Integrating Stratum-specific Likelihood Ratios with the Analysis of ROC Curves

JOHN C. PEIRCE, MD, MA, MS, RICHARD G. CORNELL, PhD

Data used to construct receiver operating characteristic (ROC) curves and to calculate the area under the curve (ROC AUC) can be used to derive stratum-specific likelihood ratios (SSLRs) with their 95% confidence intervals (95% CIs). The purpose of this study was to determine whether useful information can be obtained by adding SSLRs to the analysis of ROC curves. The authors analyzed four previously reported sets of data: 1) serum creatine kinase (SCK) for diagnosing acute myocardial infarction (AMI) in the coronary care unit (CCU); 2) SCK in the evaluation of chest pain in the emergency center (EC); 3) four predictor variables in the diagnosis of strep throat; and 4) the ordinal assessment of computed tomographic (CT) images. Use of SCK in the CCU produced four strata that had posttest probabilities that were highly discriminating, whereas SCK in the EC resulted in only two strata with limited discriminating ability. In either study the cutpoint at which the SSLR changed from less than to greater than 1.0 was higher than the reported upper normal for the test, thereby quantitating spectrum bias. The maximum number of strata of predictor signs and symptoms for strep throat was three rather than the five used in previous studies. With a larger sample size or pooling, four strata could probably be developed. With CT images, "definitely normal," "probably normal," and "questionable" were collapsed to one negative stratum. "Probably abnormal" became the true "questionable" stratum and "definitely abnormal" was the only positive stratum. The authors conclude that additional useful information is obtained by deriving stratum-specific likelihood ratios as part of the analysis of an ROC curve. *Key words:* diagnosis; decision making; ROC curves; likelihood ratios.
(Med Decis Making 1993;13:141–151)

The receiver operating characteristic (ROC) curve had its origins during World War II, when signal detection theory was applied to radar to *characterize* and evaluate how well a radar *operator* could *receive* or "see" a signal against a noisy background.[1] The ROC curve has since become an important part of a theory of human detection and recognition behavior. It was introduced into clinical medicine by Lusted in the late 1960s, with its first application in radiographic imaging,[2,3] followed by applications in clinical chemistry.[4] During the past ten years the ROC curve has been increasingly used in medical decision making and technology assessment.[5] Construction of an ROC curve is appropriate when there are more than one possible cutoffs between a positive and a negative test, i.e., when a test has a continuous scale, e.g., the serum creatine kinase level in the diagnosis of an acute myocardial infarction,[6,7] or an ordinal scale, e.g., the presence of no, one, two, three or four predictor signs or symptoms in the diagnosis of a strep throat.[8] The ROC curve is described by plotting the sensitivity (true-positive rate) on the y-axis against 1 − specificity (false-positive rate) on the x-axis for each of the several cutoffs. The area under the ROC curve (ROC AUC) represents the probability that the test correctly identifies two subjects as normal or abnormal when one is randomly chosen from the normal group and the other is randomly chosen from the abnormal group.[9,10]

The likelihood ratio (LR) comes from the likelihood function, which is defined as:

$$\text{Likelihood } (H|\text{data}) = K \cdot \Pr(\text{data}|H) \qquad (1)$$

where H is a hypothesis and K is an arbitrary constant. Likelihood differs from probability in that the data are fixed and the hypotheses are variable, whereas probabilities are calculated assuming a fixed hypothesis and random data.[11] Likelihoods do not obey the laws of probability. The likelihood ratio is defined as:

$$\text{Likelihood ratio for } H_0 \text{ vs } H_1|\text{data} = \frac{K \cdot \Pr(\text{data}|H_0)}{K \cdot \Pr(\text{data}|H_1)} \qquad (2)$$

The constants cancel and the likelihood ratio (LR) becomes equivalent to the ratio of the data's probability under one hypothesis compared with that under another hypothesis. When there are only two strata as

Received January 28, 1992, from the Department of Medical Education and Research, Good Samaritan Regional Medical Center, Phoenix, Arizona (JCP); and the Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan (RGC). Revision accepted for publication September 2, 1992. Presented (poster) in part at the thirteenth annual meeting of the Society for Medical Decision Making, October 20, 1991, Rochester, New York.

Address correspondence and reprint requests to Dr. Peirce: Department of Medical Education and Research, Good Samaritan Regional Medical Center, 1111 East McDowell Road, Phoenix, AZ 85006.

in a dichotomous test (T+ and T−) and only two disease states (D+ and D−):

$$LR+ = \frac{Pr(T+|D+)}{Pr(T+|D-)} = \frac{\text{sensitivity}}{1 - \text{specificity}} \quad (3)$$

$$LR- = \frac{Pr(T-|D+)}{Pr(T-|D-)} = \frac{1 - \text{sensitivity}}{\text{specificity}} \quad (4)$$

The LR is used in Bayesian revision of odds[12] where:

$$\text{Prior odds} \times LR = \text{posterior odds} \quad (5)$$

Conversions of odds to probability and probability to odds are performed using the formulas:

$$\text{Probability} = \frac{\text{odds}}{\text{odds} + 1} \quad (6)$$

and:

$$\text{Odds} = \frac{\text{probability}}{1 - \text{probability}} \quad (7)$$

Since the constants in the LR cancel, the ratio can be viewed as a comparison of two binomial proportions. This allows for the construction of confidence intervals about the ratio. The LR+ (equation 3) represents the slope of the line from (0, 0) to the cutpoint, and the LR− (equation 4) represents the slope of the line from the cutpoint to (1.0, 1.0) on the ROC graph.

