The interpretation of diagnostic tests

David E Shapiro

Center for Biostatistics in AIDS Research, Harvard School of Public Health, Boston, Massachusetts, USA

Laboratory diagnostic tests are central in the practice of modern medicine. Common uses include screening a specific population for evidence of disease and confirming or ruling out a tentative diagnosis in an individual patient. The interpretation of a diagnostic test result depends on both the ability of the test to distinguish diseased from nondiseased subjects and the particular characteristics of the patient and setting in which the test is being used. This article reviews statistical methodology for assessing laboratory diagnostic test accuracy and interpreting individual test results, with an emphasis on diagnostic tests that yield a continuous measurement. The article begins with a summary of basic concepts and terminology, then briefly discusses study design and reviews methods for assessing the accuracy of a single diagnostic test, comparing the accuracy of two or more diagnostic tests and interpreting individual test results.

1 Introduction

Laboratory diagnostic tests are central in the practice of modern medicine. Common uses include screening a specific population for evidence of disease and confirming or ruling out a tentative diagnosis in an individual patient. The interpretation of a diagnostic test result depends on both the ability of the test to distinguish diseased from nondiseased subjects and the particular characteristics of the patient and setting in which the test is being used. This article reviews statistical methodology for assessing laboratory diagnostic test accuracy and interpreting individual test results.

For convenience, a diagnostic test will be called continuous, dichotomous, or ordinal according to whether the test yields a continuous measurement (e.g. blood pressure), a dichotomous result (e.g. HIV-positive or HIV-negative), or an ordinal outcome (e.g. confidence rating for presence of disease – definitely, probably, possibly, probably not, definitely not). The main focus here is on continuous diagnostic tests, since the majority of laboratory diagnostic tests are of this type and since much new methodology for such tests has appeared recently in the statistical literature. Although dichotomous diagnostic tests also are common in laboratory medicine, they will only be discussed briefly here, because much of the methodology involves analysis of binomial proportions and 2 × 2 tables, which is likely to be more familiar to the reader. Many of the methods proposed for continuous tests borrow heavily from the vast literature on ordinal diagnostic tests, which are common in the radiological imaging field; methodology for ordinal tests has been reviewed in depth elsewhere, including a recent issue of Statistical Methods in Medical Research (volume 7, number 4, December 1998), and will only be discussed here when pertinent to continuous tests.

This article is organized as follows: Section 2 summarizes some basic concepts and terminology. Section 3 discusses study design and methods for assessing the accuracy
of a single diagnostic test. Section 4 reviews methods for comparing the accuracy of two or more diagnostic tests. Section 5 describes methods for interpreting individual test results. Finally, Section 6 contains some concluding remarks.

2 Basic concepts and terminology

The interpretation of a diagnostic test depends on two factors: (1) the intrinsic ability of the diagnostic test to distinguish between diseased patients and nondiseased subjects (discriminatory accuracy); and (2) the particular characteristics of the individual and setting in which the test is applied. Each of these factors will be discussed in turn.

2.1 Discriminatory accuracy

The discriminatory accuracy of a diagnostic test is commonly assessed by measuring how well it correctly identifies $n_1$ subjects known to be diseased ($D^+$) and $n_0$ subjects known to be nondiseased ($D^-$). If the diagnostic test gives a dichotomous result for each subject, e.g. positive ($T^+$) or negative ($T^-$), the data can be summarized in a $2 \times 2$ table of test result versus true disease status (Figure 1). In the medical literature, discriminatory accuracy is commonly measured by the conditional probabilities of correctly classifying diseased patients [$P(T^+ | D^+) = TP/n_1$ in Figure 1, called the sensitivity or true positive rate (TPR)] and nondiseased subjects [$P(T^- | D^-) = TN/n_0$, called the specificity or true negative rate (TNR)]; equivalently, the

![Figure 1: The $2 \times 2$ table of frequencies summarizing the results of a dichotomous diagnostic accuracy study. TP represents the number of true positives (diseased patients who test positive), FP the false positives, FN the true negatives, and FN the false negatives.](image-url)
probabilities of type I error $[P(T+ \mid D-) = FP/n_0 = 1 - \text{specificity, called the false positive rate (FPR)}]$ and type II error $[P(T- \mid D+) = FN/n_1 = 1 - \text{sensitivity, called the false negative rate (FNR)}]$ may be also be used.

