A Review on the Methodology for Assessing Diagnostic Tests

Kristian Linnet

Evaluation of diagnostic tests by the following principles are reviewed: error rates, scores based on posterior probabilities, and the excess loss considered in a decision theoretic context. Error rates or the complementary non-error rates, specificity and sensitivity, are simple measures which provide a rough indication of the discriminative value. In clinical practice, where a test serves as a decision support together with other information, conversion of test results to posterior probabilities is recommended. An aggregate score of these probabilities expresses the value of the test. Finally, in simple, well-defined cases—for example, screening situations, where the prevalence of disease and the relative consequences of false-positive and -negative classifications can be estimated—a Bayesian decision analysis is appropriate. The optimal discrimination limit is selected, and the total loss is minimized. The likelihood ratio \( LR(x) \) plays a central role in probability calculations and in the decision analysis. An example illustrates application of the procedures.

Additional Keyphrases: clinical decision-making · Bayesian decision rules · receiver-operating characteristic curves · statistics · probability calculations · screening · loss measures · error rates · scoring rules based on posterior probabilities · multivariate analysis · cutoff values · discrimination limits

The following principles for test evaluation are considered here: error rates, scoring rules based on posterior probabilities, and loss measures developed in a decision theoretic framework (1-6).

The error rate measure represents the earliest developed and most widely used method. Error rates are estimated from frequency distributions of test results in groups of reference and diseased individuals. If in addition the prevalence of disease is known, posterior probabilities can be computed, and by applying a scoring rule an aggregate measure of the probabilities may summarize the value of the test. Finally, if the consequence of erroneous classifications of non-diseased individuals relative to that of diseased subjects can be estimated, a Bayesian decision analysis with optimization of test yield can be done. Now a loss measure expresses the value of the test. Thus, the type of evaluation method should be chosen with due regard to the context in which the test is applied and the amount of information that is available.

In this review I concentrate on the simple situation where individuals are allocated into one of two groups, depending on the result of a single laboratory test. The principles, however, do cover situations with more than two groups and more than one measurement per subject (multivariate analysis). I do not deal with evaluation of prognosis or grading of disease severity, and neither am I concerned with monitoring disease over time.

Basic Concepts

A test evaluation begins with measurements on representative samples of individuals from reference (\( R \)) and diseased (\( D \)) populations. It is presumed that the true status of the individuals has been established by other means than the test being subject to evaluation, i.e., by applying a "gold standard." A diagnostic test can be qualitative, giving plus or minus results, or quantitative, yielding results distributed on a continuous scale. As an example, Figure 1 displays absolute frequencies of the logarithms of bile-acid concentrations in fasting groups of reference and diseased individuals (7). Logarithmic values have been used in order to compress the scale. Traditionally, a discrimination limit is introduced so that the results are divided into negative and positive answers, which may be summarized in a 2 × 2 table (Figure 2). Reference individuals having a negative result are called true negatives (\( TN \)), reference individuals with a positive answer are false positives (\( FP \)), diseased subjects with a negative result are false negatives (\( FN \)), and diseased subjects with a positive answer are true positives (\( TP \)).

In general, one should distinguish between a construction phase, where the diagnostic rule—for example, the location of the discrimination limit—is delineated, and an evaluation phase, where the diagnostic ability is assessed. The observations, on which the diagnostic rule is founded, constitute the construction sample. Preferably, the diagnostic performance should be evaluated on a new set of observations (validation sample) to prevent an optimistic bias.

Error Rates or Non-Error Rates

The simplest and well-known principle for test evaluation is the error rate or the complementary non-error rate

![Graph](image)

Fig. 1. Frequency histograms for log (bile acid concentrations in serum) in reference (\( R \)) and diseased (\( D \)) groups of subjects

\( D.L. \), discrimination limit (0.975-fractile of the \( R \)-distribution).

Received February 16, 1988; accepted April 13, 1988.

Department of Clinical Chemistry KK 4051, Righospitalet, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark.