Integrating Likelihood Ratios with the Analysis of the ROC Curve

Much of the work regarding ROC curves has been done with smoothed curves using the maximum-likelihood estimation to fit the data to two Gaussian curves, one each for the normals and the abnormals,[13] in order to provide the best estimate for the ROC AUC and its standard error. This has led to concerns about the optimal cutoff value, a question of whether to maximize sensitivity or specificity to rule out or rule in a disease. Forcing a single cutoff point creates a LR+ as the slope of a straight line going from (0, 0) to the optimal cutoff value and a LR− as a straight line going from the optimal cutoff value to (1.0, 1.0). The area under this trapezoid, as derived by the nonparametric method,[9] is almost always significantly smaller than the AUC derived by the maximum-likelihood method. Thus, important information is lost with the single forced cutoff. With the nonparametric method one can approximate the ROC AUC derived by maximum likelihood by creating multiple cutpoints and multiple slopes that become stratum-specific likelihood ratios (SSLRs). Several authors have suggested that this approach has merit.[7,14–16] However, with too many cutpoints the slopes become degenerate, so the task is to determine the optimum number of cutpoints and slopes (SSLRs). This paper presents a method to establish optimum cutpoints using SSLRs and their 95% confidence intervals (95% CIs) and examines previously reported studies using this method.

Development of the ROC/SSLR Macro Spreadsheet

We developed a spreadsheet on Microsoft Excel 2.2® for an Apple Macintosh computer to compute the ROC AUC, SSLRs, and their 95% confidence intervals (95% CIs). This was adapted from the method of Centor,[17] who first developed a spreadsheet on Visicalc® (the original spreadsheet) for computing the area under the ROC curve using the nonparametric method described by Hanley and McNeil.[9] The formulas and functions we use are shown in the appendix. Those modified from Centor are in cells B9 through B30. Those that we added pertaining to SSLRs and their 95% CIs are in cells B31 to B34. We tested the accuracy of our formulas and functions by entering the data from Centor[17] and comparing our results with his table 2, and by using the data in table 1 from Hanley and McNeil's published report[9] and comparing our results with theirs.

The SSLR was computed using the formula:

$$SSLR = \frac{\frac{x_{1g}}{n_1}}{\frac{x_{0g}}{n_0}} \quad (8)$$

where:

x_{1g} = the number of people with the abnormality in the gth stratum of the test

n_1 = the total number of people with the abnormality

x_{0g} = the number of people who are normal in the gth stratum of the test

n_0 = the total number of people who are normal

and where:

$\frac{x_{1g}}{n_1}$ = the probability of being in the gth stratum of the test given one has the abnormality

$\frac{x_{0g}}{n_0}$ = the probability of being in the gth stratum of the test given one is without the abnormality

The numbers of normals and abnormals in each stratum are entered in rows 6 and 7 of the spreadsheet, respectively, with the formulas for computing n_0 and n_1 in cells B14 and B15, respectively. Formula 8 is entered into cell B31.

Confidence intervals for likelihood ratios have been proposed by Simel, Samsa, and Matchar.[18] These, after algebraic simplification, result in a variance formula of:

$$\text{Var(SSLR)} = \frac{1}{x_{1g}} - \frac{1}{n_1} + \frac{1}{x_{0g}} - \frac{1}{n_0} \quad (9)$$

This is similar to the first-order variance estimator shown by Gart and Nam,[19] which is nearly unbiased, and is:

$$\text{Var(SSLR)} = \frac{1}{x_{1g} + 0.5} - \frac{1}{n_1 + 0.5} + \frac{1}{x_{0g} + 0.5} - \frac{1}{n_0 + 0.5} \quad (10)$$

This method produces results very close to the iterative Score method, considered to be the optimum method for confidence intervals in comparing two binomial proportions. The Score method is used by Centor in his ROC Analyzer, v.5.2©, to compute 95% CIs for SSLRs.[20] Since the Score method is iterative, we cannot incorporate it into a spreadsheet.

Formula 10 is entered into cell B32.

The 95% confidence interval (95% CI) is calculated by the logit method using the formula:

$$\text{SSLR}_U, \text{SSLR}_L = e^{\ln LR \pm 1.96\sqrt{\text{Var}\ln LR}} \quad (11)$$

Formula 11 is entered into cells B33 and B34.

Like other rates and ratios, the likelihood ratio is bounded by zero and infinity and is skewed to the right. The log transformation in equation 11 is used to create a distribution that approximates a normal one.

A macro of the spreadsheet was developed to allow for rapid setup of the spreadsheet. This allowed for rapid multiple analyses in the determination of the optimum number of cutoffs and the SSLRs.

Analysis of Previously Reported Studies Using the ROC/SSLR Macro Spreadsheet

We evaluated four previously reported studies using the ROC/SSLR Macro Spreadsheet in order to develop and justify the rules for creating the optimum number of strata, to demonstrate the utility of multiple SSLRs by showing how interpretations change using the SSLRs, and to suggest methods for examining pooled data.

USE OF SERUM CREATINE KINASE TO DIAGNOSE MYOCARDIAL INFARCTION IN THE CORONARY CARE UNIT

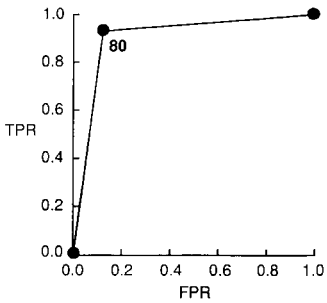
In 1967, Smith reported a study of the peak serum creatine kinase (SCK) in 400 consecutive patients admitted to the coronary care unit (CCU) of the Royal Infirmary in Edinburgh. Three hundred and sixty individuals had data sufficient for analysis, of whom 230 (64%) had the diagnosis of myocardial infarction or possible infarction and 130 had the diagnosis of coronary insufficiency or miscellaneous (non-infarction) diagnoses as determined electrocardiographically.[6] This study is particularly useful because figure 2 of the published report identifies every patient's peak SCK graphically, stratified by diagnosis, which allowed us to look at multiple cutpoints. We looked at one cutpoint with two strata, eight cutpoints with nine strata, four cutpoints with five strata, and three cutpoints with four strata, displayed in figure 1, *A* through *D*, respectively.