Many other indices of accuracy have been proposed for dichotomous diagnostic tests, including several measures of association in the $2 \times 2$ table familiar to biostatisticians, such as the odds ratio and kappa statistic. One commonly-used accuracy index, the overall percentage of correct test results $P_c = (TP+TN)/N$, deserves comment here because it can be misleading. Unlike sensitivity and specificity, $P_c$ depends on the prevalence of disease in the study sample ($n_1/N$); in fact, $P_c$ is actually a prevalence-weighted average of sensitivity and specificity. Consequently, when the prevalence is high or low, $P_c$ can be high even when the specificity or sensitivity, respectively, is quite low. For example, in screening pregnant women for HIV infection in the USA, where the prevalence in many areas is <1%, an HIV test that classified all women as HIV-negative, and therefore identified none of the HIV-positive women, would have $P_c > 99\%$ even though its sensitivity would be 0%.

Many laboratory diagnostic tests give a quantitative result $X$ rather than a dichotomous result. The degree of overlap in the distributions of test results for diseased $[f(X \mid D+)]$ and nondiseased subjects $[f(X \mid D-)]$ determines the ability of the test to distinguish diseased from nondiseased. In screening situations, where the test is used to separate patients into two groups – those requiring further investigation and/or intervention and those who do not – it is common to define a threshold or decision limit $\gamma$ and classify the patient as diseased if $X > \gamma$ and nondiseased if $X < \gamma$ (Figure 2). Each decision limit $\gamma$ yields a different $2 \times 2$ table of dichotomized test result versus true disease status; thus, sensitivity and specificity can be estimated for each decision limit $\gamma$. However, as $\gamma$ decreases, sensitivity increases but specificity

![Figure 2](https://example.com/figure2.png)

**Figure 2** Hypothetical distributions of diagnostic test results $X$ for diseased and nondiseased subjects. The vertical line at $X = \gamma$ indicates the decision limit for a positive test. The shaded area to the right of $\gamma$ is the FPR; the shaded area to the left of $\gamma$ is the FNR.
decreases; as $\gamma$ decreases, specificity increases at the expense of sensitivity. Thus, there is a trade-off between sensitivity and specificity as the decision limit varies, which must be accounted for in assessing discriminatory accuracy.

A useful graphical summary of discriminatory accuracy is a plot of sensitivity versus either specificity or FPR as the decision limit varies, which is called the receiver operating characteristic (ROC) curve (Figure 3). The upper left corner of the graph represents perfect discrimination ($TPR = 1$, $FPR = 0$), while the diagonal line where $TPR = FPR$ represents discrimination no better than chance. The ROC curve of a diagnostic test is invariant with respect to any monotone transformation of the test measurement scale for both diseased and nondiseased subjects; in other words, the ROC curve does not depend on the scale of measurement for the test, which makes it useful for comparing diagnostic tests of different scales, including tests for which the decision limit scale is latent rather than quantitative, such as medical imaging techniques. Since it is often useful to have a single number that summarizes a test’s accuracy, various summary indices based on the ROC curve have been proposed; the most popular such summaries are the area under the ROC curve (AUC), which represents both the average sensitivity over all values of FPR and the probability that test results for a randomly-selected diseased subject and nondiseased subject will be ordered correctly in magnitude; the partial AUC, which is the AUC over a range of FPR that is relevant to a particular setting; and the sensitivity at a fixed FPR.

2.2 Context-specific interpretation

The assessment of discriminatory accuracy focuses on the performance of the diagnostic test when disease status is known; however, when interpreting the diagnostic test result for an individual patient, typically the true disease status is unknown and it is of interest to use the test result to estimate the posterior probability

![Figure 3](image-url)  
**Figure 3** The ROC curve corresponding to the test result distributions shown in Figure 2. The point labelled on the curve corresponds to the decision limit shown in Figure 2.
of disease. For interpretation of a dichotomous test result, the quantities of interest are the posterior probability of disease given a positive test result \[ P(D+ | T+) = TP/m \]

1

in Table 1], called the positive predictive value (PPV), and the posterior probability of no disease given a negative test result \[ P(D− | T−) = TN/m_0 \], called the negative predictive value (NPV). Using Bayes theorem, one can calculate the PPV and NPV based on the sensitivity and specificity of the diagnostic test and the prior probability of disease (prevalence); if probabilities are converted to odds [odds = probability/(1 − probability)], this calculation has a particularly simple form: posterior odds of disease = prior odds of disease \times likelihood ratio, where the likelihood ratio is \( f(T+ | D+)/f(T+ | D−) \) = sensitivity/(1 − specificity).

