Summary Receiver Operating Characteristic Curve Analysis Techniques in the Evaluation of Diagnostic Tests

Catherine M. Jones, MBBS, BSc(Stat), and Thanos Athanasiou, MD, PhD, FETCS

The National Heart and Lung Institute, Imperial College of Science, Technology, and Medicine, Department of Cardiothoracic Surgery, St Mary's Hospital, London, United Kingdom

The number of studies in the literature using summary receiver operating characteristic (SROC) analysis of diagnostic accuracy is rising. The SROC is useful in many such meta-analyses, but is often poorly understood by clinicians, and its use can be inappropriate. The academic literature on this topic is not always easy to comprehend. Interpretation is therefore difficult. This report aims to explain the concept of SROC analysis, its advantages, disadvantages, indications, and interpretation for the cardiothoracic surgeon. We use a practical approach to show how SROC analysis can be applied to meta-analysis of diagnostic accuracy by using a contrived dataset of studies on virtual bronchoscopy in the diagnosis of airway lesions.

(Ann Thorac Surg 2005;79:16–20)

Background

Meta-analyses evaluating diagnostic test accuracy are increasingly prevalent in the medical literature. Two articles in this issue use summary receiver operating characteristic (SROC) analysis as part of the meta-analytic methodology [1, 2]. A diagnostic test is any measurement aiming to identify individuals who could potentially benefit from an intervention. This includes elements of medical history, physical examination, imaging and laboratory investigations, and clinical prediction rules. It is useful to know test performance over all available studies.

See pages 365 and 375

A meta-analysis of a diagnostic test calculates its overall accuracy, and may focus on its sensitivity or specificity. It incorporates data from different studies and populations. Guidelines have been published for meta-analysis of studies of diagnostic tests [3–7], based on evidence and the expertise of the Cochrane Collaboration [8].

The SROC analysis is applied to data which have been pooled from multiple sources. Why use SROC if simple averages will suffice? Data pooling can produce misleading results if the data sets vary between each other in terms of size, or study quality [9]. Poorly conducted or reported studies are more likely to produce outlying results, which skew the overall pooled data. A weighted average can be biased towards large studies, or studies comprised of very similar results. It can be difficult to identify outlying data and exclude it. On the other hand, more data mean wider conclusions can be reached. The SROC analysis deals with pooled data without these pitfalls.

Address reprint requests to Dr Athanasiou, Robotic and Minimally Invasive Cardiothoracic Surgery, 70 St Olaf's Rd, Fulham, London, SW6 7DN UK; e-mail: tathan5253@aol.com.

Guidelines for Systematic Review of Diagnostic Tests

It is vital to consider the aims and focus of the review before starting the literature search. Prespecified criteria for data extraction, and quality assessment of the studies by at least two reviewers, should be completed before any statistical analysis. It does not always produce material suitable for meta-analysis. Whether meta-analysis of pooled data can be conducted depends both on the number and methodological quality of the primary studies [9].

The simplest method for analyzing pooled data from multiple studies is calculating sensitivities and specificities (Table 1) and their averages. This is valid when the same criteria for a positive result have been used in each study, and each study is of similar size and quality. If different criteria, or thresholds, have been used, there will be a relationship between sensitivity and specificity across the studies. As sensitivity increases, specificity will generally drop. This is the threshold effect. In these cases, weighted averages will not reflect the overall accuracy of the test, as the extremes of threshold criteria can skew the distribution.

To demonstrate this, we have constructed a contrived data set (Table 2). The clinical setting is the use of virtual bronchoscopy (VB) in the diagnosis of airway lesions. In this dataset of 12 studies, the average sensitivity is 0.86, and average specificity 0.54. These values give an idea of the overall performance of VB as a diagnostic tool, but there is no information about the presence of outlying data points, or the weight given to each point. Each data point is derived from a different study result for sensitiv-

Published by Elsevier Inc

0003-4975/05/$30.00
doi:10.1016/j.athoracsur.2004.09.040

Table 1. Tables for Sensitivity and Specificity

		True State	
		+	−
Test result	+	TP	FP
	−	FN	TN

Sensitivity = TP / (TP + FN) = TPR.

Specificity = TN / (TN + FP).

FPR = 1 − specificity = FP / (TN + FP).

Diagnostic odds ratio = [TPR (1−FPR)] / [FPR (1−TPR)].

FN = false negative; FP = false positive; FPR = false positive rate; TN = true negative; TP = true positive; TPR = true positive rate.

ity and specificity. The relationship between sensitivity and specificity is not considered and it is not possible to evaluate if there is a threshold effect across the studies. There is a need to assess whether these pooled results are accurate.

