STATISTICS IN MEDICINE, VOL. 8, 609–618 (1989)

ASSESSING DIAGNOSTIC TESTS BY A STRICTLY PROPER SCORING RULE

KRISTIAN LINNET

Department of Clinical Chemistry KK 4051, University of Copenhagen, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark

SUMMARY

Evaluation of univariate quantitative diagnostic tests by strictly proper scoring rules is considered as an alternative to the traditional error rate measures. In principle, the posterior probability of disease as a function of the test value is estimated from training observations, and subsequently the score is assessed on a set of test samples. The same subjects may serve as training and test samples when the bootstrap procedure is applied for estimation of standard errors and correction of bias. The method is demonstrated using serum bile acids and bilirubin in patients with liver disease. The power for comparison of scores from two tests is compared with that from error rate measures for some typical situations.

KEY WORDS Bootstrap Clinical chemistry Diagnosis Scoring rules Bilirubin Bile acids Screening

INTRODUCTION

Laboratory diagnostic tests commonly yield results distributed on a continuous scale, that is they are quantitative. However the methodology for assessing diagnostic tests has been developed in the context of binary tests, using Yerushalmy's concepts of specificity and sensitivity.[1] Specificity is defined as the proportion of reference (healthy) subjects who have a negative test, and sensitivity as the proportion of subjects with disease who have a positive test. By dichotomizing a continuous scale into negative and positive test results using some cutoff limit, the terms specificity and sensitivity have been applied to quantitative tests.

In simple situations, where diagnosis or management is based solely upon a clinical chemistry test result, the error rate measures are appropriate for test evaluation. An example is the screening of a population using a laboratory test. Here the test result is referred to a discrimination limit, and a decision whether or not to institute follow-up is taken without regard to the deviation from the cutoff. In routine clinical practice a test result is interpreted differently. When confronted with a borderline test result just exceeding a cutoff limit, for example an upper limit of a reference interval, the physician gathers additional information to confirm or refute a specific diagnosis; an extreme result may establish the diagnosis with certainty. When a diagnostic test is used as a decision support as outlined here, it is sensible to convert the test results to posterior probabilities of disease and assess the validity of the test by some aggregate measure. An appropriate methodology is the family of strictly proper scoring rules developed in the field of subjective probabilities.[2,3] This has been introduced in medicine, but has attained only sparse application.[4–6] A brief introduction to the application of scoring rules in clinical chemistry has been

0277–6715/89/050609–10$05.00

Received February 1988
Revised December 1988

presented previously.[7] Scoring rules can be applied to all diagnostic systems that provide a probability conditional on test results and so may be used in univariate and multivariate settings. In this paper, I examine the statistical aspects of a strictly proper scoring rule applied to a univariate quantitative test for discrimination between a reference and a diseased population. A resampling technique, the bootstrap method, is used for estimation of standard errors and for correction of bias. The power for comparison of two tests is compared with that of the error rate measures.

STRICTLY PROPER SCORING RULES

A scoring rule is a function of the assigned probabilities of belonging to the population of origin and the actual population of origin for the subjects tested. A strictly proper scoring rule (SPRS) is characterized by having a maximum expected value when the true probabilities of population assignment are used.[5] Examples are the quadratic and the logarithmic scoring rules, which for the case of two populations are given by

$$Q_{\mathrm{qu}\,i} = 1-(1-P_i)^2$$

and

$$Q_{\ln\,i} = \ln(P_i),$$

where P_i is the assigned probability of belonging to the population of origin for the ith individual; that is, $P_i = P(R|x_i)$ for a subject belonging to the reference population (R), and $P_i = P(D|x_i)$ for a subject with disease (D), where x_i is the test result. Figure 1 depicts the relation between this probability and the two scores. The logarithmic score has been truncated at 0·01 and rescaled to the interval [0; 1].[8]

