Estimating Diagnostic Accuracy from Multiple Conflicting Reports:

A New Meta-analytic Method

BENJAMIN LITTENBERG, MD, LINCOLN E. MOSES, PhD

Reports of diagnostic accuracy often differ. The authors present a method to summarize disparate reports that uses a logistic transformation and linear regression to produce a summary receiver operating characteristic curve. The curve is useful for summarizing a body of diagnostic accuracy literature, comparing technologies, detecting outliers, and finding the optimum operating point of the test. Examples from clinical chemistry and diagnostic radiology are provided. By extending the logic of meta-analysis to diagnostic testing, the method provides a new tool for technology assessment. *Key words:* meta-analysis; sensitivity and specificity; decision support; data interpretation, statistical; regression analysis; diagnostic accuracy. **(Med Decis Making 1993;13:313–321)**

A clinician or policy maker who is concerned about the value of a diagnostic test may decide to go to the library to find out how accurate the test is. There, he or she will often discover that much has already been done to characterize the accuracy of the index test by comparing it with a reference test or "gold standard." Some index tests have been evaluated dozens of times. Naturally, the various reports do not agree perfectly. Success in using quantitative syntheses to summarize research reports[1,2] might prompt our analyst to attempt meta-analysis as a means of summarizing and understanding the variety of published reports on diagnostic accuracy.

We address two problems in applying meta-analysis to diagnostic technologies. First, there is no one number that represents accuracy. Second, there are many reports of diagnostic accuracy, and they don't agree. This paper presents a method for summarizing discrepant data on the accuracy of diagnostic technologies, gives case studies of its application, and provides guidelines for its use.

Rationale for Using a Curve to Summarize Accuracy

A therapeutic technology is generally evaluated in a homogeneous population of persons all of whom have the disease in question. A single number, perhaps the cure rate, can be used to describe the efficacy of the therapy in that population. (If there are many outcomes of interest, a family of numbers may be needed.) In contrast, a diagnostic technology is applied to a mixed population of patients with and without the disease. At least two parameters are needed to describe this situation. On the one hand, the test should perform well in detecting sick patients. This is represented as the true-positive rate (TPR), or sensitivity. On the other hand, it should also be accurate in identifying the well. This kind of accuracy is measured as the false-positive rate (FPR) (one minus the specificity) of the test.

The TPR and the FPR can be made to vary in any given test in any given population by changing the criterion for a positive test. This criterion is called the threshold. More extreme values of the test are interpreted as indicating disease. Any test can be made to look good in identifying the well (in other words have high specificity) if the threshold for a positive test is set very high. However, with a very high threshold, few of the diseased patients will be detected and sensitivity will suffer. *Identifying the well and detecting the sick are in constant tension*. There is a family of pairs of true-positive and false-positive rates that describe the functioning of any diagnostic test.

When arrayed on a graph of true-positive vs false-positive rates, this family of numbers generally traces out a curve, called the receiver operating characteristic (ROC) curve (fig. 1). Any report that provides only a single TPR and a single FPR is not providing a full picture of the test's accuracy. It is stating the two kinds of accuracy at a particular threshold only. Often, as in imaging studies, the threshold is not well defined, but the trade-off between TPR and FPR applies all the same.

Received May 4, 1992, from the Technology Assessment Program, Department of Medicine, Dartmouth–Hitchcock Medical Center, Lebanon, New Hampshire (BL); and the Department of Statistics, Stanford University, Stanford, California (LM). Revision accepted for publication March 25, 1993. Presented in part at the meeting of the American Federation for Clinical Research, May 6, 1990. Supported in part by the John A. Hartford Foundation. Dr. Littenberg is the recipient of an American College of Physicians George Morris Piersol Teaching and Research Scholarship.

Address correspondence to Dr. Littenberg: Dartmouth–Hitchcock Medical Center, Lebanon, NH 03756. Reprints are not available.

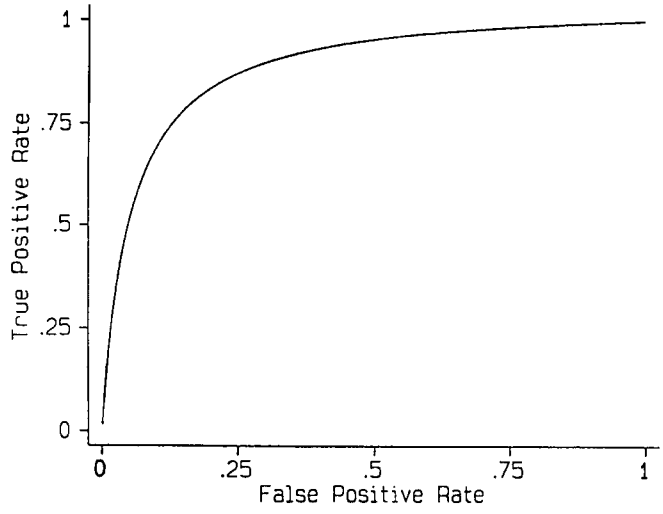

FIGURE 1. A typical receiver operating characteristic curve. The vertical axis depicts the true-positive rate (TPR) of the test. The false-positive rate (FPR) varies along the horizontal.

Need for Meta-analysis

Unfortunately, full ROC curves are rarely published. More often there are a number of studies, all purporting to report on the same test, and all giving just a pair of numbers: the true-positive rate and the false-positive rate. Usually, the threshold at which accuracy was measured is not mentioned or is subject to substantial inter-operator variation.

