993; **13**:253–256.
8. Moses LE, Shapiro DE, Littenberg, B. Combining independent studies of a diagnostic tests into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* 1993; **12**: 1293–1316.
9. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology* 1969; **6**:487–496.
10. Hanley, JA. The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine* 1996; **15**:1575–1585.
11. McCullagh, P. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1980; **42**:109–142.
12. Ma G, Hall WJ. Confidence bands for receiver operating characteristics curves. *Medical Decision Making* 1993; **13**:191–197.
13. Green DM, Swets J. *Signal Detection Theory and Psychophysics*. Wiley: New York, 1966.
14. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging* 1989; **29**:307–335.
15. Ogilvie JC, Creelman CD. Maximum likelihood estimation of ROC curve parameters. *Journal of Mathematical Psychology* 1968; **5**:377–391.
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
17. Hilden J. The area under the ROC curve and its competitors. *Medical Decision Making* 1991; **11**:95–101.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver-operating characteristic curves: a non-parametric approach. *Biometrics* 1988; **44**:837–845.
19. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.
20. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap smear accuracy. *American Journal of Epidemiology* 1995; **141**:680–689.
21. Walter SD. Small sample estimation of log odds ratios from logistic regression and fourfold tables. *Statistics in Medicine* 1985; **4**:437–444.
22. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. *Journal of the American Medical Association* 1997; 96–1101.
23. McClish DK. Analyzing a portion of the ROC curve. *Medical Decision Making* 1989; **9**:190–195.
24. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Statistics in Medicine* 1989; **8**: 1277–1290.
25. Wieand S, Gail MH, James BR. A family of nonparametric statistics for comparing diagnostic tests with paired or unpaired data. *Biometrika* 1988; **76**:585–592.
26. Jiang Y, Metz CE, Nishikawa, RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 1996; **201**:745–750.
27. Lee WC, Hsiao CK. Alternate summary indices for the receiver operating characteristic curve. *Epidemiology* 1996; **7**:605–611.
28. Lee WC. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Statistics in Medicine* 1999; **18**:455–471.
29. Bucher HC, Guyatt GH, Walter SD, Griffith L. The results of direct and indirect treatment comparisons in meta-analysis of randomized clinical trials. *Journal of Clinical Epidemiology* 1997; **50**:683–691.
30. Walter SD, Jadad AR. Meta-analysis of screening data: a survey of the literature. *Statistics in Medicine* 1999; **18**:3409–3424.
31. Reid MC, Lachs S, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *Journal of the American Medical Association* 1995; **274**:645–651.
32. Lijmer JG, Mol, BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Muelen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association* 1999; **282**:1061–1066.