A simple example illustrates the difference between a naive scoring rule such as $Q_{\mathrm{l}i} = P_i$ and the strictly proper scoring rules. Suppose that 'no test' is applied to a sample of 1000 subjects from a population in which the prevalence of disease $P(D)$ is known to be 0·6, and that no further information is available. Ignoring sampling fluctuations, the evaluation sample consists of 400 reference and 600 diseased subjects. An investigator, who correctly assigns $P(D)$ a value of 0·6 (and so $P(R) = 0{\cdot}4$) to each subject, attains the following average scores:

$$Q_{\mathrm{l}} = [1/(400+600)]\,\{400\times 0{\cdot}4 + 600\times 0{\cdot}6\} = 0{\cdot}52$$
$$Q_{\mathrm{qu}} = [1/(400+600)]\,\{400[1-(1-0{\cdot}4)^2] + 600[1-(1-0{\cdot}6)^2]\} = 0{\cdot}76$$
$$Q_{\ln} = [1/(400+600)]\,\{400\ln 0{\cdot}4 + 600\ln 0{\cdot}6\} = -0{\cdot}673.$$

If the investigator had erroneously assigned $P(D)$ the value 1·0 (and $P(R) = 0$), he would have obtained the following scores:

$$Q_{\mathrm{l}} = [1/(400+600)]\,\{400\times 0 + 600\times 1\} = 0{\cdot}60$$
$$Q_{\mathrm{qu}} = [1/(400+600)]\,\{400[1-(1-0)^2] + 600[1-(1-1)^2]\} = 0{\cdot}6$$
$$Q_{\ln} = [1/(400+600)]\,\{400\ln 0 + 600\ln 1\} = -\infty.$$

Thus, by contrast to the strictly proper scoring rules, the naive scoring rule does not have maximum value corresponding to the true probabilities of population assignment, and so is inappropriate.

Evaluation of a diagnostic test by a scoring rule begins with estimation of the relation between the posterior probability and the test result x, that is $\hat{P}(D|x) = 1-\hat{P}(R|x)$. A set of training observations from the respective populations are sampled, and from the frequency distributions of

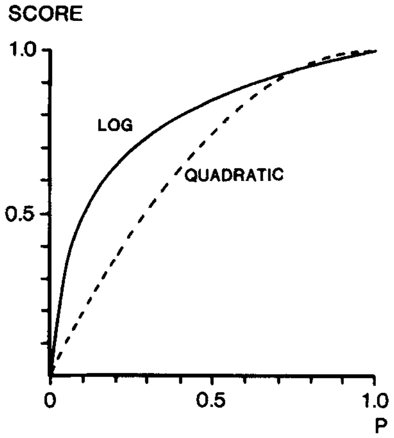

Figure 1. The logarithmic and quadratic scores as functions of the posterior probability P of belonging to the population of origin. The logarithmic score is truncated at $\varepsilon = 0{\cdot}01$ and rescaled to $[0; 1]$

test values and the prevalence of disease $P(D)$, the function $\hat{P}(D|x)$ is computed using Bayes' rule. Subsequently, a new set of (test) observations are sampled, and Q_i is computed for each individual. The average score of the set of test observations is an estimate of the value of the test. If separate sampling of reference and diseased populations is performed, the proportion of subjects with disease in the set of test observations does not reflect disease prevalence in the population, and therefore standardization of the score to the prevalence $P(D)$ may be carried out:[5]

$$\hat{Q}' = [1 - P(D)]\ 1/N_R \sum_{i=1}^{N_R} Q_i + [P(D)]\ 1/N_D \sum_{i=1}^{N_D} Q_i$$

where N_R and N_D denote the sample sizes drawn from the reference and diseased populations respectively.

EXAMPLES OF SCORES FOR DIAGNOSTIC TESTS

In Tables I and II some examples of the relation between the traditional (non-)error rate measures and the standardized scores ($P(D) = 0{\cdot}5$) are presented. Scores for binary tests with specificity and sensitivity equal to 0·5, 0·75 or 0·95 are given in row 1 of Table I. Using Bayes' formula, the posterior probabilities of belonging to the reference (R) or diseased (D) populations given a positive (+) or negative (−) test result, $P(R|-)$, $P(R|+)$, $P(D|+)$ and $P(D|-)$, have been computed from the specificity (Sp), sensitivity (Se), and prevalence $P(D) = 0{\cdot}5$.[7] The standardized quadratic score is given by

$$\begin{aligned} Q' = {} & [1 - P(D)]\,[Sp\{1 - [1 - P(R|-)]^2\} + (1 - Sp)\{1 - [1 - P(R|+)]^2\}] \\ & + P(D)\,[Se\{1 - [1 - P(D|+)]^2\} + (1 - Se)\{1 - [1 - P(D|-)]^2\}]. \end{aligned}$$