In figure 2, there are seven points. Each represents one study of a diagnostic test. We've plotted each according to its reported true-positive rate on the vertical axis and its reported false-positive rate on the horizontal axis. As is often the case, the reports don't agree.

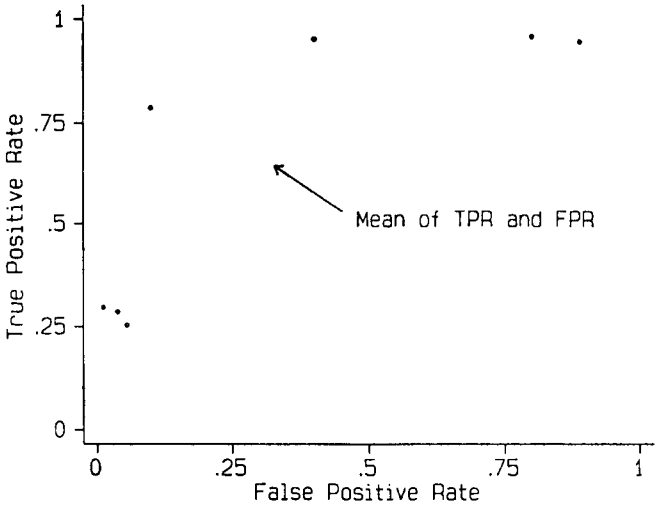

FIGURE 2. Seven hypothetical estimates of test accuracy. Each point represents a single report of a hypothetical test's accuracy. The simple mean of the true-positive rate and the false-positive rate is shown by the arrow. Notice that the joint mean does not fall near any of the data.

How should we summarize these numbers? A first guess would be to take the average true-positive rate and the average false-positive rate, but this is often very misleading. For these seven studies, the average falls far away from any of the seven. It does not seem to provide an adequate summary because the two averages form a single point that does not take into account the tension between TPR and FPR that is generated by varying threshold.[3]

A Method for Estimating a Summary ROC Curve

How can we plot a summary receiver operating characteristic (SROC) curve? We would like to perform some kind of analysis that estimates a smooth curve through (or near) all the data points. The raw data consist of the fourfold tables of true positives, false positives, false negatives, and true negatives from each relevant report. First, we transform the vertical and horizontal scales in a way that will reasonably allow us to fit a straight-line regression. Next, we estimate the slope and intercept of that line. Then we reverse the transformation to find the SROC curve. Some properties of this method are described in somewhat more detail elsewhere.[4–6]

Following methods widely used in the analysis of binary data, we convert the TPR and FPR from each study to their logistic transforms. The logit of the TPR is the logarithm of the TPR divided by one minus the TPR:

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

Likewise, for the FPR:

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

If either the TPR or the FPR is exactly zero or one (as happens when the fourfold table of test data contains a zero cell), then equations 1 and 2 are undefined. We avoid this by adding one half to all counts in all the tables (including the non-zero counts) that are used to calculate TPR and FPR. We now define:

$$S = \text{logit(TPR)} + \text{logit(FPR)} \quad (3)$$

$$D = \text{logit(TPR)} - \text{logit(FPR)} \quad (4)$$

S is the sum of the two transforms and is related to how often the test is positive, which is related to the test threshold. D is the difference between the two transforms and is a measure of how well the test discriminates between the two populations of well and sick subjects.

In general, TPR and FPR tend to increase (and decrease) together. Many circumstances that render true-positive judgements more likely also make false-positive judgements more likely. Chief among these, but not alone, is variation in threshold. This positive covariation, after conversion to the logit scale, is likely to contain an important component of linear regression. It is that which our methods address. At first it seems natural to regress logit TPR on logit FPR—or vice versa. These two regressions offer themselves, but neither is especially natural. However, linear regression of either on the other implies that D has a linear regression upon S, and D *is* the natural dependent variable. In fact, D is the log-odds ratio, a direct measure of discriminatory power. S is large and positive if both TPR and FPR are large, and negative if they are small, so the plot of data points in (S, D) space displays how the discriminatory power, as captured in the log-odds ratio, may vary with the stringency of the test criteria (or other causes of high or low positive test rates). We have found that we are following here in footsteps more than 20 years old (Cox,[7] example 6.1).

Next, we need to estimate the relationship between D and S. We fit a linear model:

$$D = bS + i \tag{5}$$

In the appendix, three alternative methods of fitting the straight line are illustrated with a numerical example. Two of them are easily executed with a pocket calculator. (We defer to a later section discussion of how the statistical properties of the various approaches differ.)

Once we know the slope and intercept of the transformed line, we can use equation 6 to back-transform it to the more familiar representation:

$$TPR = \frac{1}{1 + \dfrac{1}{e^{i/(1-b)} \cdot \left(\dfrac{FPR}{1 - FPR}\right)^{(1+b)/(1-b)}}} \tag{6}$$

When the slope, b, is near zero, this equation yields a nice smooth curve that is concave down. The higher the intercept, i, the closer the curve will be to the upper left-hand corner. Put another way, the height of the transformed line is a measure of how well the test discriminates between the sick and the well. The farther apart these two populations are, the greater is i. The slope, b, is also interesting. When the populations of sick and well have similar variances, the slope of the line, b, is near zero and the resulting SROC curve is nearly symmetrical. If the two populations (the sick and the well) have very different variances, the slope of the line will be far from zero, and the SROC curve will have a distorted appearance. For practical purposes, if $-0.5 < b < +0.5$, then the SROC curve looks reasonably like a typical ROC curve.