For tests that yield a quantitative result \( X \), one could determine the PPV and NPV as in the dichotomous case by dichotomizing the test result at a fixed decision limit \( \gamma \). However, this method has two limitations: it discards information about the likelihood of disease, since it ignores the distance of \( X \) from \( \gamma \), i.e. it yields the same estimate of PPV or NPV whether \( X \) just exceeds \( \gamma \) or is well beyond \( \gamma \); also, this method requires selecting a decision limit \( \gamma \) that is optimal in some sense for the particular setting. An alternative approach that avoids these difficulties is to calculate the likelihood ratio \( L(X) = f(X | D+)/f(X | D−) \) corresponding to the actual test result, and use it to convert the patient’s prior probability of disease into a posterior probability of disease \( P(D+ | X) \).

It is important to remember that PPV, NPV, and posterior probability of disease are specific to a particular patient or setting, and cannot necessarily be generalized to a different population or setting, because these measures depend strongly on the prevalence of disease. For example, an HIV screening test that is 82% sensitive and > 98% specific would have a PPV of 88% if the prevalence were 10%, as in a high risk setting, but a PPV of < 3% if the prevalence were only 0.044%, as might be the case in screening the general population.

3 Assessing the accuracy of a single diagnostic test

This section begins with a brief overview of design considerations for studies of discriminatory accuracy. It then reviews methods for continuous diagnostic tests, including estimation of the ROC curve, summary indices, and incorporating covariates, and concludes with a brief discussion of methods for dichotomous diagnostic tests. Context-specific assessment of diagnostic test performance will be discussed in Section 5.

3.1 Study design

Study design considerations for diagnostic test accuracy studies have received much attention in the literature, but have been reviewed elsewhere, so only a brief summary of the major issues will be given here.

Boyd reviews common study designs for diagnostic test accuracy studies, including prospective and retrospective designs. Results of such studies are subject to several potential biases, some of these biases can be avoided by careful study design, and some can be corrected to some degree in the analysis. Begg identifies the most
prominent and important biases as those concerned with issues relating to the reference test, or ‘gold standard’, used to determine true disease status. Verification bias in the reported sensitivity and specificity estimates occurs if the selection of patients to receive the reference test is influenced by the result of the diagnostic test being studied. A prospective study design in which all subjects in a defined cohort receive the reference test is preferred, but selective verification may be unavoidable for some diseases, such as when the reference test is invasive; methods for correction of verification bias exist and have been reviewed recently by Zhou. Another important source of bias is an imperfect reference test; many bias-correction methods have been proposed. The interpretation of the reference test and diagnostic test results should be independent and blinded. Failed or uninterpretable test results should be included in the analysis, because such results may affect the test’s usefulness. Selection of appropriate samples of diseased and nondiseased subjects is critical, since sensitivity and specificity can vary according to the ‘case-mix’ or disease spectrum in the study cohort. Adequate sample size is also critical; methods for determining sample size for diagnostic accuracy studies have been reviewed recently by Obuchowski.

Many diagnostic accuracy studies in the medical literature have serious methodologic flaws. A recent review of diagnostic test studies published in four prominent general medical journals between 1978 and 1993 found that the percentage of studies that fulfilled criteria for each of seven methodological standards ranged from 8% (reported test indexes for clinical subgroups) to 46% (avoided verification bias). The proportions of studies meeting individual standards and meeting multiple standards increased over time during the review period, but in the most recent interval, 1990–93, only one methodological standard was fulfilled by more than 50% of the studies, and only 44% of the studies met at least three of the standards.

3.2 ROC curve estimation

Let \( X \) and \( Y \) denote the continuous diagnostic test result for nondiseased and diseased populations with cumulative distribution functions \( F \) and \( G \) respectively. Then, the sensitivity at decision limit \( t \) is the survivor function \( F(t) = 1 - F(t) \), and FPR is \( G(t) = 1 - G(t) \). The ROC curve of the diagnostic test is a plot of \((F(t), G(t))\) for all possible values of \( t \), or, equivalently, a plot of \((p, q)\), where \( p = \text{FPR} \) and \( q = \text{TPR} \).

Suppose diagnostic test results \( x_1 \ldots x_{n_0} \) are available for \( n_0 \) nondiseased and \( y_1 \ldots y_{n_1} \) for \( n_1 \) diseased subjects. Many approaches have been proposed for estimating the ROC curve, including parametric, semiparametric, and nonparametric methods; each category will be reviewed in turn.