The SROC analysis is the best method for examining these issues [10]. The reasoning behind this is set out in the following sections. The easiest way to understand SROC is to look first at its predecessor, the receiver operating characteristic (ROC) method.

Principles of ROC

The ROC analyses the accuracy of a single test in a single population. It compares test accuracy over different thresholds for positivity [7, 10]. The ROC curves graph sensitivity (or true positive rate [TPR]) against (1-specificity) (or false positive rate [FPR]), where each point is derived from a different threshold.

Once the test and true results are known for a diagnostic threshold, sensitivity and specificity are calculated. Another threshold is chosen, and sensitivity and specificity calculated again. Eventually there is a number of (sensitivity, specificity) pairs for the test, each corresponding to a different diagnostic threshold. (Sensitivity, 1-specificity) pairs are then calculated and plotted. Sensitivity is on the vertical axis, and (1-specificity) on the horizontal. These points make up the basis of the ROC graph (Fig 1).

Table 2. Results From the VB Example (Contrived Dataset)

Study	TP/FN	FP/TN	Sensitivity	1-Specificity
1	17/6	1/19	0.74	0.05
2	20/1	12/3	0.95	0.80
3	19/1	1/3	0.95	0.25
4	16/2	5/7	0.89	0.42
5	44/3	2/5	0.94	0.29
6	20/9	10/8	0.69	0.55
7	23/6	3/1	0.79	0.75
8	13/1	8/8	0.93	0.50
9	11/4	1/5	0.73	0.17
10	20/1	1/2	0.95	0.33
11	21/2	1/2	0.91	0.33
12	13/4	3/3	0.76	0.50

Total number of subjects = 391; Pooled sensitivity = 0.86 (95% CI 0.81–0.90); Pooled specificity = 0.54 (95% CI 0.44–0.65).

CI = confidence interval; FN = false negative; FP = false positive; TN = true negative; TP = true positive; VB = virtual bronchoscopy.

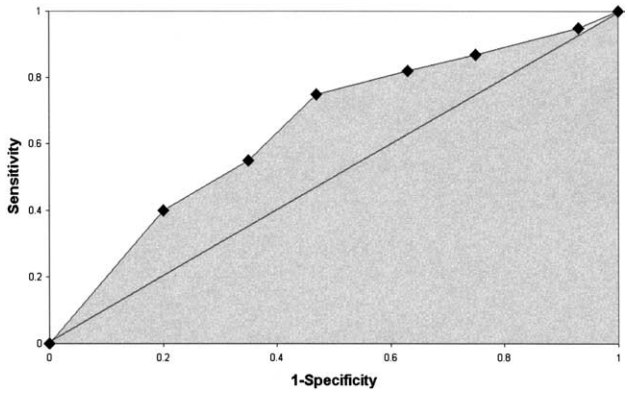

Fig 1. Schematic receiver operating characteristic (ROC) curve also showing the diagonal (the random test with area under the curve [AUC] of 0.5). The area under the ROC curve is the AUC (shaded).

Low thresholds produce more true and false positives. Strict criteria for a positive result produce fewer positives, with low sensitivity and high specificity.

To demonstrate its use, consider the contrived VB example. An airway stenosis is distinguished from a normal airway by the degree of airway narrowing. The investigator nominates a minimum degree of narrowing required for a positive diagnosis. By changing the threshold, the number of positive and negative results on VB changes. For example, if a minimum of 5% narrowing is required for diagnosis of a stenosis, many airways will be positive for stenosis (high sensitivity, low specificity), whereas a minimum threshold of 90% narrowing produces far fewer positive results (low sensitivity, high specificity).

The advantage of ROC is that accuracy is plotted for different thresholds and compared. Overall test accuracy is measured by the closeness of the graph to the top left corner, which represents high sensitivity and specificity. This is more easily visualized by a curve placed over the points. The closer the curve to the unit square, the better the overall accuracy. The curve is made either by fitting the points together in a straight line, or using a smoothing function [11–15].

Sensitivity and specificity values lie between zero and one inclusive. This makes the area under the curve (AUC) for a perfect test equal to one. The random test, allocating positive results half the time, has an AUC of 0.5.

The AUC can be calculated for different diagnostic tests, and then compared to each other [11–13]. An AUC closer to one indicates a better test. It is the probability of a randomly selected pair of a true positive and a true negative being ranked as such by the diagnostic test [16–18].