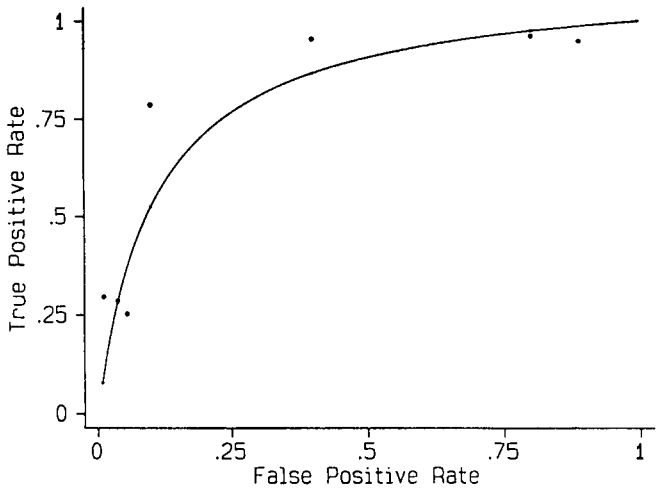

FIGURE 3. A summary receiver operating characteristic (SROC) curve for the seven hypothetical estimates of figure 2, derived by the method described in the text.

Since D_k is the log-odds ratio of the kth study, a summary of D (by equally weighted mean, weighted mean, or median, for example) can serve as a reasonable estimate of the discriminating power of the test. D is a more satisfactory summary when b is near zero, that is, when the discriminatory power does not vary much as stringency varies.

Figure 3 is the curve generated from the seven sample data points we averaged earlier. Unlike a traditional ROC curve, which describes the impact of threshold in a single patient population, this curve describes the test in many populations. Just as the mean can summarize a set of numbers, this curve summarizes the central tendency of a set of accuracy reports.

Because this method involves regression analysis, it is more or less susceptible to the untoward effects of extreme outliers. For instance, if we are examining a test that is used to "rule in" serious disease, we may insist that it be operated with a threshold that results in very high specificity (low FPR), even if the sensitivity (TPR) is reduced. If one or more studies report very high TPR and FPR, the estimated SROC curve will be heavily influenced by data that have little bearing on the clinical problem. We advise restricting analysis to those data that lie within a clinically determined "relevant range," determined prior to the data analysis. Likewise, we do not extrapolate the SROC curve beyond the range of the included data. Examining only a relevant range introduces a bias that slightly inflates the apparent accuracy of the index test, but this bias is not large.[5]

The curve represents a simplifying approach that allows us to approach this question: Why do the different estimates of accuracy vary in both TPR and FPR? There are several possible answers:

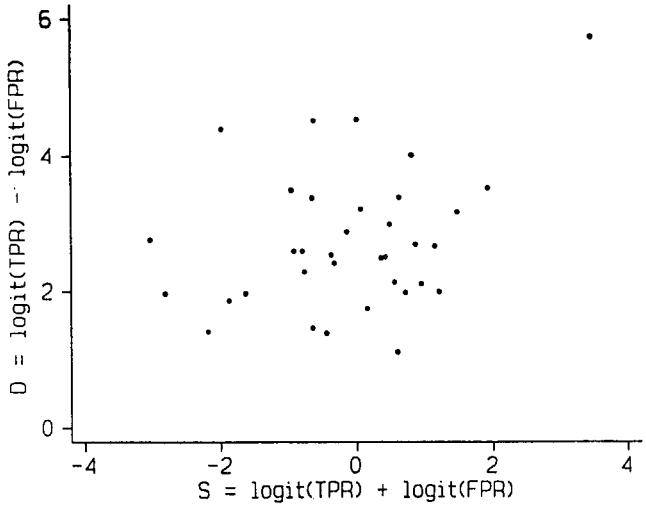

FIGURE 4. Leukocyte esterase dipstick test data. The horizontal axis depicts the quantity S (see equation 3) and the vertical axis represents D (equation 4) for 35 reports of the accuracy of the leukocyte esterase urine dipstick test as compared with urine culture.

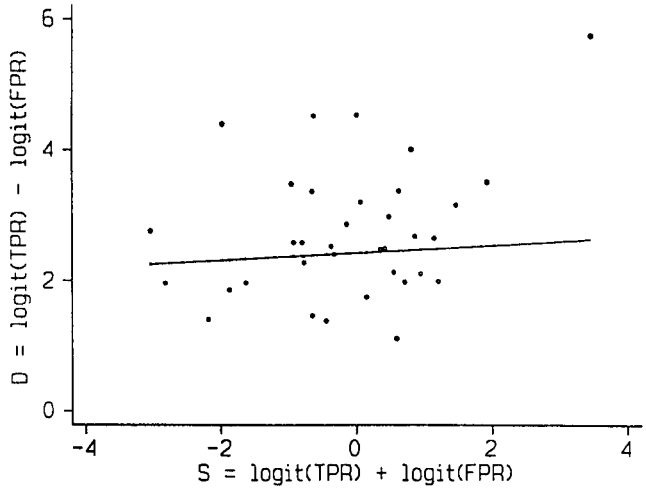

FIGURE 5. Leukocyte esterase dipstick test data regression. The data and axes are the same as those in figure 4. The line represents the least-squares regression line through the 35 data points weighted by the inverse of the variance of each study.