A similar expression applies for the truncated logarithmic score. The scores for the quantitative tests were obtained by simulation, using the NAG subroutines GO5DDF or GO5DFF with 5×10^5 pseudo-random numbers generated from each of the specified distributions.[9] The probability assigned to the population of origin for each observation was computed from the probability densities of the distributions, using the normal or log-normal density functions with the true population parameters. The probability of belonging to the diseased population for an

Table I. Relations between specificity, sensitivity and the standardized ($P(D)=0{\cdot}5$) quadratic and logarithmic scores for binary and quantitative tests. The logarithmic score is truncated at $\varepsilon=0{\cdot}01$ and rescaled to [0; 1]. The quantitative tests have normal distributions of test values with equal dispersions ($\sigma_D/\sigma_R=1$) in reference (R) and diseased (D) populations

	Specificity = sensitivity					
	0·5		0·75		0·95	
	Quadratic	Log	Quadratic	Log	Quadratic	Log
Binary test	0·750	0·850	0·813	0·878	0·953	0·957
Quantitative test	0·750	0·850	0·832	0·890	0·963	0·972

Table II. Relations between specificity, sensitivity and the standardized ($P(D)=0{\cdot}5$) quadratic and logarithmic scores for binary and quantitative tests. Examples with specificity of 0·95 and sensitivities of 0·5 or 0·75. The quantitative tests have normal distributions of test values in the reference (R) population and normal or log-normal distributions in the diseased (D) population

	Specificity = 0·95			
	Sensitivity			
	0·50		0·75	
	Quadratic	Log	Quadratic	Log
Binary test	0·814	0·897	0·878	0·919
Quantitative tests:				
Normal/normal:				
$\sigma_D/\sigma_R=1$	0·858	0·905	0·911	0·938
$\sigma_D/\sigma_R=2$	0·827	0·890	0·891	0·926
Normal/log-normal:				
$\sigma_R=1$, $\sigma_{\log}=0{\cdot}7719$	0·880	0·923	0·917	0·945

observation x from this population is given by

$$P(D|x)=\frac{P(D)\,pd_D(x)}{P(D)\,pd_D(x)+[1-P(D)]\,pd_R(x)},$$

where pd_D and pd_R denote the probability density distributions of the diseased and reference populations respectively. In all examples the reference distribution is standard normal, $N(0; 1^2)$, and the diseased population distributions are either normal or log-normal (Figure 2). In row 2 of Table I scores are presented for tests with normal distributions with equal dispersions ($\sigma_D/\sigma_R=1$). Mean values μ_D for the diseased population distribution have been selected so that specificity and sensitivity with a cutoff at $\mu_D/2$ are equal to those of the binary test examples. The scores are slightly different from those of the equivalent binary tests. For the non-discriminative test with $1-Sp=Se$ the quadratic score is 0·75, and the logarithmic score is 0·85.

In Table II the specificity of the binary tests is fixed at 0·95, and the sensitivity is 0·50 or 0·75. The location parameters of the quantitative tests were chosen so that, with a cutoff point at 1·645 (giving a specificity of 0·95), sensitivities were equal to those of the binary examples. Normal distributions with $\sigma_D/\sigma_R=1$ (row 2) or $\sigma_D/\sigma_R=2$ (row 3) were selected. A log-normal distribution with $\sigma_{\log}=0{\cdot}7719$ was chosen for the examples of row 4. For $\mu_{\log}=\ln(1{\cdot}645)$ this distribution has a