CLINICAL CHEMISTRY, Vol. 34, No. 7, 1988 1379
These error rates or non-error rates are estimates of unknown population values. For a qualitative test and a quantitative test with an a priori known discrimination limit, the standard errors are the binomial ones, for example, \(\sqrt{Sp(1-Sp)/N_R}\) for the specificity. The 0.95-confidence interval ranges from minus to plus two standard errors around the point estimate.

\(\frac{1}{2}(Sp + Se)\) is an overall non-error rate that is independent of \(N_D/N_R + N_D\), i.e., the prevalence of disease in the evaluation sample. Notice, however, that

\[
\frac{n(TP) + n(TN)}{n(TP) + n(FP) + n(FN) + n(TN)}
\]

depends on the prevalence. This fraction is sometimes called the accuracy or efficiency.

When no a priori known discrimination limit for a quantitative test exists, one has to establish a limit on the basis of the data. Commonly, an extreme fractile of one of the distributions is selected, for example the 0.95- or the 0.975-fractile of the reference distribution. Nonparametrically, the 0.975-fractile is estimated as the 0.975 \((-1)(N_R + 1)th\) ordered observation. For a gaussian distribution a more precise estimate is the mean plus two standard deviations. After this construction phase, the value of the test is judged from the sensitivity. When the discrimination limit has been estimated, the standard error of the sensitivity is the square root of the sum of two variance components. One component arises from the uncertainty associated with the estimation of the discrimination limit, and the other is the usual binomial variance of a proportion (8, 9). Taking the bile-acid example, a Kolmogorov–Smirnov test for goodness of fit revealed that the reference distribution did not deviate significantly from normality (\(P > 0.20\)), and so the 0.975-fractile was estimated parametrically to 2.27, which corresponds to 9.7 \(\mu\)mol/L for the original bile-acid values. The sensitivity was estimated to 0.64, with a standard error of 0.06.

Finally, the discrimination limit may be selected, taking the distributions of both groups into account. This approach is considered in the section on decision theory.

**Posterior Probabilities and Scoring Rules**

Given a patient with test result \(x\), one may ask: what is the probability that the patient is diseased (or healthy)? Before we perform a test, the probability of disease \(P(D)\), the so-called a priori probability, is equal to the prevalence of disease in the target population. Having obtained the result \(x\) for the patient, we revise the a priori probability to a posterior probability, \(P(D|x)\), i.e., the probability conditional on test result \(x\) (10, chapter 1). According to Bayes’ rule (11):

\[
P(D|x) = \frac{P(D) \cdot P(x|D)}{P(D) \cdot P(x|D) + [1 - P(D)] \cdot P(x|R)}
\]

For the two-group case, \(P(R|x)\) is equal to 1 \(- P(D|x)\). When a test is qualitative, \(x\) takes the values minus or plus, and we are primarily interested in \(P(D|+)\) and \(P(R|-)\). Recognizing that \(P(D|+) = Se\) and \(P(R|-) = 1 - Sp\), we observe that \(P(D|+)\) can be computed from specificity, sensitivity, and prevalence of disease by using formula 1. Analogously, \(P(R|-) = 1 - P(D|-)\) can be calculated from 1, noting that \(P(-|D) = 1 - Se\) and \(P(R|-) = Sp\). We might also obtain the results in a direct way from the \(2 \times 2\) table (Figure 2):

\[
P(D|+) = \frac{P(TP)/[n(TP) + n(FP)]}{P(TP)/[n(TP) + n(FP)] + P(TN)/[n(TN) + n(FN)]}
\]

\[
P(R|-) = \frac{P(TN)/[n(TN) + n(FN)]}{P(TP)/[n(TP) + n(FP)] + P(TN)/[n(TN) + n(FN)]}
\]

The direct calculation method presumes that the prevalence of disease in the total validation sample \([N_D/(N_R + N_D)]\) equals the prevalence of the target population. This will be the case (within the limits of random fluctuations) if subjects suspected of having the disease in question have been included without selection. On the other hand, if 100 reference and 100 diseased subjects have been sampled separately, primarily to assess specificity and sensitivity, the \(N_D/(N_R + N_D)\) ratio of the total validation sample obviously does not reflect the prevalence of disease in the target population, and formula 1 should be applied.