The reports study different diseases

The reports study different reference tests

The reports study different types of patients

The reports study populations with different prevalences

The reports study different index tests

The reports use different study methodologies

The reports have random error

The reports use different thresholds (threshold effect)

By arraying the reports as points in ROC space and searching for an SROC curve, we investigate the possibility that the threshold effect (or other factors that affect TPR and FPR together) explains much or all of the variation in reported accuracy. If all the points fall near a single ROC curve, their diversity can be explained, in large part, by differences in threshold (or other such factors). The strength of this approach is not that we are sure that the threshold effect is always the cause of the variation in accuracy, but rather that it nearly always accounts for at least some of the variation. If we do not account for the threshold effect, we have no hope for understanding the other causes of variation.

Case Study 1: Using the SROC to Summarize Accuracy Data

We analyzed 35 reports that each compared a leukocyte esterase dipstick with culture for bacteriuria.[4] We converted the data to their logistic transforms to obtain figure 4. Notice that the data tend to form a horizontal line. In the next step, we used weighted least-squares regression to obtain a line (fig. 5). The slope (b) is 0.06 and the intercept (i) is 2.4. (The weights for this regression happen to be higher for data points with lower values of D. Therefore, the regression line falls below the apparent midpoint of the data.) Then we back-transformed the line to obtain the curve in figure 6. Notice that we do not extrapolate the curve past the range of empiric data.

The curve has several advantages. First, it appears to represent the central tendency of all the included data. Second, it is completely specified by just two parameters, b and i. Third, it has the familiar concave-down form that represents the tension between the

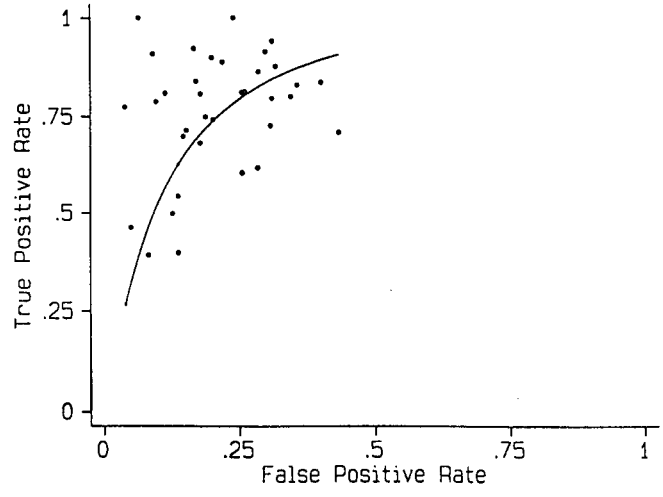

FIGURE 6. Leukocyte esterase dipstick test data back-transformed. The data and curve from figure 5 were back-transformed (using equation 6) to allow representation in the familiar receiver operating curve space.

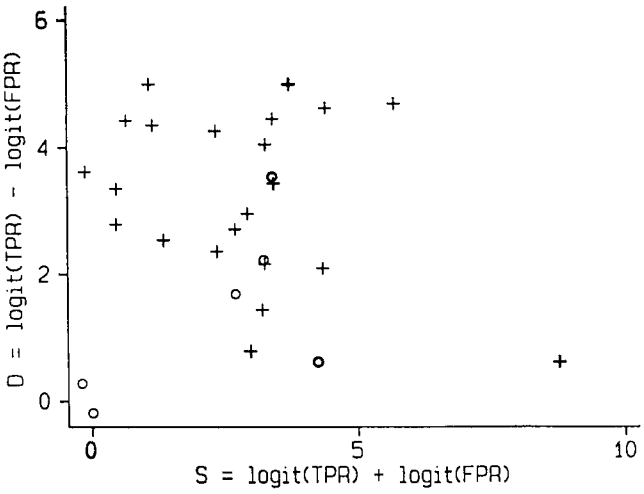

FIGURE 7. Thermography study data. Twenty-eight studies of the accuracy of thermography in back pain are presented. The format is the same as that in figure 4. The crosses represent 22 studies of infrared thermography. The six circles represent studies of liquid crystals thermography.

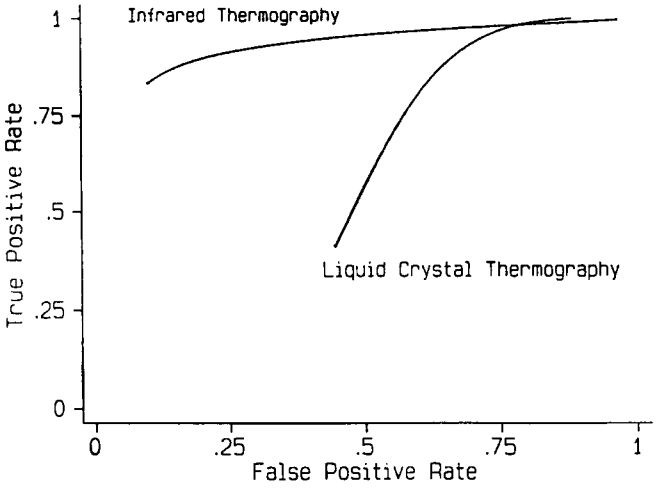

FIGURE 8. Thermography study data back-transformed. The weighted least-squares–estimated summary receiver operating curve for each modality is shown. The two curves are both clinically and statistically different.

well and the sick. Fourth, *it was constructed without knowledge of the specific threshold for positivity employed in each report*. We interpret the curve as more or less consistent with each of the included reports of accuracy. It serves as the best available summary of the studies of the diagnostic accuracy of the technology.