**Parametric approaches**

One parametric approach is to assume that \( F \) and \( G \) follow a parametric family of distributions, e.g. normal or lognormal, and fit these distributions to the observed (possibly transformed) test results. A major advantage of this distributional approach is that it yields a smooth ROC curve that is completely specified by a small number of parameters, which form the basis for statistical inference; hypothesis tests,
variance estimates, and confidence intervals for the fully parametric binormal ROC curve and its parameters are well-developed. The parametric method, however, is quite sensitive to departures from the distributional assumptions.\(^{15}\)

An alternative parametric approach is to specify the functional form of the ROC curve, e.g. hyperbolic\(^{16}\) or weighted exponential.\(^ {17}\) This approach is closely related to the distributional approach, in that specifying the underlying test result distributions determines the form of the ROC curve; e.g. the symmetric ‘bilogistic’ ROC curve obtained by assuming that F and G follow logistic distributions with equal scale parameters is hyperbolic in form.\(^5\) Note that in fitting an ROC curve directly in ROC space, the fitting algorithm needs to take into account the errors in both FPR and TPR; e.g. simple least squares minimizing vertical deviations would be inappropriate.\(^2\)

Semi-parametric approaches

Metz et al.\(^ {18}\) propose a semi-parametric algorithm, LABROC4, that groups the ordered continuous test measurements into runs of diseased and nondiseased subjects and then fits a binormal ROC curve using the Dorfman-Alf maximum likelihood algorithm for ordinal data.\(^ {19}\) The LABROC4 approach assumes that the underlying distributions of the grouped nondiseased and diseased test measurements can be transformed to normal distributions by an unspecified monotone transformation of the test measurement axis. The resulting binormal ROC curve is of the form \(\hat{TPR} = \Phi(a + b\Phi^{-1}(FPR))\), where \(\Phi\) is the standard normal cumulative distribution function, \(a\) is the scaled difference in means and \(b\) the ratio of standard deviations of the latent normal distributions of diseased and nondiseased. Large-sample normal-theory standard errors and confidence intervals for the ROC curve parameters are provided.

The assumption of latent normal distributions for grouped data in the LABROC4 method is less strict than the assumption of explicit distributional form for the continuous measurements in the fully parametric approach. However, the LABROC4 method does still assume that a single monotone transformation exists that would make both latent distributions normal simultaneously; this assumption can be checked graphically by assessing linearity of the raw data on normal probability axes (\(\Phi^{-1}(FPR)\) versus \(\Phi^{-1}(TPR)\)). Hanley\(^ {20}\) demonstrated that by transformation one can make many highly non-normal pairs of distributions resemble the binormal model; however, Zou et al.\(^ {21}\) constructed a simulated dataset in which \(G\) was bimodal but \(F\) was unimodal (as might occur if the diagnostic test identified only one of two possible manifestations of a disease), and found that the LABROC4 ROC curve departed substantially from the ROC curves obtained by nonparametric methods (described next). Hsieh and Turnbull\(^ {22}\) proposed a generalized least squares procedure for fitting the binormal model to grouped continuous data and compared it to the Dorfman-Alf algorithm;\(^ {19}\) they also described a minimum-distance estimator of the binormal ROC curve that does not require grouping the data.

Nonparametric approaches

The simplest nonparametric approach involves estimating \(F\) and \(G\) by the empirical distribution functions \(\hat{F}_n(t) = \#\{x_i < t\}/n_0\) and \(\hat{G}_n(t) = \#\{y_i < t\}/n_1\) for the non-diseased and diseased subjects, respectively. This approach is free of distributional
assumptions in that it depends only on the ranks of the observations in the combined sample, but the resulting empirical ROC curve is a series of horizontal and vertical steps (in the absence of ties), which can be quite jagged. Hsieh and Turnbull derive asymptotic properties of this estimator. If the data contain ties between test measurements of nondiseased and diseased subjects, corresponding sections of the empirical ROC curve are undefined; Le proposes a modified estimator of the empirical ROC curve based on mid-ranks that overcomes this difficulty and shows that it converges uniformly to the true ROC curve.

An alternative nonparametric approach is to fit a smoothed ROC curve using kernel density estimation of $F$ and $G$:

$$
\hat{F}(t) = \frac{1}{n_0 h} \sum_{i=1}^{n_0} K\left(\frac{t - x_i}{h}\right)
$$