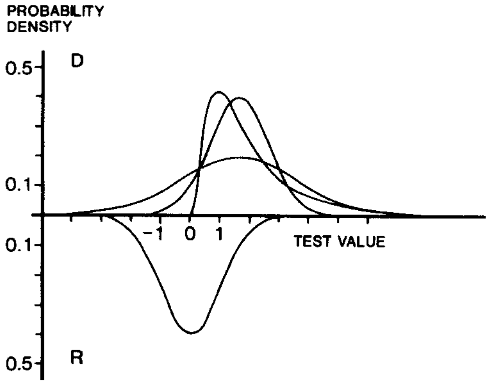

Figure 2. Probability density distributions in populations of reference subjects (R) and patients with disease (D) for hypothetical diagnostic tests. The reference distribution is standard normal ($N(0; 1^2)$), and the patients' distributions are normal ($N(1{\cdot}645; 1^2)$ and $N(1{\cdot}645; 2^2)$) or log-normal with $\mu_{\log} = \ln(1{\cdot}645)$ and $\sigma_{\log} = 0{\cdot}7719$

standard deviation of 2,[10] and skewness is 3·4. We notice considerable variation in the score for a given (non-)error rate. Actually, the quadratic score of the quantitative test with normal/log-normal distributions with sensitivity 0·5 is greater than the score of the binary test with sensitivity 0·75. Thus, a ranking of tests according to the respective measures may be quite different. In the following sections we concentrate on the quadratic score.

EVALUATION OF THE QUADRATIC SCORE OF A TEST USING THE BOOTSTRAP PROCEDURE

In the previous section scores were calculated on the basis of the true population distributions and probabilities. In practice we have to rely on sample estimates. If we estimate $\hat{P}(D|x)$ and evaluate the quadratic score using the same set of observations, a positive bias (ω) arises. This bias is analogous to that which occurs when the error rate of a prediction rule is evaluated on the training samples. The quadratic score can be evaluated on the training samples by use of a statistical method that reduces the bias, such as cross-validation or the bootstrap principle.[11–13] In general, the bootstrap method provides smaller root-mean-square errors of estimates than the cross-validation method, and so we chose the former technique. The procedure consists of repeated resampling of the original observations with replacement. Probability masses of $1/N_R$ and $1/N_D$ are assigned to respective test values, and separate random pseudo-samples are then drawn by a computer from the observed values in the reference and diseased population samples. The pseudo-samples are considered as training samples, and the function $\hat{P}^*(D|x)$ is estimated. With the original samples as test samples, we compute the quadratic score $\hat{Q}'^*_{\text{test}}$ using the function $\hat{P}^*(D|x)$. Evaluation of the quadratic score on the pseudo-training samples using the function $\hat{P}^*(D|x)$ yields a biased estimate $\hat{Q}'^*_{\text{train}}$. An estimate of the bias, $\hat{\omega}^*$, can be obtained from the difference:

$$\hat{\omega}^* = \hat{Q}'^*_{\text{train}} - \hat{Q}'^*_{\text{test}}.$$

By repetition of the procedure, say a hundred times, an average estimate of this bias is obtained ($\bar{\omega}^*$). A bias-corrected estimate $\hat{Q}'_{\text{corr}}$ of the original quadratic score $\hat{Q}'_{\text{train}}$ can then be derived:

$$\hat{Q}'_{\text{corr}} = \hat{Q}'_{\text{train}} - \bar{\omega}^*.$$

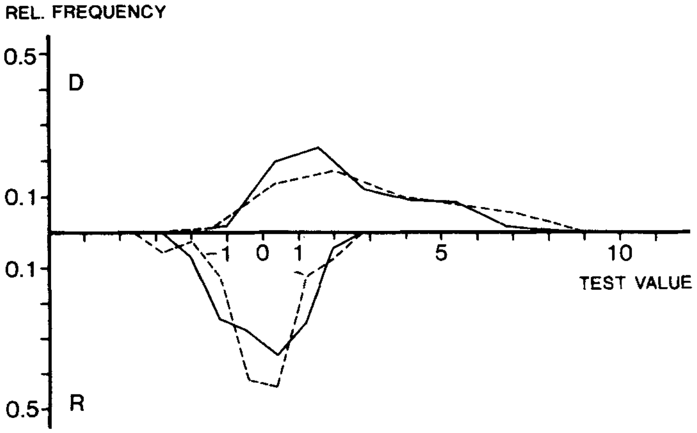

Figure 3. Frequency polygons of logarithmically transformed serum bilirubin (———) and serum bile acids (– – – – – – –), standardized to mean zero and unit standard deviation for the reference groups

At the same time, the standard deviation of $\hat{\omega}^*$ for the 100 bootstrap replications ($SD(\hat{\omega}^*)$) provides an estimate of the root-mean-square error of $\hat{Q}'_{\text{corr}}$ as an estimator of Q'_{test}, the population value of the quadratic score conditional on the estimated $\hat{P}(D|x)$ function.[11-13] Thus, an approximate 95 per cent confidence interval for Q'_{test} is $\hat{Q}'_{\text{corr}} \pm 2SD(\bar{\omega}^*)$.