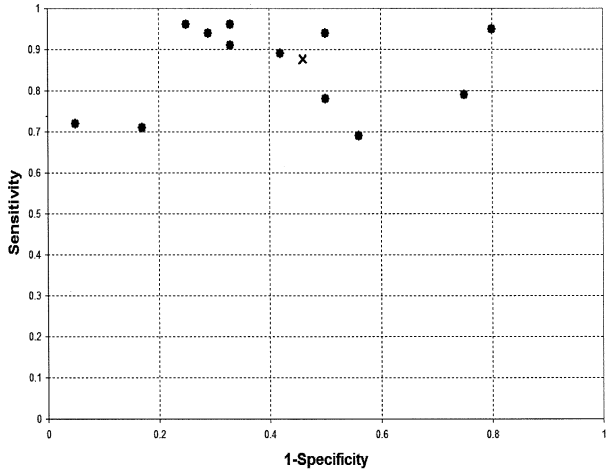

Fig 2. Sensitivity and (1-specificity) results from 12 studies (●) of accuracy of virtual bronchoscopy (contrived data). Pooled sensitivity and (1-specificity) are shown as X.

Principles of SROC Curves

The SROC graph is conceptually very similar to the ROC. However, each data point comes from a different study, not a different threshold. Diagnostic thresholds should be similar for each study, so threshold effect does not influence the shape of the curve. The curve is shaped solely by the results across the studies. Each study produces values for sensitivity, specificity and therefore TPR and FPR. A graph is made from the (TPR, FPR) points.

The SROC curve is placed over the points to form a smooth curve. It is calculated from a number of possible formulas. The most commonly used is a regression model [10] where sensitivity and (1-specificity) are transformed into complex logarithmic variables and graphed (for purposes of discussion, log variables 1 and 2). A regression equation is calculated, and the variables are manipulated to achieve sensitivity as a function of (1-specificity). This is the equation for the SROC curve, which is then plotted over the original (sensitivity, 1-specificity) points on the original axes.

The VB example demonstrates these steps. The sensitivity and (1-specificity) values in Table 2 are plotted on the normal axes (Fig 2). The points are transformed into the complex logarithmic variables and plotted with a regression line (Fig 3). The variables in the regression equation are transformed back into TPR and FPR. The regression line is transformed from a straight regression line (Fig 3) into the SROC curve on the original axes (Fig 4). The SROC curve shows that the points do not all lie on the curve. The pooled sensitivity and specificity values (the X in Figs 2 and 4) do not lie on the curve; they appear to underestimate the accuracy of the test.

The (TPR, FPR) points will not lie on a smooth curve. Smoothing formulas place the final curve as close as possible to the overall data set, treating each pair equally. It can be placed closer to certain points by weighting their importance according to study size, study variance, population characteristics or other study characteristics

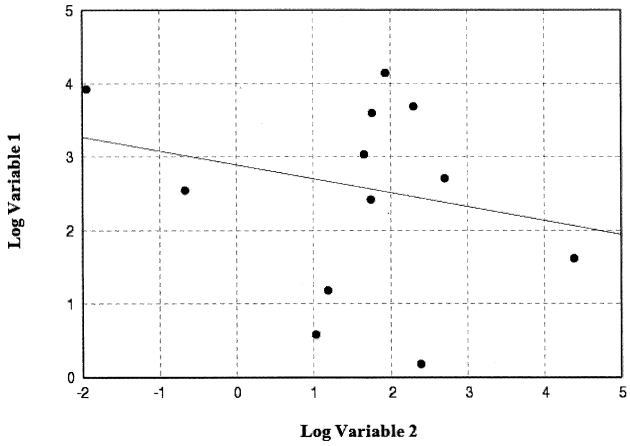

Fig 3. The regression model of the transformed data points (●) from Figure 2, shown on logarithmic axes, with the regression line shown.

[18]. The differences in study characteristics across the primary studies, called heterogeneity, should be minimized to prevent influencing the shape of the curve.

The AUC is calculated for SROC as for ROC. The diagnostic test is constant throughout the studies, so the AUC reflects overall performance of that test. The perfect test will again have an AUC of one. The SROC is reproducible and thus the AUC can be used to compare accuracy of different diagnostic tests [18].