With two strata (fig. 1*A*), which is the equivalent of determining the optimal cutpoint, the ROC AUC of 0.90 with a SE of 0.02 is more than two standard errors away from the ROC AUC with four strata of 0.95, demonstrating a significant loss of information. With nine strata (fig. 1*B*), we have an ROC curve that approximates that seen with smoothing; however, the upper five strata have SSLRs that are degenerate, i.e., two are at infinity and the other three are not monotonically related. This leads us to collapse these five into one stratum, producing a curve with five strata (fig. 1*C*). With five strata, the third and fourth strata have SSLRs that approximate each other, leading us to ask: Is there really a difference between stratum 3 and stratum 4, with SSLRs of 2.12 and 3.28, respectively? Since the lower 95% CI in stratum 4 includes the SSLR of stratum 3, and the upper 95% CI of stratum 3 includes the SSLR of stratum 4, we conclude that the SSLRs are not different, so we collapse them into one stratum, leading to an ROC curve with four strata. This is shown in figure 1*D*. Now all of the 95% CIs do not contain the SSLRs of the adjacent strata.

We now take the final step of revising the odds and probabilities of patients in the coronary care unit who have SCKs in each of the four strata using equations 5, 6, and 7. The prior odds of a myocardial infarction is 0.639/0.361 = 1.769. Those with peak SCKs below 40 IU/L have a posttest probability of 2% and those at 160 IU/L or above have a posttest probability of 98%, both near certainty. Those with peak SCKs between 40 and 79 IU/L have a posttest probability of 35%, one that clearly does not rule out a myocardial infarction. Those between 80 and 159 IU/L have a posttest probability of 82%. The upper and lower strata are much more discriminating in ruling in or out an acute myocardial infarction than the middle two.

In his article Smith states:

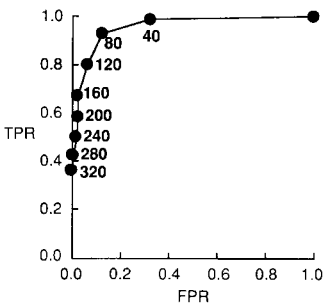
A



Stratum	1 to 79	80+
Normals	114	16
Abnormals	16	214
TPR	0.93	0
FPR	0.12	0
Likelihood Ratio	0.08	7.56
LR Lower 95% CI	0.05	4.81
LR Upper 95% CI	0.12	11.88

AUC	= 0.90	N norm	= 130
SE(AUC)	= 0.02	N Abn	= 230

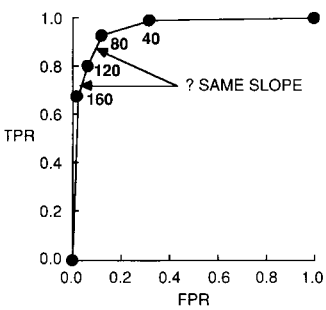
B



Stratum	1 - 39	40 - 79	80 - 119	120 - 159	160 - 199	200 - 239	240 - 279	280 - 319	320+
Normals	88	26	8	5	0	1	1	1	0
Abnormals	2	14	30	29	21	19	18	13	84
TPR	0.99	0.93	0.80	0.67	0.58	0.50	0.42	0.37	0
FPR	0.32	0.12	0.06	0.02	0.02	0.02	0.01	0	0
Likelihood Ratio	0.01	0.30	2.12	3.28	∞	**10.74**	**10.17**	**7.35**	∞
LR Lower 95%CI	0.00	0.17	1.02	1.35	∞	**2.07**	**1.95**	**1.38**	∞
LR Upper 95%CI	0.04	0.56	4.39	7.94	∞	**55.74**	**52.98**	**39.16**	∞

AUC	= 0.96
SE(AUC)	= 0.01

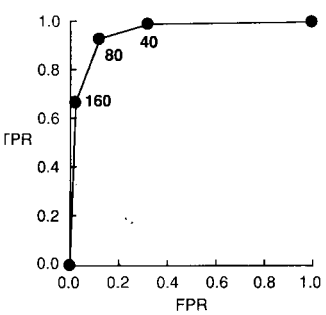
C



Stratum	1-39	40-79	80-119	120-159	160+
Normals	**88**	26	8	5	3
Abnormals	**2**	14	30	29	155
TPR	0.99	0.93	0.8	0.67	0
FPR	0.32	0.12	0.06	0.02	0
Likelihood Ratio	0.01	0.30	**2.12**	**3.28**	29.20
LR lower 95%CI	0.00	0.17	**1.02**	**1.35**	10.35
LR upper 95%CI	0.04	0.56	**4.39**	**7.94**	82.41

AUC	= 0.96
SE(AUC)	= 0.01

D



Stratum	1-39	40-79	80-159	160+
Normals	88	26	13	3
Abnormals	2	14	59	155
TPR	0.99	0.93	0.67	0
FPR	0.32	0.12	0.02	0
Likelihood Ratio	0.01	0.30	2.57	29.20
LR lower 95%CI	0.00	0.16	1.48	10.35
LR upper 95%CI	0.04	0.56	4.45	82.41

AUC	= 0.95	Prevalence	
SE(AUC)	= 0.01	230/360	= 0.639

FIGURE 1. Establishing the optimum number of strata of peak serum creatine kinase levels in IU/L in predicting acute myocardial infarction in patients in a coronary care unit, based on data from Smith's study.[6] *A*, two strata; *B*, nine strata; *C*, five strata; *D*, four strata. TPR = true-positive rate; FPR = false-positive rate; LR lower 95% CI = lower 95% confidence interval of the likelihood ratio; LR upper 95% CI = upper 95% confidence interval of the likelihood ratio; AUC = area under the receiver operator characteristic curve; SE(AUC) = standard error of the area under the receiver operator characteristic curve; N Norm = total number of normals; N Abn = total number of abnormals. Bold values are those with overlap and/or those that are degenerate.