Evaluation of the quadratic score by the bootstrap method is demonstrated for two clinical chemistry tests for detection of liver disease, serum concentrations of bilirubin (A) and of bile acids (B).[14, 15] A total of 143 hospitalized patients, in whom liver disease had been ruled out by clinical and paraclinical investigations, served as a reference sample, and 157 patients with verified liver disease constituted the sample of diseased subjects. The two groups were similar with respect to the distributions of age. In this example, a hospitalized reference group was preferred to a group of healthy, usually younger subjects such as blood donors. It is clear that any selection bias in the sampling of the training set of observations will be reflected in the point estimates and confidence intervals obtained by the bootstrap method. Figure 3 shows frequency polygons of logarithmically transformed test values for the two samples; the transformation has been applied to compress the scale. The sample distributions for the patients with liver disease are positively skewed, with coefficients of skewness $\hat{\gamma}_1$ equal to $+0{\cdot}7$ (A) and $+0{\cdot}4$ (B), respectively; the distributions of the reference group are slightly skewed in the opposite direction; with $\hat{\gamma}_1$ equal to $-0{\cdot}3$ for test A and $-0{\cdot}4$ for test B. A Kolmogorov–Smirnov test of normality for the reference distributions revealed that the transformed bile acid distribution deviated significantly from normal, $0{\cdot}025 < P < 0{\cdot}05$, and so a non-parametric approach for estimation of the function $\hat{P}(D|x)$ was selected (for both tests). Optimal class intervals for the frequency polygons of a skewed distribution (with standard deviation s in a sample of size N) have been determined from a formula given by Scott:[16]

$$h = k2{\cdot}15\, s\, N^{-0{\cdot}2}$$

where

$$k = 1/(1 - 0{\cdot}0060|\gamma_1| + 0{\cdot}27\gamma_1^2 - 0{\cdot}0069|\gamma_1|^3)$$

The test values were subjected to 200 bootstrap replications, and bias-corrected standardized ($P(D) = 0{\cdot}5$) quadratic scores ($\hat{Q}'_{\text{corr}}$) of 0·843 (A) and 0·867 (B) were recorded. These are roughly intermediate between a worthless and a perfect test. The standard errors conditional on the

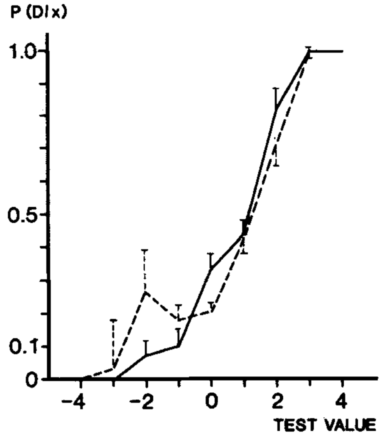

Figure 4. The function $\hat{P}(D|x)$ for logarithmically transformed serum bilirubin (———) and serum bile acids (– – – – – – –), x standardized to mean zero and unit standard deviation for the reference groups. Standard errors of the $\hat{P}(D|x)$ functions were obtained from 200 bootstrap replications

estimated functions $\hat{P}(D|x)$ were 0·0090 and 0·0099, respectively. The estimated biases were negligible, 0·0025 and 0·0028. Finally, the bootstrap technique also provides standard errors of $\hat{P}(D|x)$ evaluated at selected points. In each bootstrap replication, the value of $\hat{P}^*(D|x)$ was recorded at fixed values of x dispersed over the relevant interval. The standard deviation of the 200 values of $\hat{P}^*(D|x)$ is an estimate of the standard error of $\hat{P}(D|x)$ (Figure 4). Standard errors and so confidence intervals for the posterior probabilities are valuable, because knowledge of how close the *estimated* probabilities are likely to be to the true values is important for the proper use of a test as a decision support.