The AUC for the VB example is 0.82. While reasonable, it shows that in this example, VB accuracy needs improvement before being adopted as first line investigation in these patients. Weighted analysis to assess the impact of study numbers showed no difference in AUC

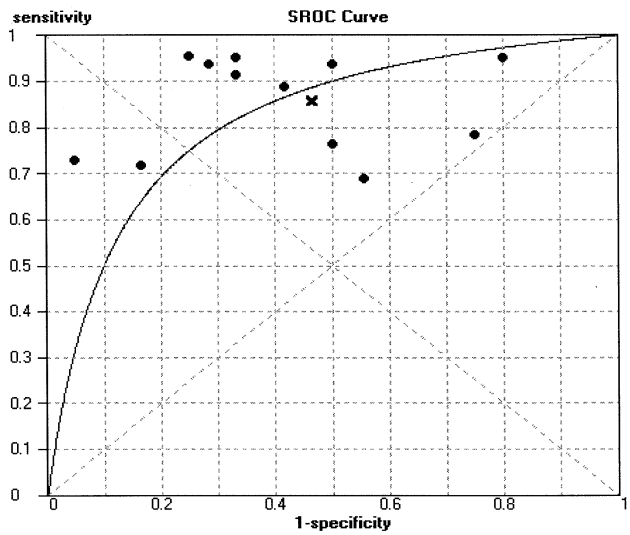

Fig 4. The points (●) from Figure 2, shown with the SROC curve superimposed. The regression line in Figure 3 has been transformed into the SROC curve, and the points in Figure 3 have been reverted back to the points from Figure 2. Pooled sensitivity and (1-specificity) are shown as X. (SROC = summary receiver operating characteristic.)

values. This implies that there was no significant bias of study size.

If the FPR values are limited to part of the range, the SROC curve and AUC calculation will only be accurate for this range. A partial AUC (for the range of experimental FPR points) is one possible solution [14, 15, 18], although it cannot be used for comparison with other tests.

Q and Diagnostic Odds Ratio

Q is the intercept of the SROC and the antidiagonal line through the unit square. Its value indicates overall accuracy by finding where sensitivity and specificity are the same. The closer the curve to the top left corner (perfect sensitivity and specificity), the better the accuracy. The antidiagonal will cut the curve at a higher level, giving higher Q, in a more accurate test.

Q is appropriate provided high sensitivity and high specificity are equally desirable. If one is clinically more important than the other, Q does not address the clinical usefulness of the test. In these cases, overall accuracy is not as relevant as overall sensitivity or specificity.

For the VB example, a positive result may lead to further investigation or surgery, while a negative result may prompt no further immediate or invasive action. High sensitivity is clinically more important than high specificity, although that is also desirable. Generally, failing to recognize a potentially malignant airway lesion has greater implications than performing an unnecessary investigation. As well as overall accuracy, we are interested in overall sensitivity. In this case, the antidiagonal crosses the SROC curve at (0.25, 0.75), giving a Q of 0.75. This shows overall accuracy is reasonable, but it does not indicate its overall sensitivity.

Diagnostic odds ratio (DOR) is calculated from sensitivity and specificity (Table 1). It is another measure of the overall diagnostic power of the test. A high DOR implies that the test shows good diagnostic accuracy in all patients. For the VB example, DOR was 8.98, significantly greater than one. This implies that VB is better than the random test at diagnosing airway lesions.

How Many Studies?

The assumption of SROC that the data points are normally distributed depends on the number of studies. Generally, it is assumed that ten or more data points are needed for this assumption to be valid. Outlying data points need to be assessed for validity. When studies produce dramatically different results, this can be random error or due to differences in study methodology, population, or test characteristics. Small studies, for example, may produce extreme results from a small population.

Comment

The SROC can be appropriate for analyzing overall accuracy of a diagnostic test from multiple study results. For SROC to be valid, there should be similarity of study methodology and quality, diagnostic threshold, study endpoints, and test variables. In practice, not all articles using SROC analysis meet all these criteria. The reader must decide whether to dismiss the analysis results.

Study numbers are important, small studies are prone to producing outlying results, and shift the overall outcome. Calculations can be weighted for study size and the results compared to nonweighted calculations. If there is a difference, then study size may be contributing to the different sensitivity and specificity results. Heterogeneity of the studies may contribute to the different sensitivity and specificity results. Its significance can be assessed through graphical exploration (Galbraith plots), or metaregression of DOR against relevant study or patient variables.

A fair test shows better than average accuracy, and has an AUC above 0.5. To demonstrate excellent accuracy, the AUC should be in the region of 0.97 or above. An AUC of 0.93 to 0.96 is very good; 0.75 to 0.92 is good. Less than 0.75 can still be reasonable, but the test has obvious deficiencies in its diagnostic accuracy, and is approaching the random test. It is important to remember that the AUC must be interpreted according to the context of the individual analysis and that these guidelines are not absolute.