The normal range of S.C.K. was determined in 53 males and 58 females over 25 years old attending a general-practice surgery: those ambulant patients had no serious illness, nor any condition known to be associated with abnormal S.C.K. levels. The means (±S.D.) were: 21.6 (±7.7) I.U. per litre for males and 17.0 (±7.0) I.U. per litre for females. A value of 40 I.U. per litre was taken as the upper limit of normal for both sexes.

In their paper on test bias,[21] Ransohoff and Feinstein used "... the term 'spectrum' to denote the range of features found in patients used to challenge a test's sensitivity and specificity." In this study the SSLRs changed from less than to greater than 1.0 at 80 IU/L. We conclude that the patients in this study, who were consecutive patients studied in a CCU, constitute a more clinically relevant spectrum of patients to challenge the test's sensitivity and specificity than do subjects from a "... general-practice surgery" who "... had no serious illness, nor any condition known to be associated with abnormal S.C.K. levels." The magnitude of the difference between the two "normal limits" is how we quantify spectrum bias. In this study it was 40 IU/L.

When these four SSLRs are coupled with prior probabilities developed from the workup of patients with chest pain in the emergency department using nine clinical and two electrocardiographic findings, as reported by Goldman et al.,[22] the resultant posterior probabilities are quite discriminating. Of 14 patient groups, all with different clinical patterns, six exceeded the threshold probability of 7% for admission to the CCU for further observation. The prior probabilities of acute myocardial infarction in these six groups were 0.11, 0.17, 0.21, 0.26 (2), and 0.77. With a peak SCK less than 40 IU/L, the posterior probabilities would be 0.001, 0.002, 0.003, 0.004, and 0.03, respectively, and with a SCK greater than 160 IU/L the posterior probabilities would be 0.78, 0.86, 0.89, 0.91, and 0.99, respectively. The mid-ranges of SCK are less discriminating, with posterior probabilities for a peak SCK of 40 to 79 IU/L being 0.04, 0.06, 0.07, 0.10, and 0.50, and

Table 1 • Five Strata and Two Strata of Serum Creatine Kinase (SCK) Levels as Predictors of Acute Myocardial Infarction in Patients Seen in the Emergency Center*

	SCK Stratum (IU/L)				
	1–120	121–240	241–360	361–480	480+
Five strata					
Normals	471	201	24	12	14
Abnormals	23	6	7	6	9
TPR	0.55	0.43	**0.29**	0.18	0
FPR	0.35	0.07	**0.04**	0.02	0
Likelihood ratio	**0.69**	**0.42**	**4.13**	**7.08**	**9.10**
LR lower 95% CI	**0.51**	**0.20**	**1.91**	**2.86**	**4.22**
LR upper 95% CI	**0.94**	**0.88**	**8.90**	**17.49**	**19.61**

AUC	= 0.66	N norm	= 722
SE(AUC)	= 0.05	N abn	= 51

	Stratum	
	1–240	241+
Two strata		
Normals	672	50
Abnormals	**29**	22
TPR	**0.43**	0
FPR	**0.07**	**0**
Likelihood ratio	**0.61**	**6.23**
LR lower 95% CI	**0.48**	**4.14**
LR upper 95% CI	**0.77**	**9.37**

AUC	= 0.68	Prevalence	
SE(AUC)	= 0.05	$\frac{51}{773}$	= 0.07

TPR = true positive rate; FPR = false positive rate; LR lower 95% CI = lower 95% confidence interval of the likelihood ratio; LR upper 95% CI = upper 95% confidence interval of the likelihood ratio; AUC = area under the receiver operator characteristic curve; SE(AUC) = standard error of the area under the receiver operator characteristic curve; N Norm = total number of normals; N Abn = total number of abnormals. Bold values are those with overlap and/or those which are degenerate.

*Based on data from the study of Radock et al.[7]

Table 2 • Five Strata and Three Strata of Predictor Variables in the Prediction of Strep Throat*

	Stratum (No. of Predictor Variables)				
	0	1	2	3	4
Five strata					
Normals	35	68	49	29	12
Abnormals	2	3	8	16	12
TPR	0.95	0.88	0.68	0.29	0
FPR	0.82	0.47	0.21	0.06	0
Likelihood ratio	0.27	0.21	0.77	2.60	4.71
LR lower 95% CI	0.08	0.07	0.40	1.57	2.32
LR upper 95% CI	0.93	0.58	1.47	5.28	9.56

AUC	= 0.78	N norm	= 193
SE(AUC)	= 0.04	N abn	= 41

	Stratum		
	0, 1	2	3, 4
Three strata			
Normals	103	49	41
Abnormals	5	8	28
TPR	0.88	0.68	0
FPR	0.48	0.21	0
Likelihood ratio	0.23	0.77	3.21
LR lower 95% CI	0.10	0.40	2.29
LR upper 95% CI	0.50	1.47	4.51

AUC	= 0.78
SE(AUC)	= 0.04

TPR = true positive rate; FPR = false positive rate; LR lower 95% CI = lower 95% confidence interval of the likelihood ratio; LR upper 95% CI = upper 95% confidence interval of the likelihood ratio; AUC = area under the receiver operator characteristic curve; SE(AUC) = standard error of the area under the receiver operator characteristic curve; N Norm = total number of normals; N Abn = total number of abnormals. Bold values are those with overlap and/or those which are degenerate.

*Based on data from the study of Centor et al.[8] Predictor variables were: none and one to four of a specific set of four signs and symptoms.

for a peak SCK of 80 to 159 IU/L being 0.24, 0.34, 0.41, 0.47, and 0.90.