Vecchio (12) introduced the term “predictive value,” \(PV_{pos} = P(D|+)\) and \(PV_{neg} = P(R|-)\). I prefer the terms “posterior probability” or “conditional probability,” which are valid for both qualitative and quantitative tests and are in accord with conventional statistical notation.

Dealing with quantitative tests, \(P(x|D)\) and \(P(x|R)\) become probability densities, \(f_D(x)\) and \(f_R(x)\), respectively, which are estimated as relative frequencies from relative frequency polygons (Figure 3 and the Appendix). We have:

\[
P(D|x) = \frac{P(D) \cdot f_D(x)}{P(D) \cdot f_D(x) + [1 - P(D)] \cdot f_R(x)}
\]

Introducing the likelihood ratio (10, 11), \(LR(x) = f_D(x)/f_R(x)\), and multiplying by \(1/f_R(x)\) in 2 we obtain:

\[
P(D|x) = \frac{P(D) \cdot LR(x)}{P(D) \cdot LR(x) + [1 - P(D)]}
\]
This operation has not simplified the formula, but if we replace probabilities with odds a reduction is possible. In general:

\[
\text{odds} = \frac{\text{probability}}{1 - \text{probability}},
\]

and 3 can be rearranged to

\[
\text{odds}(D|x) = \frac{P(D)}{1 - P(D)} \cdot LR(x)
\]

or

\[
\text{odds}(D|x) = \frac{\text{odds}(D) \cdot LR(x)}{\text{odds}(D) + 1}
\]

In words, the equation states that the posterior odds are equal to the prior odds multiplied by the likelihood ratio for the result x. Thus, this form is an easily memorizable expression of Bayes' rule for the two-group case. Finally, after having computed the posterior odds one may revert to a posterior probability by using the relation:

\[
P(D|x) = \frac{\text{odds}(D|x)}{[\text{odds}(D|x) + 1]}
\]

A simple distribution-free method to estimate the \( LR(x) \) function is based on relative frequency polygons for the reference and diseased groups (Figure 3). A frequency polygon is constructed by connecting the class centers of a histogram with straight lines. An optimal class division for a given sample size and dispersion of test values is computed as described in the Appendix. \( LR(x) \) is then obtained in each point as the ratio between the relative frequencies per unit (Figure 4).

Once one has constructed the \( LR(x) \) or \( P(D|x) \) function, the next step is to evaluate the test on the basis of the posterior probabilities. A new set of validation observations are sampled, and the estimated probability of correct population of origin is recorded for each individual, i.e., \( P(R|x) \) for the reference and \( P(D|x) \) for the diseased individuals of the validation samples. Naively, one might consider a simple average of the estimated probabilities of correct population of origin as a measure of the value of the test:

\[
Q_I = \frac{1}{N_R + N_D} \left[ \sum_{i=1}^{N_R} P(R|x_i) + \sum_{i=1}^{N_D} P(D|x_i) \right]
\]

\( N_R \) and \( N_D \) refer to the numbers of individuals in the validation samples. Such a linear score would range from zero for the worst possible test, which assigns a probability of zero to the population of correct origin for each individual, to 1 for the ideal test, which assigns a probability of 1 for the population of correct origin for each subject. The score so obtained is not appropriate, because for a given degree of overlap between the distributions of reference and diseased individuals one achieves a higher score when the probability for belonging to the greatest group is exaggerated rather than assigned the true value (4).

The following simple example concerning the application of a "no test" illustrates the point. Suppose that the prevalence of disease \( P(D) \) is known to be 0.6, and that no further information is available. If the investigator is presented for an evaluation sample of 1000 individuals, 400 reference and 600 diseased individuals, and correctly assigns a \( P(D) \) value of 0.6 [and so \( P(R) = 0.4 \)] to each subject, he achieves the following score:

\[
Q_I = \frac{1}{400 + 600} \left[ (400 \cdot 0.4) + (600 \cdot 0.6) \right] = 0.52
\]

If the investigator erroneously had assigned a \( P(D) \) value of 1.0 (and \( P(R) = 0 \)) to each individual, he would have attained the following score:

\[
Q_I = \frac{1}{400 + 600} \left[ (400 \cdot 0) + (600 \cdot 1) \right] = 0.60
\]

i.e., a larger score for an overestimated probability for the greatest group.