Summary ROC vs Single-population ROC

The new SROC curve from the meta-analysis differs from the traditional single-population ROC analysis in some important ways. In traditional ROC analysis, each curve represents a single population. The summary ROC curve is derived from several independent populations. A single-population ROC curve describes how TPR and FPR vary as the threshold varies, all else being held constant. The new method summarizes many reports without specifying which variables are different from report to report.

The SROC curve serves principally as a compact description of the accuracy of a diagnostic test. Where two different tests for the same purpose are under consideration, the two summaries capture the information at hand, and statistical significance can be tested in various ways, one of which we show in case study 2. Another use of the SROC is in constructing decision models; a fuller discussion appears in case study 4.

Case Study 2: Using SROC to Compare Technologies

The summary curve has been useful to analyze the impacts of index test characteristics upon reported accuracy. For instance, figure 7 shows transformed data from 28 studies of thermography in low back pain.[8] The 22 crosses represent studies that used infrared thermography. The six circles come from reports on liquid crystal thermography. Notice that the crosses sit higher on the graph than the circles. The average difference in height represents the difference in discrimination between the two test methods. We tested the difference between the two groups by using Student's t-test. The difference is statistically significant at $p = 0.003$.

After back-transforming the data, we can see (fig. 8) that the two curves they generate are very different. We conclude that the accuracy of thermography is highly sensitive to the type of equipment used. *We are able to draw this conclusion without knowing the exact threshold for positivity used in each of the reports.*

Case Study 3: Using SROC to Detect Outliers

Summaries may also be useful to detect discrepant points, or outliers. Figure 9 shows data from 11 reports of the accuracy of technetium bone scanning for osteomyelitis in the foot.[9] Although there is a lot of heterogeneity among these points, the most disturbing data come from the study marked as an outlier. This study suggests that bone scans are worse than a random coin flip for diagnosing osteomyelitis! A careful review revealed that this study was done with about half the dose of technetium that the other investigators used. Since this dose is not generally used anymore, it may be reasonable to exclude this study from analysis.

With this study excluded, the curve shifts upward. Looking at figure 10, we conclude that the dose of technetium is important. *Again, this conclusion does not require specific knowledge of the precise threshold employed in each study.*

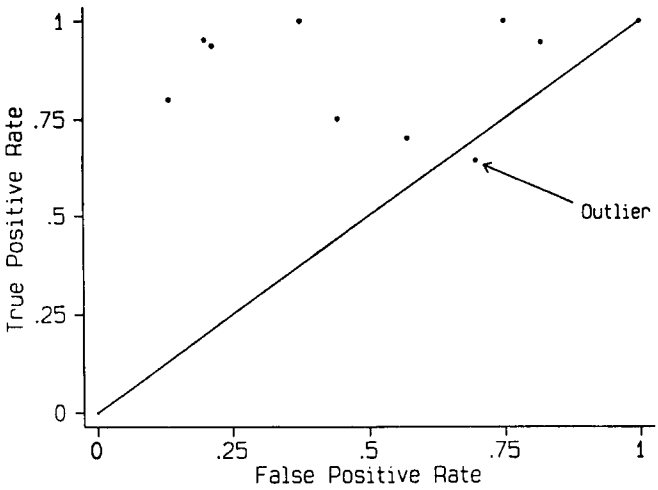

FIGURE 9. Bone scanning data. The accuracies from 11 published reports are displayed in receiver operating curve space. Notice that the outlier has a false-positive rate that is greater than its true-positive rate.

Case Study 4: Using SROC to Specify the Optimum Operating Characteristics of a Test

Receiver operating characteristic curves, including meta-analytic SROC curves, do not specify the exact operating point (the threshold and its associated TPR and FPR) that is best. This depends upon factors besides the accuracy of the test. These factors include the prevalence of disease, the risks and benefits of the four possible test results (TP, FP, FN, and TN), and all of their consequences. If one knows the relative importance of each of these factors and the shape of the ROC curve, it is possible to derive the optimum operating point on the ROC curve.[10,11]

A modification of this approach starts with a decision tree that models the use of a diagnostic test. The model requires estimates of TPR and FPR to calculate the expected utility of testing. The usual technique is to review the available data and make a point estimate of TPR and FPR for the model. You can do sensitivity analysis, but it is not clear whether to vary TPR and FPR independently or simultaneously. So, instead of specifying an exact pair of TPR and FPR, use equation 6. First, estimate b and i (and the other probabilities and utilities in the tree) as well as possible. Wherever FPR must be specified, put it in as a variable. For TPR, use equation 6, which makes TPR a function of FPR. You can subject the tree to a sensitivity analysis by systematically varying FPR and recalculating the expected utility of testing at each FPR. Equation 6 will ensure that for each value of FPR, the decision model uses an appropriate value for TPR. You can simultaneously vary both aspects of the test's accuracy by changing only one variable: FPR. Typically, if you plot the expected utility of testing on the vertical axis and FPR on the horizontal axis, the resulting curve will have an inverted U shape. The peak of the curve is the optimum operating point. One should choose a threshold for the test that yields the FPR (and TPR) associated with the highest expected utility.

For example, the decision to use a technetium bone scan to diagnose osteomyelitis in the inflamed foot of a diabetic was modeled using a decision tree and the summary ROC curve from figure 10.[12] To detect the optimum operating point, we performed a sensitivity analysis of the effect of FPR on expected utility (fig. 11). The curve of expected utility reaches a maximum at FPR = 0.47 and a utility of 0.983. At FPR = 0.47, the TPR is 0.84 as fixed by the use of equation 6 in the decision model.