COMPARISON OF TWO TESTS BY THE BOOTSTRAP METHOD

Suppose that we wish to test whether the score for test A (0·843) is significantly different from that of test B (0·867). We distinguish between two purposes for carrying out such a test of significance. First, we may intend to apply the diagnostic test with the highest score as a decision support. Here we are interested in testing the difference between the scores conditional on the particular estimated functions $\hat{P}(D|x)$ which are to be used in the future. Therefore, we should apply the standard error of the difference conditional on the estimated functions $\hat{P}(D|x)$. The other purpose is to infer whether the score of one of the tests is higher than that of the other averaged over all estimated functions $\hat{P}(D|x)$; for example, is the score of serum bile acids higher than that of serum bilirubin? Here we should use the unconditional standard error of the difference.

In the first situation, we proceed as in the previous section. Pairs of random pseudo-values, one for each test, are resampled by putting probability masses of $1/N_R$ or $1/N_D$ on each pair of the original values. For each bootstrap replication we compute:

$$\Delta\hat{\omega}^* = \hat{\omega}_B^* - \hat{\omega}_A^*,$$

and then $SD(\Delta\hat{\omega}^*)$ serves as an estimate of the root-mean-square error of the difference between the bias-corrected quadratic scores conditional on the estimated functions $\hat{P}(D|x)$. The conditional standard error of the difference between test A and test B is 0·0107, and an approximate 95 per cent confidence interval for the true difference is 0·004 to 0·045. Thus, superior performance as a decision support is expected for the bile acid assay using estimated function $\hat{P}(D|x)$.

The unconditional variance of $\hat{Q}'_{\text{test}}$ is

$$\text{var}(\hat{Q}'_{\text{test}}) = \text{var}[E(\hat{Q}'_{\text{test}}|\hat{P})] + E[\text{var}(\hat{Q}'_{\text{test}}|\hat{P})].$$

An estimate of the unconditional standard error of the difference between the scores is obtained by slight modification of the resampling procedure. Instead of using the original set of values as test samples, we resample randomly a second set of N_R pairs of pseudo-values from the reference sample and N_D pairs of pseudo-values from the diseased sample, and regard these pseudo-samples as test samples. In the same way as before, corrected scores $\hat{Q}'_{\text{corr A(B)}} = \hat{Q}'_{\text{train A(B)}} - \bar{\omega}^*_{\text{A(B)}}$ are recorded. The standard deviation of $\hat{Q}'^*_{\text{test B}} - \hat{Q}'^*_{\text{test A}}$ for the bootstrap replications is used as an estimate of the unconditional standard error of the difference between the bias-corrected estimates $\hat{Q}'_{\text{corr B}} - \hat{Q}'_{\text{corr A}}$. The unconditional standard error is 0·0119, yielding a standard normal deviate z of 2·02. Thus, averaged over all estimated $\hat{P}(D|x)$ functions, the difference is just significant.

For comparison, the sensitivities of serum bile acids and bilirubin were 0·63 and 0·54, respectively, at a specificity of 0·95 (non-parametrically determined discrimination limit). The standard error of the difference was 0·056 and $z = 1{\cdot}61$, a non-significant result.[17, 18] For completeness, we mention that if transformations optimal for the reference distributions had been selected, a combined parametric/non-parametric approach for comparison of sensitivities could have been applied, resulting in a significant difference between the sensitivities.[18]

SOME POWER STUDIES

The power for comparison of two tests on the basis of the quadratic scores was compared with the power of the error rate measures for some typical situations (Table III). The power for non-parametric comparison of sensitivities of 0·50 and 0·75 with a specificity of 0·95 was computed for normal/normal models ($\sigma_D/\sigma_R = 1$ or 2) and for the normal/log-normal model considered previously.[18] The sample sizes were $N_R = N_D = 100$. The functions $\hat{P}(D|x)$ were estimated non-parametrically, and the unconditional standard error of the difference between the scores was estimated from 50 bootstrap replications in each trial. The power was obtained from 1000 trials. We observe that the power for the scoring rule approach exceeds that of the error rate approach for the normal/normal models. The powers were about equal in the normal/log-normal case. Thus the power relations for the two methods depend on the model and its parameters.