Instead, we should use a score whose maximum for a given extent of overlap is achieved when the true probabilities are substituted. Theoretically, it can be proved (4) that the following two scores have the desired property:

\[
Q_{eq} = \frac{1}{N_R + N_D} \left[ \sum_{i=1}^{N_R} 1 - (1 - P(R|x_i))^2 + \sum_{i=1}^{N_D} 1 - (1 - P(D|x_i))^2 \right]
\]

and

\[
Q_{ln} = \frac{1}{N_R + N_D} \left[ \sum_{i=1}^{N_R} \ln(P(R|x_i)) + \sum_{i=1}^{N_D} \ln(P(D|x_i)) \right]
\]

\( Q_{eq} \) has been estimated from the relative frequency polygons of Fig. 3.
In the "no test" example the quadratic score would have been 0.76 or 0.6 when a \( P(D|x) \) value of 0.6 or 1.0, respectively, was assigned to each individual. The logarithmic score would have taken the values \(-0.673 \) or \(-\infty \), respectively. Scoring rules whose maximum for a given degree of overlap is achieved when the true probabilities are substituted are called "strictly proper scoring rules" (SPSR) (4).

For the remainder of this article, the focus is on the quadratic scoring rule. The quadratic score ranges from zero for the worst possible test to 1 for the ideal test. Given a disease prevalence of 0.5, a nondiscriminative test, which just assigns the prevalence as a posterior probability to each individual ("no test"), has a score of 0.75, and discriminative tests have scores exceeding this value. Because the score depends on the prevalence \( P(D) \), one may operate with a score standardized to a given prevalence:

\[
Q_{ow} = \frac{1 - P(D)}{1/N_D} \sum_{i=1}^{N_D} 1 - [1 - P'(R_{ix})]^2
\]

\[+ P(D)(1/N_D) \sum_{i=1}^{N_D} 1 - [1 - P(D|x_i)]^2\]

where \( P \) indicates the assigned probability of correct origin for the given \( P(D) \) value. As an example, one may standardize the score to a prevalence of 0.5.

Examples of the relation between the traditional error rate measure and the quadratic score standardized to \( P(D) = 0.5 \) are shown in Table 1. In row 1 the quadratic scores are presented for qualitative tests with specificity 0.95 and sensitivities 0.5 or 0.75. The score is calculated as follows for the example with sensitivity of 0.5. An evaluation sample of 1000 subjects consists of 500 reference and 500 diseased individuals. Of the reference subjects, 475 have a negative result and 25 a positive one. \( P'(R_{i-}) = 0.655 \), and \( P'(R_{i+}) = 0.091 \) (formula 1). The contribution to the score from the reference individuals is

\[
(1 - 0.5) \cdot 1/500 \cdot [475 \cdot (1 - (1 - 0.655)^2) + 25 \cdot (1 - (1 - 0.091)^2)] = 0.423. \]

Likewise, the contribution from the 500 diseased individuals is calculated to be 0.391, so that the standardized score is 0.81. In rows 2 and 3 the scores are given for quantitative tests with gaussian distributions of test values for both groups. The distributions have dispersion ratios \( \sigma_R^2/\sigma_D^2 \) equal to 1 or 2, and the location difference has been selected in such a way that for a discrimination limit giving a specificity of 0.95, the sensitivities become 0.5 or 0.75. \( LR(x) \) and so \( P'(D|x) \) have been determined from \( f_R(x) \) and \( f_D(x) \) by using the

- The posterior probability \( P(D|x) \) as a function of the log ( bile acid concentration)