Discussion

HISTORY AND ORIGIN OF THESE IDEAS

This paper might be thought of as a contribution to meta-analysis of ROC studies. Despite the growing interest in meta-analysis, we found only one other such paper,[13] plus an important letter.[3] But tools for such meta-analysis have long been at hand. Dorfman and Alf[14] gave a method for combining different 2 × 2 tables representing independent studies of a single discrimination task with differing thresholds. That method (maximum likelihood estimation using normal rather than logistic distributions) could be used to address the problems treated here. We have already mentioned plotting sums and differences of logistic transforms as a data-analytic method (Cox,[7] example 6.1), but one not yet in the context of diagnostic testing.

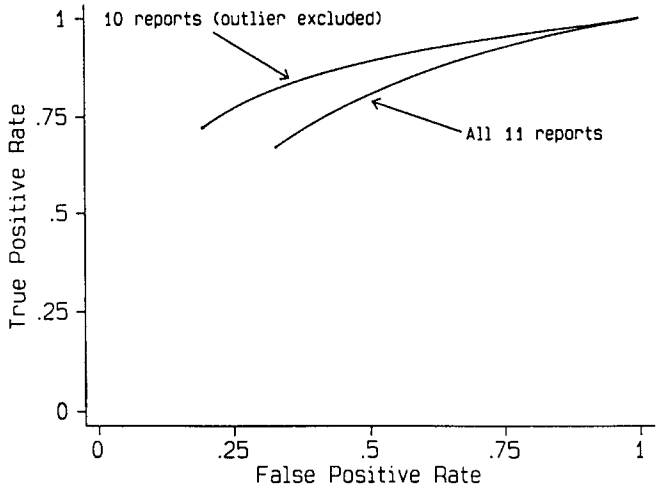

FIGURE 10. Bone scanning data sensitivity analysis. The lower curve is the summary receiver operating curve (SROC) derived from all 11 reports of the accuracy of technetium bone scanning in pedal osteomyelitis. The upper curve is the SROC with the one outlier study eliminated.

CHOICE OF TRANSFORMATION METHOD

We chose to use the logistic transform rather than the probit transform not from considerations of theory, but because we find statistical analysis is then somewhat more straightforward. We agree with Hanley[15] that the choice between the two approaches is hardly likely ever to produce important differences in interpretation. But the logit has one special virtue: kinship of the log-odds ratio to the logistic transform is a real advantage in medical applications because the log-odds ratio enjoys a wide currency in epidemiology and clinical medicine. It would be possible to replace the logit transformation by the probit in our technique, and the outcomes would presumably be very similar. The difference in convenience is clear but not great. When the regression of D upon S is flat, i.e., when b can be taken as zero, then each study's D_i is an estimate of the true average log-odds ratio. Here, use of the logistic transform provides access to a well-understood body of theory that treats odds ratios, and this is an argument of some weight favoring the logistic approach.

UNCERTAINTY

This issue arises in several ways. First, the (asymptotic) standard error of both S and D as calculated from a table with frequencies, a, b, c, d may be taken as:

$$\sqrt{\frac{1}{a + 1/2} + \frac{1}{b + 1/2} + \frac{1}{c + 1/2} + \frac{1}{d + 1/2}}$$

and if the frequencies were all (say) quadrupled, the "standard error of the point" (i.e., of both its two coordinates) would shrink by half. If sample sizes at each study were to grow very large, the uncertainty about each point would disappear.

Second, *maybe* the points would also get closer and closer to a common ROC curve (as a maximum-likelihood analysis presumes). If different studies lay on somewhat different ROC curves, then we would not see the (S, D) points for the studies approach a single ROC curve as the studies grew very large in size. In this latter case we would be confronted not only with binomial variation, but also with a between-studies component of variation. A thorough treatment of these complexities is beyond the scope of this paper, but it can be said that use of the t-test to compare deviations (in two groups) from a single line fitted to the combined groups is a robust procedure, giving useful inferences whether there is or is not a between-studies component of variation.

Third, some linear regression methods will produce standard errors and confidence intervals for both i and b. We have not found these to be very useful, nor do they transform nicely into confidence bands for the SROC. But if taking b as zero is acceptable, then confidence bands for the SROC become available, for in that case each D_k is an estimate of the "true" D corresponding to the collection of studies. Confidence intervals can be based on the sample average and standard deviation of D, or upon the weighted average, or upon the median D_k for which standard nonparametric methods will produce a confidence interval.

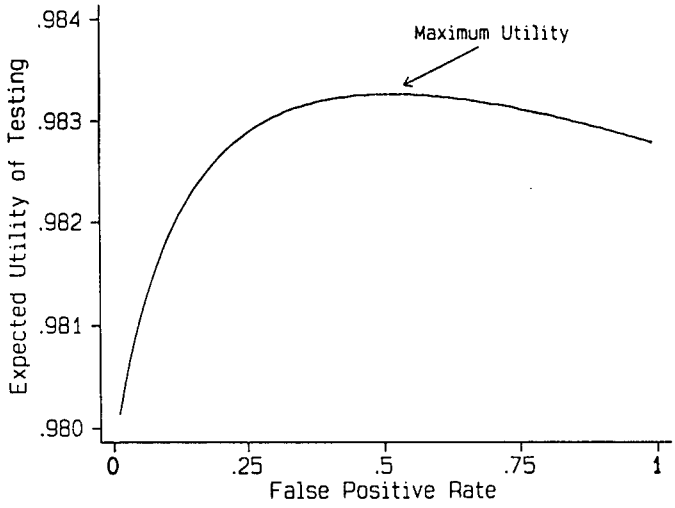

FIGURE 11. Finding the optimum operating point of the bone scan test. The horizontal axis depicts false-positive rate (FPR). For each FPR, a decision model calculated the true-positive rate (TPR) by using equation 6. The model used FPR, TPR, prevalence, and the values of the expected outcomes to calculate the expected utility of testing.