DISCUSSION

In a medical context, scoring rules based on posterior probabilities have been applied to evaluate particular clinical prediction rules.[4, 5, 19–21] In clinical chemistry several authors have suggested the conversion of test results to posterior probabilities, but they have not considered validation of the tests on the basis of these probabilities.[22–24] Application of the scoring rule concept to clinical chemistry tests has been briefly considered,[7] and here the focus has been on the statistical aspects. Tables I and II demonstrate that the (non-)error rate and the scores based on posterior probabilities measure different aspects of test performance. An aggregate measure of the posterior probabilities should reflect more closely the validity of a test as a decision support than the (non-)error rate which seems more appropriate for rating of screening tests.

The bootstrap method is useful for estimating the sampling variation of a score. As mentioned previously, a standard error conditional on a given $\hat{P}(D|x)$ function and an unconditional standard error can be obtained. Some simulation studies based on the normal/normal models for distribution of test values confirmed that the bootstrap estimates of standard errors were quite accurate. The unconditional standard error requires a Monte Carlo approach for computation,

Table III. Comparison of the powers of the error rate and scoring rule methods for some typical situations. Non-parametric approaches are used for both. Column 1 displays the model and parameters for tests A and B. Specificity, sensitivity and standardized ($P(D)=0·5$) quadratic scores (Q') are shown in columns 2 and 4, respectively. The powers appear in columns 3 and 5. The type I error rate is 0·05

1	2	3		4	5	
	Specificity = 0·95	Power*			Power*	
Model	Sensitivity	$\rho=0$	$\rho=0·75$	Q'	$\rho=0$	$\rho=0·75$
Normal (R)/normal (D):						
A: $N(0; 1^2)$-$N(1·645; 1^2)$	0·50	0·48	0·71	0·858	0·77	0·99
B: $N(0; 1^2)-N(1·645+0·675; 1^2)$	0·75			0·911		
A: $N(0; 1^2)$-$N(1·645; 2^2)$	0·50	0·72	0·97	0·827	0·90	0·99
B: $N(0; 1^2)$-$N(2·993; 2^2)$	0·75			0·891		
Normal (R)/log-normal (D):						
A: $N(0; 1^2)$-log-normal (ln 1·645; $0·7719^2$)	0·50	0·60	—	0·880	0·59	—
B: $N(0; 1^2)$-log-normal (1·01879; $0·7719^2$)	0·75			0·917		

$N_R=N_D=100$, Correlation (ρ) between tests A and B is assumed equal in the two populations

whereas the conditional standard error can also be computed in a direct way. The total sample is split at random into a training and a test set of observations. The variance of the standardized score conditional on the training function $\hat{P}(D|x)$ is then

$$[1-P(D)]^2\, 1/N_R^2 \sum_{i=1}^{N_R} (Q_i-Q_R)^2+[P(D)]^2\, 1/N_D^2 \sum_{i=1}^{N_D} (Q_i-Q_D)^2,$$

where N_R and N_D are the numbers of test observations, and Q_R and Q_D are the mean scores for the reference and disease groups. However, this method is less efficient than the bootstrap approach.

The bootstrap method also served for correction of bias, but for the univariate situation studied here the bias is rather negligible. In multivariate cases this aspect gains importance.[13] Finally, the bootstrap principle is convenient for estimation of standard errors for the posterior probabilities estimated non-parametrically. Parametrically, the standard errors $\hat{P}(D|x)$ can be calculated as shown by Machin *et al.*[25] In addition to random variation, the estimated posterior probabilities may be subject to systematic deviations from the true values. An incorrect parametric model or training samples that are non-representative of the background population cause biased posterior probabilities. In the latter case, independent training and test samples are necessary to disclose the error Splitting the quadratic score into sharpness and calibration components is also informative.[6]

ACKNOWLEDGEMENT

This work has been supported by the Danish Hospital Foundation for Medical Research, Region of Copenhagen, The Faroe Islands, and Greenland.

REFERENCES

1. Yerushalmy, J. 'Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques', *Public Health Reports*, **62**, 1432–1449 (1947).
2. Winkler, R. L. 'The quantification of judgment: some methodological suggestions', *Journal of the American Statistical Association*, **62**, 1105–1120 (1967).
3. Savage, L. J. 'Elicitation of personal probabilities and expectations', *Journal of the American Statistical Association*, **66**, 783–801 (1971).