The prevalence \( P(D) \) is supposed to be 0.5
consequences of false-positive and false-negative classifications, one may select an optimal decision rule. In decision theory we operate with losses or the complementary utilities (5, 6, 14, 15). Returning to the 2 × 2 table (Figure 2), we associate a certain loss with each type of classification: L(TN), L(TP), L(FN), and L(FP). Loss can be measured in various units—for example, mortality rates, years of life, or money. When we diagnose disease as being present for a diseased individual (TP), we apply a beneficial treatment, making on the average L(TP) less than L(FN). The consequence of making a false-negative classification is an amount of excess loss:

$$\Delta L_{FPN} = L(FN) - L(TP)$$

When testing reference subjects, we may falsely declare disease as being present (FP) and perform a superfuous treatment, which may represent a loss L(FP) because of side effects. L(FP) is generally greater than L(TN), and the excess loss is:

$$\Delta L_{FP} = L(FP) - L(TN)$$

A perfect test makes no false classifications, and the total excess loss is zero. For an imperfect test the total excess loss per subject being tested is:

$$\Delta L = \frac{1}{N_R + N_D} [\Delta L_{FP} \cdot n(FP) + \Delta L_{FPN} \cdot n(FN)]$$

Starting with the simple example $$\Delta L_{FP} = \Delta L_{FPN}$$, we have from Bayesian decision theory (6, chapter 4, and 16) that the classification rule that minimizes $$\Delta L$$ for the given test is:

$$\text{odds}(D|x) > 1: \text{allocate to } D$$
$$\text{odds}(D|x) < 1: \text{allocate to } R$$

In the next example we presume that the excess loss per false-negative classification is nine times greater than the excess loss per false-positive classification, i.e., $$\Delta L_{FP} / \Delta L_{FPN} = 1/9$$. Intuitively, we would try to prevent false-negative allocations to a greater extent than false-positive ones, and this goal is achieved by using an odds limit that is lower than 1/9 as above. More precisely, it can be shown (6, 16) that $$\Delta L$$ is minimized by the following rule:

$$\text{odds}(D|x) > \Delta L_{FP} / \Delta L_{FPN}: \text{allocate to } D$$
$$\text{odds}(D|x) < \Delta L_{FP} / \Delta L_{FPN}: \text{allocate to } R$$

From formula 4 we derive the optimal discrimination limit ($$d.L_{opt}$$) from the following relation:

$$LR(d.L_{opt}) = \Delta L_{FP} / \Delta L_{FPN} \cdot \frac{1 - P(D)}{P(D)}$$

This rule agrees with intuitive expectations. The larger $$\Delta L_{FP}$$ is, the smaller is $$d.L_{opt}$$, and as $$P(D)$$ increases $$d.L_{opt}$$ declines (assuming that the test values of diseased subjects tend to be larger than those of the reference individuals, making LR(x) an increasing function). Once having estimated $$d.L_{opt}$$ from the construction set of observations, the test is evaluated on an independent set of validation observations, and the total excess loss per subject $$\Delta L$$ expresses the value of the test. $$\Delta L$$ depends on the prevalence, and a standardization to a given prevalence may be performed:

$$\Delta L_{st} = (1 - P(D)) \cdot \Delta L_{FP} / (1/N_D) \cdot n(FP) + P(D) \cdot \Delta L_{FPN} / (1/N_D) \cdot n(FN)$$

The discrimination limit should be the optimal one for the given prevalence.

Using the bile-acid values, I now illustrate the problem of test optimization and the subsequent evaluation of performance. The conditions are hypothetical and not necessarily very realistic.

Subjects with a malignant disease receive a cytotoxic drug, which induces liver damage in a proportion of subjects equal to the prevalence in the construction set [95/(95 + 68) = 0.58]. The concentration of bile acids in serum serves as a test of liver function. The drug is withdrawn when the test is positive, and we suppose that the liver damage is reversible at the time of testing. False-negative subjects develop progressive liver dysfunction, resulting in a mortality rate of 0.8 (= L(FN)). True positives have a mortality rate of 0.5 (= L(TP)) because of the (untreated) malignant disease. The excess loss of lives per false-negative classification is $$\Delta L_{FPN} = 0.8 - 0.5 = 0.3$$ Reference subjects with a negative result receive a full treatment and have a mortality rate of 0.05 (= L(TN)). False positives stop receiving the drug, which results in a mortality rate of 0.5 (= L(FP)). $$\Delta L_{FP} = 0.5 - 0.05 = 0.45$$. We have:

$$LR(d.L_{opt}) = (0.45/0.3) \cdot \frac{1 - 0.58}{0.58} = 1.07$$

Figure 4 shows that $$d.L_{opt} = 1.79$$, which corresponds to 6.0 µmol/L on the original scale.