GOODNESS OF FIT AND OUTLIERS

This issue is related to "overdispersion," the existence of a between-studies component of variation, but also includes the problem of identifying and dealing with outliers. This is another large area incapable of full treatment here. Our outlook is pretty well represented in case study 3. The analyses including and excluding one study differed sharply, and brought the question forcefully to the fore. When there appeared a clear substantive explanation of *why a particular value might well be invalid* we felt easy in choosing to omit that point. A more formal treatment would not necessarily be more convincing, especially as it would require choosing unverified assumptions to construct a model in terms of which to conduct the formal analysis.

VERY ACCURATE TESTS

In some clinical settings, we have come upon log-odds ratios much larger than 3.0 (which corresponds to odds ratio 20.0). With high odds ratios (100 or more), zero frequencies become quite common, and the size of the continuity correction (we have offered 1/2 here)

becomes quite influential. We are not satisfied with our approach where zero frequencies are quite common. We expect that recourse to the bootstrap may prove valuable in such situations, and are investigating that possibility.

We have described a method to summarize and analyze multiple reports of the accuracy of a diagnostic test. We choose to represent accuracy as a summary ROC curve, rather than as a point estimate, because we believe that different thresholds are generally in use in different reports, which causes TPR and FPR to covary. Other factors can also induce such covariation. The logistic transform provides a convenient vehicle for applying linear regression methods to a curvilinear problem. The resultant summary ROC curve is useful for summarizing a body of diagnostic accuracy literature, comparing technologies, detecting outliers, and finding the optimum operating point of the test. Extending the logic of meta-analysis to diagnostic testing provides a new tool for technology assessment.

Data for the case examples were graciously provided by Drs. Richard M. Hoffman, Terry A. Hurlbut III, Daniel L. Kent, and Alvin I. Mushlin. Drs. Daniel Rabinowitz and David Shapiro were influential contributors of many of the analytic ideas presented here. Dr. Robert F. Nease provided helpful comments.

References

1. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. Ann Intern Med. 1987;107:224–33.
2. Light RJ, Pillemer DB. Summing Up: The Science of Reviewing Research. Cambridge, MA: Harvard University Press, 1984.
3. Swets JA. Sensitivities and specificities of diagnostic tests. JAMA. 1982;248:548–9.
4. Hurlbut TA, Littenberg B. The diagnostic accuracy of rapid dipstick tests in predicting urinary tract infection. Am J Clin Pathol. 1991;96:582–8.
5. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993;12:1293–316.
6. Littenberg B, Moses LE, Rabinowitz D. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Clin Res. 1990;38:415a.
7. Cox DR. The Analysis of Binary Data. London: Methuen and Co., Ltd., 1970.
8. Hoffman RM, Kent DL, Deyo RA. Diagnostic accuracy and clinical utility of thermography for lumbar radiculopathy. A meta-analysis. Spine. 1991;16:623–8.
9. Littenberg B, Mushlin AI, The Diagnostic Technology Assessment Consortium. Bone scans in the diagnosis of osteomyelitis: a meta-analysis of test performance. J Gen Intern Med. 1992;7:158–64.
10. Swets JA, Pickett RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York: Academic Press, 1982.
11. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. Med Decis Making. 1988;8:279–89.
12. Mushlin AI, Littenberg B. The Diagnostic Technology Assessment Consortium. Diagnosing pedal osteomyelitis: testing choices and their consequences. J Gen Intern Med. 1993 (in press).
13. Kardaun JWPF, Kardaun OJWF. Comparative diagnostic performance of three radiologic procedures for the detection of lumbar disk herniation. Meth Inform Med. 1990;29:12–22.
14. Dorfman DD, Alf EA Jr. Maximum likelihood estimation of signal detection theory—a direct solution. Psychometrika. 1988;33:117–24.
15. Hanley JA. The robustness of the "binormal" assumptions used in fitting ROC curves. Med Decis Making. 1988;8:197–203.
16. Kafadar K. Robust-resistant line. *In* Kotz S, Johnson NL (eds). Encyclopedia of Statistical Sciences. 1988; Volume 8; 169–70.
17. Li G. Robust regression. *In* Hoaglin DC, Mosteller F, Tukey JW (eds). Exploring Data Tables, Trends, and Shapes. New York: John Wiley and Sons, 1985.

Appendix

Example of Calculations

As an example of the calculations required by this method, we demonstrate its application to a previously published data set. Table A shows results of nine studies of myelography used for detecting lumbar disk herniation, as presented by Kardaun and Kardaun.[13] The frequencies of true positives, false negatives, etc., are given. These frequencies are reconstructed from the sample sizes and rates in the Kardaun article. Analysis requires computing the values of S and D from the first four columns. We illustrate the procedure for the top line:

$$Q = \frac{TP + 1/2}{TP + FN + 1} = \frac{80.5}{107} = 0.7523$$

$$V = \ln \frac{Q}{1 - Q} = \ln \frac{0.7523}{0.2477} = 1.111$$

$$P = \frac{FP + 1/2}{FP + TN + 1} = \frac{3.5}{30} = 0.1167$$

$$U = \ln \frac{P}{1 - P} = \ln \frac{0.1167}{0.8833} = -2.024$$

$$S = V + U = 1.111 - 2.024 = -0.91$$

$$D = V - D = 1.111 + 2.024 = +3.14$$

These values appear for Hudgin's study in table A. The nine studies are represented in figure 12 by dots that display their values of S and D.