USE OF SERUM CREATINE KINASE TO DIAGNOSE MYOCARDIAL INFARCTION IN THE EMERGENCY CENTER

Radack, Rouan, and Hedges reported using likelihood ratios for SCKs to better discriminate patients with possible myocardial infarction seen in the emergency center (EC).[7] They recommended five strata with SSLRs as shown in table 1, rather than the cutoff of 120 IU/L used by the laboratory at their institution as the upper limit of normal. Before calculating the 95% CIs for the SSLRs (which they did not provide), we can see that the lower two SSLRs are degenerate, since they are not monotonic. After collapsing these two strata into one lower stratum, its upper 95% CI (0.77) does not include 1.0. We collapsed the SSLRs of the upper two strata because they approximate each other, leaving three strata with the SSLRs (95% CIs) of the mid and upper strata of 4.13 (1.91, 8.90) and 8.17 (4.67, 14.29), respectively. Because of the small number of subjects in the middle stratum, its upper 95% CI included the SSLR of the upper stratum. With more subjects in the middle stratum, the 95% CIs would probably narrow such that three distinct strata could be maintained. However, all we can confidently say with this information is that there are two strata with the cutoff at 240 IU/L. Since the SSLRs changed from less than to greater than 1.0 at 240 IU/L, this, rather than 120 IU/L used by their laboratory, constituted the upper limit of normal for this test in that setting, creating a spectrum bias of 120 IU/L.

When looking at the odds and probability revision with the reported prevalence for acute myocardial infarction of 6.6% (prior odds = 0.071) in this study, a negative test adds little, with the probability for acute myocardial infarction decreasing only to 4%. With a positive test, the probability for acute myocardial infarction increases to 31%. Most importantly, the prevalence of acute myocardial infarction in the EC is substantially lower than that in the CCU (6.6% vs 63.9%).

Table 3 • Five Strata and Three Strata of Ratings of Computed Tomographic Images in Predicting Abnormalities*

	Stratum				
	Definitely Normal	Probably Normal	Questionable	Probably Abnormal	Definitely Abnormal
Five strata					
Normals	33	6	2	11	2
Abnormals	3	2	2	11	33
TPR	**0.94**	0.90	**0.87**	0.65	0
FPR	**0.43**	0.33	**0.22**	0.03	0
Likelihood ratio	**0.10**	**0.38**	**0.38**	1.14	18.76
LR lower 95% CI	**0.04**	**0.09**	**0.09**	0.55	5.48
LR upper 95% CI	**0.29**	**1.55**	**1.55**	2.35	64.15

AUC	= 0.89	N norm	= 58
SE(AUC)	= 0.03	N abn	= 51

	Stratum		
	≤Questionable	Probably Abnormal	Definitely Abnormal
Three strata			
Normals	45	11	2
Abnormals	7	11	33
TPR	0.86	0.65	0
FPR	0.22	0.03	0
Likelihood ratio	0.18	1.14	18.76
LR lower 95% CI	0.09	0.54	5.48
LR upper 95% CI	0.35	2.35	64.15

AUC	= 0.88
SE(AUC)	= 0.04

TPR = true positive rate; FPR = false positive rate; LR lower 95% CI = lower 95% confidence interval of the likelihood ratio; LR upper 95% CI = upper 95% confidence interval of the likelihood ratio; AUC = area under the receiver operator characteristic curve; SE(AUC) = standard error of the area under the receiver operator characteristic curve; N Norm = number of normals; N Abn = number of abnormals. Bold values are those with overlap and/or those which are degenerate.

*Based on data from the study of Hanley and McNeil.[9]

The marked difference in prevalences, the substantially lower ROC AUC with only two strata for SCK in the EC compared with four in the CCU, shows the importance of defining the clinical population and the purpose for which the test is being used. In this case, one cannot "transport" the test from one clinical population to the other.

PREDICTOR VARIABLES IN THE DIAGNOSIS OF STREP THROAT

Centor et al. studied signs and symptoms that could be used to predict strep throat in the emergency department setting. Using logistic regression, they showed that four signs and symptoms, tonsillar exudates, swollen tender anterior cervical nodes, lack of a cough, and a history of fever, predicted a positive throat culture for group A β-hemolytic streptococcus.[8] Centor, in developing the Visicalc® program,[17] showed that the area under the ROC curve using zero through four predictor variables was 0.7797, with a standard error of 0.0404. In table 2, we replicate his analysis with five strata and their SSLRs and 95% CIs. Zero or one positive predictor variable clearly has a SSLR less than 1.0, and thereby constitutes a negative test. Since the SSLRs of these lower two strata were not monotonically related and the 95% CIs of each stratum easily included the SSLR of the other stratum, these two were collapsed to one stratum. Two predictor variables each had a SSLR of 0.78 with 95% CIs that included 1.0, indicating that this was an indeterminate stratum for which we assigned the SSLR a value of 1.0. The 95% CIs of two findings did not include the SSLR of either the stratum below or the stratum above, so this remained as a separate stratum.

Three and four positive variables had 95% CIs that encompassed the SSLRs of each other, but not by much. It may be that with a greater number of subjects these would have constituted distinctly different strata. When we collapse the third and fourth strata into one, we have a stratum whose 95% CIs do not overlap with the SSLR of the adjacent strata, as depicted in table 2.

In revising the odds and probabilities, the probability goes from 18% to 5% with no or one predictor

variable, while it remains the same with two predictor variables. With three or four predictor variables the posterior probability of a strep throat is 41%. If increased cases allowed for two distinct positive strata (three and four clinical findings) with SSLRs of 2.60 and 4.71, then the posterior probabilities would be 36% with three predictor variables and 50% with four predictor variables.

INTERPRETING CT IMAGES

Hanley and McNeil[9] used a single radiologist's interpretation of CT images to demonstrate the nonparametric estimate of the area under the ROC curve in table 1 of their published report. Their original data are shown in table 3 under "five strata," to which we have added the SSLRs and their 95% CIs. "Definitely normal," "probably normal," and "questionable" all had likelihood ratios below 1.0, but because of the small numbers of subjects in the latter two categories, the upper 95% CIs included 1.0. The 95% CIs of all of these three strata included the SSLRs of the other two, so we collapsed all three into one stratum. This collapsed lower stratum became the "normal" reading, with a SSLR below 1.0 and a sufficient number of cases to create an upper 95% CI that did not include 1.0. "Probably abnormal" became the true "questionable" and was assigned a value of 1.0, since its SSLR was 1.14 and had 95% CIs that included 1.0. The "positive" reading was when the radiologist read the CT as "definitely abnormal." With three strata there are no 95% CIs that include the adjacent SSLRs.