By using this estimated $$d.L_{opt}$$ we validate test performance on an independent sample of individuals (13). Of 75 reference subjects, four were false positive; and of 62 diseased individuals, 18 were recorded as false negatives. The estimated total excess loss of lives per subject standardized to $$P(D) = 0.58$$ is:

$$\Delta L_{st} = (1 - 0.58) \cdot 0.45 \cdot (4/75) + 0.58 \cdot 0.3 \cdot (18/62) = 0.06$$

The standard error is 0.01 (Appendix).

Evaluation of Multivariate Tests

In general, multivariate tests can be assessed by the same methods as univariate tests. The best-known multivariate techniques are linear discriminant analysis and logistic discriminant analysis (17, 18). In linear discriminant analysis, a score y is calculated for each subject from p separate measurements:

$$y = a_0 + a_1 x_1 + \ldots + a_p x_p$$

From a construction set of observations the coefficients $$a_0$$, $$a_1$$, ..., $$a_p$$, are estimated on the basis of certain mathematical principles (17). Subsequently, a score value y may be calculated for each subject of an independent set of validation observations and related to a discrimination limit just as in the univariate case. The value of the multivariate test can be expressed by the error rates. Further, if $$P(D)$$ is known, posterior probabilities can be calculated from the score values y, and the test can be assessed by a scoring rule. The various statistical computer programs available for performing discriminant analysis (BMDP, SPSS, SAS, and others) usually have options for expression of posterior probabilities. Finally, if the excess losses of false-positive
and -negative classifications are known, one may estimate an optimal odds limit from formula 5 and minimize the total excess loss.

Logistic discriminant analysis expresses the posterior probability as:

\[
P(D|x_1, x_2, ..., x_p) = \frac{\exp(b_0 + b_1x_1 + ... + b_px_p)}{1 + \exp(b_0 + b_1x_1 + ... + b_px_p)}
\]

where \(b_0\) takes the \(P(D)\) value into account. Whereas linear discriminant analysis performs optimally for multivariate normal distributions with equal dispersions, the logistic principle is optimal for a wider range of distributions because the coefficients are estimated in a different way (17). A test based on logistic discriminant analysis is assessed in the same way as a linear discriminant analysis test.

Other Test Measures

Consideration of receiver-operating characteristic (ROC) curves has become popular in recent years (2, 10 chapter 4), (19, 20). The ROC curve displays the sensitivity as a function of the false-positive rate (= 1 — specificity). Figure 6 shows the ROC curve for the bile-acid construction set of observations. By inspection of the curve one may select a combination of specificity and sensitivity that is suitable for the particular diagnostic problem in question. Having chosen a point on the curve, one uses the corresponding discrimination limit in the future. In a medical context, the ROC curve was introduced in radiology (10) to evaluate the diagnostic consequences of selecting various cutoff points. Because an exact continuous scale for cutoff points seldom is available for an image-diagnostic test, i.e., the cutoff may refer to various patterns of findings in a picture, a ROC curve is a convenient way of expressing the error rates for various cutoff measures. In clinical chemistry, however, test values are usually expressed on a continuous interval scale, or at least on an ordinal scale, and plotting both the sensitivity and the specificity as a function of the cutoff value displayed on the abscissa seems to be more instructive (Figure 7).

The area under the ROC curve has been advocated as a measure of the diagnostic value of a test (21). A test with a ROC curve coinciding with the diagonal is worthless, and the more the curve deviates upwards towards the left corner, the better the test is. Thus, the area under the ROC curve is related to the discriminative ability of the test. One may question whether such an area is an appropriate measure of the value of a test. Actually, the area expresses the average sensitivity obtained for a distribution of discrimination limits giving a uniform distribution of specificities. In practice, however, one selects a particular cutoff point for future use, and the value of the test should then be expressed by the specificity and sensitivity for this particular cutoff point and not for some hypothetical situation where the point is continuously changed. If a specified cutoff point is not desired, one should operate with posterior probabilities and summarize the value of the test by a scoring rule.