A major advantage of SROC analysis is that data pooling problems are overcome. It takes a greater effect to shift a curve than to change an average. Outliers are easier to spot on a graph. It can be used for different tests, and the same statistic (AUC, Q, DOR, as appropriate) used to compare their accuracies. Finally, other variables can be used to weight the analysis if clinically indicated.

Limitations of SROC

The SROC modeling has several limitations. When the calculations are weighted, there is a bias towards studies with lower DOR [8], leading to underestimation of accuracy. The variables in the regression for the SROC curve are themselves functions of sensitivity and specificity. They are not independent of each other, and theoretically should not be used to calculate the SROC curve formula without accounting for this interdependence.

The most important drawback of SROC is its assumption that the primary studies are random samples of one large common study, and that differences in results are random error. It does not account for patient variables, study variables, physician experience and training, and institutional characteristics. One solution to this is hierarchical SROC, which accounts for variation both within and between studies [19]. Both threshold and accuracy are included in the model. In practice, it is rarely used because the calculations and interpretations are complex and few software packages include it. Its use may increase as software is developed and general understanding of SROC analysis widens.

Sometimes the sensitivity and specificity will be available for different thresholds within the same study. Depending on the predetermined diagnostic threshold, and amount of literature available, the most appropriate threshold should be chosen for the analysis. With enough

literature available, it is possible to perform SROC analysis for different thresholds of the same test. The AUC or Q would be used, where appropriate, to compare the accuracy of the same test for different thresholds. This requires multiple analyses which are often published separately.

The range of commercial software calculating SROC statistics is limited and difficult for nonstatisticians. Freeware software such as Meta-Test [20], developed by Dr Joseph Lau, are commonly used to generate curves and statistics. Important steps forward would include user-friendly software, individual patient data analysis [21], and methods to allow conditional dependence between multiple test results in individuals.

Conclusion

Summary receiver operating characteristic analysis is increasingly popular for meta-analyses of diagnostic test validity. It is suited to this purpose with its use of sensitivity and specificity. However, it is only meaningful when similar endpoints, diagnostic threshold, study quality, and test characteristics are compared. The AUC and Q statistics are used to compare results from different SROC analyses. Partial AUC may be used if specificity values are limited, with interpretation on an individual meta-analysis basis.

There is a lack of understanding by clinicians of the concepts and interpretation of SROC. It is our hope that as SROC becomes more popular and understanding grows, interpretation of SROC analysis will become easier.

The authors would like to thank Dr Gary Grunkemeier for his contribution to the editing of this manuscript.

References

1. Jones CM, Athanasiou T. Is virtual bronchoscopy an efficient diagnostic tool for the thoracic surgeon? Ann Thorac Surg 2005;79:365–74.
2. Birim O, Kappetein AP, Stijnen T, Bogers AJJC. Meta-analysis of positron emission tomographic and computed tomographic imaging in detecting mediastinal lymph node metastases in nonsmall cell lung cancer. Ann Thorac Surg 2005;79:375–82.
3. Irwig L, Tosteson ANA, Gatsonis C, et al. Guidelines for meta-analysis evaluating diagnostic tests. Ann Intern Med 1994;120:667–76.
4. Shapiro, DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. Acad Radiol 1995;2:S37–S47.
5. Deville WL, Buntinx F, Bouter LM, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002;2:9.
6. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Radiology 2003;226:24–8.
7. Vamvakas EC. Meta-analyses of studies of the diagnostic accuracy of laboratory tests. Arch Pathol Lab Med 1998;122:675–86.
8. Cochrane Collaboration. http://www.cochrane.org.
9. Rutter CM, Gatsonis GA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol 1995;2:S48–56.
10. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13:313–21.
11. Beck JR, Shultz EK. The use of receiver operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 1986;110:13–9.
12. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. J Math Psychol 1975;12:387–415.
13. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating curve (ROC) curves from continuously distributed data. Stat Med 1998;17:1033–53.
14. Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. Med Decis Making 1988;8:204–15.
15. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989;9:190–5.
16. Centor RM, Schwartz JS. An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. Med Decis Making 1985;5:149–56.
17. Hilden J. The area under the ROC curve and its competitors. Med Decis Making 1991;11:95–101.
18. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med 2002;21:1237–56.
19. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865–84.
20. Lau, J. Meta-Test software. http://www.medepi.org/meta/MetaTest.html.
21. Khan KS, Bachmann LM, ter Riet G. Systematic reviews with individual patient data meta-analysis to evaluate diagnostic tests. Eur J Obstet Gynecol Reprod Biol 2003;108:121–5.