Discussion

In the process of producing a smoothed ROC curve to improve the estimation of the AUC for a test, investigators can lose sight of the discriminating power of that test. Feinstein notes that increasing categories is a method of increasing a test's discriminating power.[23] In the Smith study of SCK in the CCU, the optimum cutpoint on a smoothed ROC curve would have been approximately 80 IU/L where the tangent to the curve equaled 1.0. Had we used this to create a dichotomous negative (below 80 IU/L) and positive (at or above 80 IU/L) result, we would have had only two likelihood ratios of 0.008 and 7.56, respectively (see fig. 1*A*), requiring us to interpret a SCK of 100 IU/L the same as one of 300 IU/L. Intuitively clinicians know that an acute myocardial infarction is more likely with a higher SCK. The task then is to quantify this qualitative discriminatory judgment. Creating stratum-specific likelihood ratios as shown here is a defensible method. In constructing decision trees, the decision analyst could now use four branches instead of two, depending upon the result of the SCK.

We propose a method of determining an optimum number of strata by calculating likelihood ratios specific to different strata (SSLRs) along with their 95% confidence intervals. Although the likelihood function is not a probability and does not obey the rules of probability, the likelihood ratio is calculated using two binomial probabilities with the same constant in the numerator and denominator as shown in equation 2. In each stratum the two probabilities are the "true-positive rate" (actually the true-positive *proportion*) and the "false-positive rate," the two proportions used in constructing the ROC curve. To achieve the optimum number of strata we have developed the following rules based upon the analyses in this paper:

1. Provide sufficient abnormal and normal cases in each stratum to allow the SSLRs to be monotonically related.

2. Collapse those strata where the SSLRs are close to one another and their 95% CIs easily overlap.

3. When the SSLRs appear to be clinically different and the overlaps of the 95% CIs with their adjacent SSLRs are small, consider performing a new study with a larger sample size or pooling similar studies to see whether data support the increased number of strata.

4. An indeterminate stratum, i.e., one that includes 1.0 within the 95% CIs, should be assigned an SSLR of 1.0, i.e., having no information. These are more common when the test has an ordinal scale defining the strata, but are also possible with tests using a continuous scale to delineate the strata.

Gart and Nam[19] reviewed issues of bias and skewness of confidence intervals on ratios of two binomial proportions. They conclude that the Score method, requiring iterative computations, is the least biased and least likely to produce degenerative results. The logit method described in our equations 10 and 11 is the best of the noniterative methods. Centor uses the Score method in his ROC Analyzer, v.5.2©. Using exact confidence intervals on odds ratios computed with StatXact©, and transforming these to the corresponding limits on likelihood ratios according to the approach of Thomas and Gart,[24] we compared the logit 95% CIs, Score 95% CIs, and exact 95% CIs in table 4. The differences between these three methods are small. Because the discreteness of the data leads to confidence limits equal to or greater than the nominal confidence limits, the exact method is the most conservative (wider interval) for both the lower and the upper interval. For the lower interval the logit method is the least conservative; for the upper interval the Score method is the least conservative, showing mild skewness to the right of the logit CIs relative to the Score

Table 4 • Comparison of Logit, Score, and Exact 95% Confidence Intervals for Strata of Predictor Variables in Four Studies

Test and Stratum	Stratum-specific Abnormals (x_1)	Stratum-specific Normals (x_0)	Total Abnormals (n_1)	Total Normals (n_0)	Stratum-specific Likelihood Ratio	Logit $95CI_L$	Logit $95CI_U$	Score $95CI_L$	Score $95CI_U$	Exact $95CI_L$	Exact $95CI_U$
Coronary care unit study[6]											
$SCKCC_{1-79}$	16	114	230	130	0.08	0.05	0.12	0.05	0.13	0.05	0.12
$SCKCC_{80+}$	214	16	230	130	7.56	4.81	11.88	4.87	12.07	5.19	11.16
$SCKCC_{1-39}$	2	88	230	130	0.01	0.00	0.04	0.00	0.05	0.00	0.05
$SCKCC_{40-79}$	14	26	230	130	0.30	0.17	0.56	0.17	0.56	0.15	0.58
$SCKCC_{80-119}$	30	8	230	230	2.12	1.02	4.39	1.03	4.45	0.98	5.22
$SCKCC_{120-159}$	29	5	230	130	3.28	1.35	7.94	1.36	8.08	1.29	10.64
$SCKCC_{80-159}$	59	13	230	130	2.57	1.48	4.45	1.49	4.50	1.46	4.92
$SCKCC_{160+}$	155	3	230	130	29.20	10.35	82.41	10.23	85.80	10.23	139.2
Emergency room study[7]											
$SCKER_{1-120}$	23	471	51	722	0.69	0.51	0.94	0.49	0.90	0.48	0.92
$SCKER_{121-240}$	6	201	51	722	0.42	0.20	0.88	0.20	0.85	0.16	0.87
$SCKER_{241-360}$	7	24	51	722	4.13	1.91	8.90	1.87	8.70	1.55	9.23
$SCKER_{361-480}$	6	12	51	722	7.08	2.86	17.49	2.81	17.17	2.24	19.26
$SCKER_{480+}$	9	14	51	722	9.10	4.22	19.61	4.15	19.30	3.61	21.10
$SCKER_{1-240}$	29	672	51	722	0.61	0.48	0.77	0.46	0.75	0.47	0.76
$SCKER_{240+}$	22	50	51	722	6.23	4.14	9.37	4.05	9.22	3.89	9.29
$SCKER_{360+}$	15	26	51	722	8.17	4.67	14.29	4.57	14.70	4.26	14.60
Strep throat study[8]											
$Strep_0$	2	35	41	193	0.27	0.08	0.93	0.07	0.92	0.03	0.98
$Strep_1$	3	68	41	193	0.21	0.07	0.58	0.07	0.56	0.04	0.59
$Strep_2$	8	49	41	193	0.77	0.40	1.47	0.39	1.42	0.33	1.48
$Strep_3$	16	29	41	193	2.60	1.57	5.28	1.53	4.22	1.44	4.33
$Strep_4$	12	12	41	193	4.71	2.32	9.56	2.28	9.48	2.07	10.36
$Strep_{0,1}$	5	103	41	193	0.23	0.10	0.50	0.10	0.49	0.08	0.50
$Strep_{3,4}$	28	41	41	193	3.21	2.29	4.51	2.25	4.49	2.20	4.33
Computed tomography study[9]											
$CT_{DefNorm}$	3	33	51	58	0.10	0.04	0.29	0.03	0.29	0.02	0.30
$CT_{PrbNrm\text{-}Quest}$	2	6	51	58	0.38	0.09	1.55	0.09	1.56	0.04	2.01
$CT_{ProbAbn}$	11	11	51	58	1.14	0.55	2.35	0.55	2.36	0.49	2.64
CT_{DefAbn}	33	2	51	58	18.76	5.48	64.15	5.42	68.87	5.28	148.19
$CT_{DefNrm\text{-}Quest}$	7	45	51	58	0.18	0.09	0.35	0.09	0.34	0.08	0.34