The ROC curve has also been used for test optimization in a decision theoretic framework. The slope of the curve is equal to the likelihood ratio \(\text{LR}(x) = f_D(x)/f_D(x)\), where \(x\) is the discrimination limit for the given point on the curve (2, 6). According to formula 6 the optimal discrimination limit is the value corresponding to the point on the curve where the slope is:

\[
(\Delta L_{FP}/\Delta L_{FN})(1 - P(D)/P(D))
\]

In my opinion, a more straightforward approach for identifying the optimal discrimination limit is the previously outlined procedure where the plot of \(\text{LR}(x)\) as a function of \(x\) (Figure 4) is used.

Finally, Youden's index and Shannon's concept of entropy are briefly considered.

Youden's index (22) is an overall non-error rate measure:

\[
J = \text{specificity} + \text{sensitivity} - 1
\]

This index ranges from zero for a worthless test to 1 for the perfect test. The index is a linear transformation of \(1/2 (Sp + Se)\) and so is equivalent to this average non-error rate.

Shannon's entropy concept was developed in the field of communication theory (23), and Böttner (24) has introduced the principle in clinical chemistry. Essentially, the entropy measure for a diagnostic test corresponds to the logarithmic scoring rule.

Discussion

In this presentation, the principle for test evaluation has been related to the situation in which the test is applied and
to the available information on disease prevalence and consequences of erroneous classifications.

In the everyday clinical situation, a clinical chemical test result enters into the decision process along with other information. If the doctor has some idea of the value of the prior odds for disease, he may judge the posterior odds or probability on the basis of the likelihood ratio for the test result. If the posterior probability for one of the groups is high, the doctor makes a decision—disease present or absent—and institutes ± therapy, accordingly. For a posterior probability of intermediate level he will probably gather more information before he makes a decision. When a test is used as a decision support as outlined here, a natural way to assess the validity of the test is to use an aggregate measure of the posterior probabilities, i.e., to apply a scoring rule. Several clinical chemists have advocated the conversion of test results to posterior probabilities (11, 25, 26). They have not, however, taken the consequence of expressing the value of the test from these posterior probabilities. In recent years, some examples of test evaluation by scoring rules have appeared (3,4,27,28), and the principle will probably be applied increasingly in the future.

In completely well-defined diagnostic situations, where only a clinical chemical test result enters the decision process and where \( P(D) \) and the consequences of wrong classifications are known, a Bayesian decision analysis for optimization of test yield seems appropriate. An example of such a case is screening of newborns for a metabolic disease such as hypothyroisys. \( P(D) \) is here known to be about \( 2 \cdot 10^{-4} \), and the excess losses can be measured in money. \( L(TN) \) is zero (we ignore the costs of screening), and \( L(TP) \) is the cost of additional testing for establishing a definitive diagnosis, say $100. Thus, \( \Delta L_{TP} = 100 \cdot L(FN) \) is the cost of lifelong institutional care, say $10^5. \( L(TP) \) is the cost of additional testing to verify the diagnosis and the cost of lifelong substitution therapy, say $10^4. \( \Delta L_{TP} = 10^5 - 10^4 = 10^5 \). Using formula 6, we estimate that \( LR(d.i.e.p) \) is equal to 0.5. From a \( LR(x) \) function we would be able to estimate \( d.L_{opt} \). Using an independent validation sample, the validity of the test might be expressed as the excess loss \( \Delta L \) (in money) per subject being tested. Petersen et al. (31) have partly performed such an analysis.

Up to now the cost of the screening procedure in itself has not been taken into account, because this factor does not enter the Bayesian analysis for test optimization. The cost of testing should be recognized when we examine whether testing is worthwhile for society or when various tests for the same diagnostic task are compared (14, 32). The total loss without testing should be compared with the total losses for different test procedures (32). Beck (33) has in part performed such an analysis for various combinations of tests of thyroid function.