Now the problem is to estimate a line through the points described by S and D. Various regression techniques are available. Equally weighted least squares is

Table A • Kardaun and Kardaun[13] Myelography Data*

Study	True Positive	False Negative	False Positive	True Negative	S†	D‡	Weight
Hudgins	80	26	3	26	−0.91	3.14	2.68
Macnab	35	2	3	10	1.55	3.75	1.24
Cook	50	2	3	7	2.24	3.77	1.19
Meyenhorst	64	15	4	68	−1.30	4.15	3.16
Claussen	17	4	1	1	1.36	1.36	0.62
Fries	151	23	2	16	−0.02	3.75	1.96
Haughton	28	2	9	16	1.88	2.99	1.66
Jepson	44	5	1	5	0.79	3.39	0.95
Schipper	190	39	10	24	0.73	2.42	6.00

*We omitted four studies from analysis because they report FPR > 0.5, which we consider out of the relevant range. Kardaun and Kardaun excluded one study from their analysis because of apparent bias. We omit it here. For more information about these studies, see Kardaun and Kardaun.[13]

†S = the sum of the two transforms; it is related to how often the test is positive, which is related to the test threshold.

‡D = the difference between the two transforms; it is a measure of how well the test discriminates between the two populations of well and sick subjects.

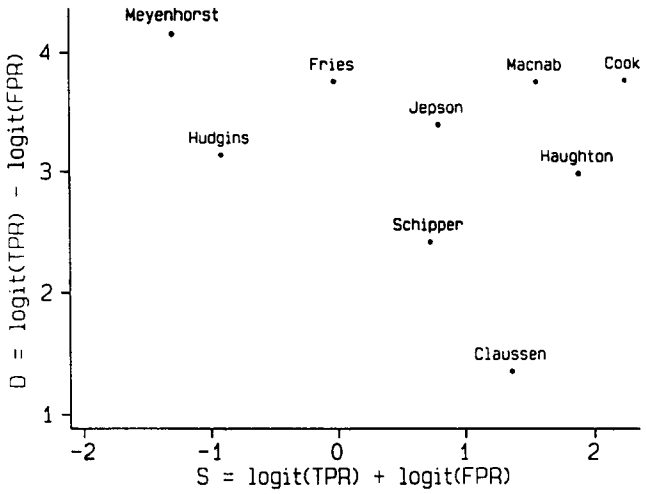

FIGURE 12. Data from myelography studies. The vertical axis represents D, the difference between logit(TPR) and logit(FPR). The horizontal axis represents S, the sum of the two logits. The nine data points (from Kardaun and Kardaun[13]) are displayed with the names of the first authors of the original reports.

readily obtained by many pocket calculators. For these data, $i = 3.32$ and $b = -0.19$.

Another attractive regression technique is called robust resistant fitting.[16,17] It offers protection against outliers and provides these estimates: $i = 3.48$ and $b = -0.26$.

Weighted least squares can be used by finding and then using weights in a weighted regression program. The weights are given in the last column of table A. Each is the reciprocal of the asymptotic variance of D. Again we illustrate the calculation using the first row of data from the table:

$$W = \frac{1}{\frac{1}{80.5} + \frac{1}{26.5} + \frac{1}{3.5} + \frac{1}{26.5}}$$

$$W = \frac{1}{0.3736}$$

$$W = 2.68$$

The weighted regression fitted line has parameters $i = 3.26$ and $b = -0.26$.

The ROC curve in the unit square is constructed from b and i by using equation 6, which for each given value of FPR produces the corresponding value of TPR, once b and i are given. The three ROC curves, obtained by backmapping the three lines, appear in figure 13, together with the FPR and TPR for each of the nine studies.

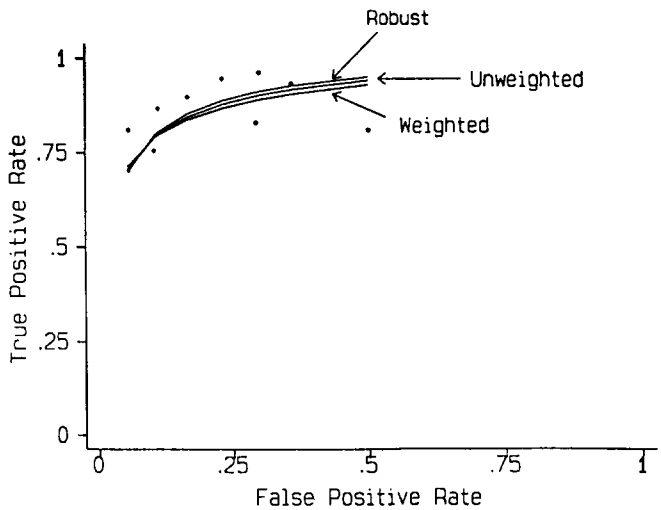

FIGURE 13. Three regression methods compared. The three lines represent back-transformed summary ROC curves for the nine data points in figure 12. The top line was generated using robust regression techniques. The middle line was generated by unweighted least-squares regression. The bottom line is from weighted least-squares regression.