$SCKCC_{1-79}$ = serum creatine kinase of 1 through 79 IU/L in the coronary care unit study; subscripts for SCKCC denote the range of values for the test in that stratum. $SCKER_{1-120}$ = serum creatine kinase of 1 through 120 IU/L in the emergency room study; subscripts for SCKER denote the range of values for the test in that stratum. $Strep_0$ = the stratum with no predictor variable for strep throat; one subscript denotes the number of predictor variables in that stratum. $Strep_{0,1}$ = the stratum that contains zero or one predictor variable for strep throat. $Strep_{3,4}$ = the stratum that contains three or four predictor variables for strep throat. $CT_{DefNorm}$ = computed tomography read as definitely normal. $CT_{ProbNorm\text{-}Quest}$ = computed tomography read as probably normal or questionable. $CT_{ProbAbn}$ = computed tomography read as probably abnormal. CT_{DefAbn} = computed tomography read as definitely abnormal. $CT_{DefNorm\text{-}Quest}$ = computed tomography read as definitely normal, probably normal, or questionable. $95CI_L$ = lower 95% confidence interval for that stratum-specific likelihood ratio. $95CI_U$ = upper 95% confidence interval for that stratum-specific likelihood ratio.

CIs. The greatest disparity in the three methods occurred in the upper CI with the exact method when the SSLR was uppermost and large with very few normal cases in that stratum, e.g., $SCKCC_{160+}$ and CK_{DefAbn}. With both SSLRs, the 95% CIs of the logit and Score methods were much closer to each other. From this comparison we conclude that the logit and Score methods are comparable. We found no evidence of the logit method's becoming degenerate as reported by Gart and Nam.

Most laboratories establish normal limits for a test by determining the distribution of values for that test in a sample of subjects known to be free of disease, i.e., "normal subjects," taking the upper limit of normal at the 95th or 99th upper confidence limit. However, such a sample will not challenge the test's sensitivity and specificity in a variety of diseases and conditions where it will be used because the sample does not contain patients who have the different but commonly confused disorders needed for such a challenge.[14] This is a common problem with many tests and points up the importance of studies such as Smith's, where consecutive patients being evaluated for a given condition or disease comprised the sample used for evaluating

the test. Smith's "no-disease" group contained patients who had the different but commonly confused disorders needed to challenge the test's sensitivity and specificity. When there is a difference between the "upper normal" determined by conventional means and the cutpoint where the likelihood ratios change from below to above 1.0, we choose the latter as the "true upper normal" and the magnitude of the difference to be the quantity of the spectrum bias.

We saw a spectrum bias of 120 IU/L in the Radock, Rouan, and Hedges study of SCK in the EC. Aside from the scaling differences, there were several other important features that differed from those in Smith's study of SCK in the CCU. The settings and prevalences were clearly different: one in the EC with a prevalence of 0.07, the other in a CCU with a prevalence of 0.64. The ROC AUCs—0.95 for the CCU and 0.68 for the EC; the magnitudes of the upper LR—29.2 for the CCU and 6.23 for the EC—and lower LR—0.01 for the CCU and 0.61 for the EC; and the numbers of strata (two for the EC and four for the CCU) were strikingly different. We would be ill-advised to try to pool the results of these studies. Not only are the settings clearly different, with different actions resulting from a diagnosis, but the test characteristics are very different, as noted above.

The number of strata, which defines the discriminating power of the test, is a function of the ROC AUC, the number of normal subjects, and the number of abnormal subjects. In the strep throat study, the third and fourth strata had distinctly different SSLRs, with only the lower 95% CI of the fourth stratum containing the SSLR of the third stratum. Using the method described by Simel, Samsa, and Matchar,[18] a sample size of 290 normals and 62 abnormals would be needed to tighten the lower 95% limit for stratum 4 to 2.60 while maintaining a SSLRs of 2.57 and 4.71 in strata 3 and 4, respectively. Achieving this sample size would require a repeat study after estimating the prevalence of strep throat in the community, or an evaluation of other similar studies to see whether results could be pooled to achieve sufficient power.

Wigton, Connor, and Centor[25] and Poses et al.[26] showed that the decision rule initially proposed by Centor could be "transported" to other similar settings with as much variation in prevalence as 26% in the former study and 5% in the latter. The latter authors used likelihood ratios to deal with the variability in the prevalence of the disease. Of note are the similarities in their ROC AUCs, the settings in which the rules were applied, and the actions resulting from the test. These studies suggest that requirements for pooling of data that need to be met are the similarities of the ROC AUCs, the settings, and the actions taken because of the results of the test. To these we would add the similarities of the numbers of strata and their SSLRs, and the cutpoints at which the SSLRs change from less than to greater than 1.0.