Expressing the value of a diagnostic test by the error rates constitutes the simple method that can be applied without knowledge of \( P(D) \) or the excess losses. This approach may represent the first step in the evaluation process of a newly developed test, providing guidance on whether the discriminative value is low or high. In the case of a quantitative test, selection of an extreme fractile for one of the groups as the discrimination limit will often be suitable. A high specificity is attained when a high fractile of the reference distribution is chosen, and a positive result verifies with a high probability that disease is present, whereas a negative result leaves the question "plus or minus disease" unanswered. On the contrary, selection of a low fractile of the diseased group as a cutoff point results in a high sensitivity, and now a negative result excludes disease with great certainty, whereas a positive result is inconclusive (these considerations cover situations where \( P(D) \) is neither extremely low nor high).

Finally, some remarks concerning pitfalls in test evaluation. Representative samples of reference and diseased individuals are important for achieving realistic measures of test validity (34). A schematic representation of a tumor marker as a screening test for cancer. The diseased group should be in an early phase of the disease, one at which it is still curable. The reference subjects should be of about the same age as the diseased individuals and include subjects with benign conditions likely to influence the test results. For example, subjects with prostatic hyperplasia are appropriate reference individuals in case of validating a screening test for prostatic carcinoma.

When the same observations serve for construction and evaluation of a diagnostic rule, an optimistic bias arises (17, 35). Special statistical methods, such as cross validation or bootstrapping procedures, are available for bias correction so that an independent validation sample may be spared (17, 25–38). Bias correction is most important when dealing with a multivariate test, whereas the bias for an univariate test is rather small and may be ignored for reasonably large sample sizes (>100 observations per group). If two tests are compared, it is advisable to arrange a paired design, where all validation subjects are exposed to both tests, in order to prevent a bias because of differences in patient materials. Appropriate statistical tests for paired comparisons should be applied (9, 39).

Appendix

1. Frequency Polygons

A frequency polygon is a nonparametric estimate of the unknown underlying probability density curve for the population. One starts with calculation of mean, standard deviation, and coefficient of skewness \( \gamma_1 = 2 \left( \frac{\text{median} - \text{mean}}{\text{standard deviation}} \right)^3 \) for a symmetric distribution the optimal class width is:

\[
h = 2.15 \cdot \text{SD} \cdot N^{-0.2}
\]

"Optimal" is in the sense that the random error is minimized (39). For a skew distribution the optimal width is:

\[
h = k \cdot 2.15 \cdot \text{SD} \cdot N^{-0.2}
\]

where

\[
k = 1/[1 - 0.0060 \cdot |\gamma_1| + 0.27 \gamma_2^2 - 0.0069 \cdot |\gamma_3^3] \]

For each class \( j \), the number of observations \( n_j \) is recorded, and the relative frequency per unit is \( n_j/(N \cdot h) \). The frequency polygon is constructed by connecting the centers of each class by straight lines. Because the ordinate is the frequency per unit, it is allowable to operate with different (optimal) class divisions in reference and diseased groups when \( LR(x) \) is estimated.

2. Standard Errors of Standardized Scores and Excess Losses

Using the \( P'(D|x) \) function estimated from the set of construction values, we compute \( P'(R|x) \) and \( Q_i = 1 - [1 - \ldots ] \)
\( P(R | x)^2 \) for each reference individual of the validation sample. The average is:

\[
Q_R = (1/N_R) \sum_{i=1}^{N_R} Q_i
\]

Analogously, \( Q_D \) is calculated for the diseased validation group. The variance of the standardized score is:

\[
\text{Var}(Q_d) = \frac{1}{(N_D - 1)N_D} \sum_{i=1}^{N_D} (Q_i - Q_D)^2
\]

The standard error is the square root of the variance. The variance of the standardized loss is:

\[
\text{Var}(\Delta L_{FP}) = \frac{1}{(N_R - n(FP))N_R} \sum_{i=1}^{N_R} (N_R - n(FP))
\]

\[
+ \frac{1}{(N_D - n(FN))N_D} \sum_{i=1}^{N_D} (N_D - n(FN))
\]

This variance formula presupposes that the excess losses \( \Delta L_{FP} \) and \( \Delta L_{FN} \) are not subject to stochastic variation. When two tests have been evaluated on the same validation subjects, a paired statistical test should be applied, a paired-rank test for the quadratic score and McNemar's test for the difference of excess losses (38).

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

References