4. Shapiro, A. R. 'The evaluation of clinical predictions', *New England Journal of Medicine*, **296**, 1509–1514 (1977).
5. Hilden, J., Habbema, J. D. F. and Bjerregaard, B. 'The measurement of performance in probabilistic diagnosis III. Methods based on continuous functions of the diagnostic probabilities', *Methods of Information in Medicine*, **17**, 238–246 (1978).
6. Spiegelhalter, D. J. 'Probabilistic prediction in patient management and clinical trials', *Statistics in Medicine*, **5**, 421–433 (1986).
7. Linnet, K. 'A review on the methodology for assessing diagnostic tests', *Clinical Chemistry*, **34**, 1379–1386 (1988).
8. Shuford, E. H., Albert, A. and Massengill, H. E. 'Admissible probability measurement procedures', *Psychometrika*, **31**, 125–145 (1966).
9. Numerical Algorithms Group (NAG) *Fortran Library Manual*, Numerical Algorithms Group Ltd, Oxford, 1984.
10 Aitchison, J. and Brown, J. A. C. *The Lognormal Distribution*, Cambridge University Press, London, 1969.
11. Efron, B. 'Bootstrap methods: another look at the jackknife', *Annals of Statistics*, **7**, 1–26 (1979).
12. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, 1982.
13. Efron, B. and Gong, G. 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *American Statistician*, **37**, 36–48 (1983).
14. Linnet, K., Kelbaek, H. and Frandsen, P. 'Predictive value of the concentration in serum of total 3alfa-hydroxy bile acids in the diagnosis of hepatobiliary disease', *Scandinavian Journal of Gastroenterology*, **17**, 263–268 (1982).
15. Linnet, K., Kelbaek, H. and Bahnsen, M. 'Diagnostic values of fasting and postprandial concentrations in serum of 3alfa-hydroxy bile acids and gamma-glutamyl transferase in hepatobiliary disease', *Scandinavian Journal of Gastroenterology*, **18**, 433–438 (1983).
16. Scott, D. W. 'Frequency polygons: theory and application', *Journal of the American Statistical Association*, **80**, 348–354 (1985).
17. Greenhouse, S. W. and Mantel, N. 'The evaluation of diagnostic tests', *Biometrics*, **6**, 399–412 (1950).
18. Linnet, K. 'Comparison of quantitative diagnostic tests: type I error, power, and sample size', *Statistics in Medicine*, **6**, 147–158 (1987).
19. Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F. and Gelpke, G. J. 'Comparison of discrimination techniques applied to a complex data set of head injured patients', *Journal of the Royal Statistical Society Series A*, **144**, 145–175 (1981).
20. Schmitz, P. I. M., Habbema, J. D. F. and Hermans, J. 'The performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods', *Statistics in Medicine*, **2**, 199–205 (1983).
21. Lindberg, G., Thomsen, C., Malchow-Møller, A., Matzen, P. and Hilden, J. 'Differential diagnosis of jaundice: applicability of the Copenhagen Pocket Chart proved in Stockholm patients', *Liver*, **7**, 43–49 (1987).
22. Van der Helm, H. J. and Hische, E. A. H. 'Application of Bayes's theorem to results of quantitative clinical chemical determinations', *Clinical Chemistry*, **25**, 985–988 (1979).
23. Albert, A. 'On the use and computation of likelihood ratios in clinical chemistry', *Clinical Chemistry*, **28**, 1113–1119 (1982).
24. Gruemer, H.-D., Miller, W. G., Chinchilli, V. M., Leshner, R. T., Hassler, C. R., Blasco, P. A., Nance, W. E. and Goldsmith, B. M. 'Are reference limits for serum creatine kinase valid in detection of the carrier state for Duchenne muscular dystrophy?', *Clinical Chemistry*, **30**, 724–730 (1984).
25. Machin, D., Dennis, N. R., Tippett, P. A. and Andrews, V. 'On the standard error of the probability of a particular diagnosis', *Statistics in Medicine*, **2**, 87–93 (1983).