Conclusion

Using the information needed to construct an ROC curve and calculate the area under the curve, one can compute likelihood ratios and their confidence intervals on a spreadsheet for each of several strata. We have set forth rules using 95% CIs for SSLRs to determine the optimum number of strata. In examining four previously reported studies with this method, new levels of normal were defined, as well as the discriminating power of the test. We have also compared several methods of calculating 95% CIs, discussed methods for sample size determination when SSLRs appeared to be clinically different but had overlapping 95% CIs, and suggested rules for pooling studies to achieve sufficient statistical power.

The authors thank Dr. Robert M. Centor for helpful suggestions and Dr. Paul Stander for review of the manuscript.

References

1. Lusted LB. ROC recollected. Med Decis Making. 1984;4:131–5.
2. Lusted LB. Introduction to medical decision making. Springfield, Illinois: Charles C Thomas, 1968.
3. Goodenough DJ, Rossman K, Lusted LB. Radiographic applications of receiver operating characteristic (ROC) curves. Radiology. 1974;110:89–96.
4. Galen RS, Gambino SR. Beyond normality: the predictive value and efficiency of medical diagnoses. New York: John Wiley and Sons, 1975.
5. Centor RM. Signal detectability: the use of ROC curves and their analyses. Med Decis Making. 1991;11:102–6.
6. Smith AF. Diagnostic value of creatine-kinase in a coronary care unit. Lancet. 1967;2:178–82.
7. Radack KL, Rouan G, Hedges J. The likelihood ratio: an improved measure for evaluating diagnostic test results. Arch Pathol Lab Med. 1986;110:689–93.
8. Centor RM, Witherspoon JM, Dalton HP, et al. The diagnosis of strep throat in adults in the emergency room. Med Decis Making. 1981;1:239–46.
9. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
10. Bamber D. The area above the ordinal dominance graph and below the receiver operating characteristic graph. J Math Psychol. 1975;12:387–415.
11. Goodman SN, Royall R. Evidence and scientific research. Am J Public Health. 1988;78:1568–74.
12. Weinstein MC, Fineberg HV, Elstein AS, et al. Clinical decision analysis. Philadelphia: W. B. Saunders, 1980.
13. Dorfman DD, Alf E Jr. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. J Math Psychol 1969;6:487–96.
14. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown, 1985.
15. Thornbury JR, Fryback DG, Edwards W. Likelihood ratios as a measure of usefulness of excretory pyelograms. Radiology. 1975;114:561–5.
16. Albert A. On the use and computation of likelihood ratios in clinical chemistry. Clin Chem. 1982;28:1113–9.
17. Centor RM. A Visicalc program for estimating the area under a receiver operating characteristic (ROC) curve. Med Decis Making. 1985;5:139–48.

18. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol. 1991;44:763–70.
19. Gart JJ, Nam J. Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. Biometrics. 1988;44:323–38.
20. Centor RM. The ROC Analyzer, v.5.2©, Reference Guide and personal communication.
21. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926–30.
22. Goldman I, Cook EF, Brand DA, et al. A computer protocol to predict myocardial infarction in emergency department patients with chest pain. N Engl J Med. 1988;318:797–803.
23. Feinstein AR. Clinimetrics. New Haven, Connecticut: Yale University Press, 1987.
24. Thomas DG, Gart JJ. A table of exact confidence limits for differences and ratios and their odds ratios. J Am Stat Assoc. 1977;72:73–6.
25. Wigton RS, Connor JL, Centor RM. Transportability of a decision rule for the diagnosis of streptococcal pharyngitis. Arch Intern Med. 1986;146:81–3.
26. Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules: the case of streptococcal pharyngitis. Ann Intern Med. 1986;105:586–91

Appendix

Spreadsheet Formulas and Functions

	A	B	C
1			
5	Stratum		
6	Normals		
7	Abnormals		
8			
9	Abn>x	=SUM(C7:$Z$7) ====>	Fill right to Z
10	Norm<x	=SUM($A$6:A6) ====>	Fill right to Z
11	Mann-Whit	=(B6*B9)+((B6*B7)/2) ====>	Fill right to Z
12	Q1 num	=B7*(B10^2+(B10*B6)+(B6^2/3))====>	Fill right to Z
13	Q2 num	=B6*(B9^2+(B9*B7)+B7^2/3) ====>	Fill right to Z
14	N norm	=SUM(B6:Z6)	
15	N abn	=SUM(B7:Z7)	
16	N norm-1	=B14-1	
17	N abn-1	=B15-1	
18	AUC	=(SUM(B11:Z11))/(B14*B15)	
19	1-AUC	=1-B18	
20	Q1	=SUM(B13:Z13)/(B14*B15^2)	
21	Q2	=SUM(B12:Z12)/(B14^2*B15)	
22	Q1-SQ(AUC)	=B20-B18^2	
23	Q2-SQ(AUC)	=B21-B18^2	
24	VAR(AUC)	=(((B18*B19)+(B17*B22)+(B16*B23))/ B14)/B15	
25			
26	AUC	=B18	
27	SE(AUC)	=SQRT(B24)	
28			
29	TPR	=SUM(C7:$Z$7)/$B$15 ====>	Fill right to Z
30	FPR	=SUM(C6:$Z$6)/$B$14 ====>	Fill right to Z
31	Likelihood Ratio	=(B7/$B$15)/(B6/$B$14) ====>	Fill right to Z
32	VAR (lnLR)	=(1/(B7+0.5))-(1/($B$14+0.5)) +(1/(B6+0.5))-(1/($B$15+0.5)) ===>	Fill right to Z
33	LR lower 95%CI	=EXP(LN(B31)-1.96*(SQRT(B32)))====>	Fill right to Z
34	LR upper 95%CI	=EXP(LN(B31)+1.96*(SQRT(B32)))====>	Fill right to Z