where $K = \int k$, the kernel $k$ is a mean 0 density function, and $h$ is the ‘bandwidth’ which controls the amount of smoothing; $G(t)$ is defined similarly based on $y_1, \ldots, y_{n_1}$. Zou et al. suggested using a biweight kernel $k(t) = 15/16(1 - t^2)^2$ for $t$ in $(-1,1)$, with bandwidth $h_0 = 0.9 \min(\text{SD},\text{IQR}/1.34)n^{-1/5}$, where SD and IQR are the sample standard deviation and interquartile range, respectively; separate bandwidths $h_{n_0}$ and $h_{n_1}$ are determined for $F$ and $G$. This bandwidth is optimal for histograms that are roughly bell-shaped, so that a preliminary transformation of the test measurement scale might be required. Lloyd suggested using the standard normal kernel with bandwidths $h_{n_0}/h_{n_0} \approx (n_0/n_1)^{1/3}$; this choice of bandwidth minimizes mean square error (MSE) if the overall smoothness of the two densities is similar. Lloyd also suggests using the bootstrap to reduce bias in the smooth ROC curve. Advantages of the kernel method are that it closely follows the details of the original data and is free of parametric assumptions, but disadvantages are that it is unreliable at the ends of the ROC curve, requires some ad hoc experimentation with choice of bandwidth, and does not work well when substantial parts of $F$ or $G$ are near zero.

Simultaneous confidence bands for the entire ROC curve

For the binormal model, Ma and Hall give simultaneous confidence bands for the entire ROC curve by constructing Working–Hotelling confidence bands in probit coordinates, where the binormal ROC curve is linear, and transforming them back to ROC coordinates; the method generalizes to other location-scale parametric ROC models. Campbell constructs fixed-width confidence bands for the entire ROC curve based on the bootstrap; Li et al. provide an alternative bootstrap simultaneous confidence band for the entire ROC curve based on a vertical shift quantile comparison function.

3.3 Estimation of summary accuracy indices

The ROC curve provides a graphical summary of discriminatory accuracy, but often a one-number summary index of discriminatory accuracy is desired. Many such summary indices have been proposed; the choice of an appropriate index depends on the application. The AUC is the most popular global index for the ordinal diagnostic tests common in radiology, where the test measurement scale is usually latent rather
than an observed continuous variable; however, the AUC has several limitations that may make it less useful for continuous diagnostic tests: when two ROC curves cross, the two tests can have similar AUC even though one test has higher sensitivity for certain specificities while the other test has better sensitivity for other specificities; the AUC includes regions of ROC space that would not be practical interest (e.g., very high FPR, or very low TPR); and the AUC is actually the probability that the test results of a random pair of subjects, one diseased and one nondiseased, are ranked correctly (i.e., $P(X < Y)$), but in practice the test is not given to pairs of diseased and nondiseased subjects. Despite these limitations, since the AUC continues to be the focus of much methodologic work and it may be still be useful in certain applications, this section begins by reviewing methods for AUC estimation and inference then discusses alternative summary indices.

**AUC**

For normally distributed continuous test results, the AUC is

$$\Phi \left( \frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}} \right)$$

so the AUC can be estimated directly by substituting sample means and standard errors; Wieand et al. obtain a variance estimator via the delta method. The semiparametric LABROC algorithm estimates the AUC based on the binormal ROC curve parameters, with large-sample standard errors and confidence intervals.

An unbiased nonparametric estimate of $P(X < Y)$ is the trapezoidal area under the empirical ROC curve, which equals the Mann–Whitney form of the two-sample Wilcoxon rank-sum statistic. Several variance estimators for this nonparametric AUC estimator have been proposed; Hanley and Hajian-Tilaki recommend either using the variance based on the theory of U-statistics or jackknifing. Coffin and Sukhatme show that if the diagnostic test results are subject to measurement error, the nonparametric area estimator is biased downward, i.e., underestimates the true population area. They show that resampling methods such as the bootstrap and jackknife are not appropriate for correction of biases caused by nonsampling errors, and derive a bias-correction factor based on kernel-density estimation. Monte Carlo simulations indicated that for binormal and bigamma ROC models, with normal and non-normal measurement errors, the bias-corrected estimator had smaller bias and comparable MSE to the uncorrected estimator. However, they caution that although measurement errors also reduce the power of hypothesis tests based on the uncorrected nonparametric area estimate, it is unclear whether use of the bias-corrected estimator would result in increased power due to its larger variance.

Hajian-Tilaki et al. assessed the bias of parametric and nonparametric AUC estimates for continuous test results generated by Monte Carlo simulation from binormal and nonbinormal models. The biases in both the parametric and nonparametric AUC estimates were found to be very small with both binormal and nonbinormal data. However, for nonbinormal data, standard error estimates by both parametric and nonparametric estimates exceeded the true standard errors.
Tilaki et al. did not evaluate the effect of the small biases and inflated standard errors on the power and size of hypothesis tests; however, in another simulation study in which the binormal model was applied to simulated ordinal rating data from the bilogistic model, Walsh observed small biases in area estimates and inflated standard errors that together substantially altered the size and power of statistical tests comparing independent areas, suggesting that model misspecification can lead to altered test size and power loss.

Zou et al. obtain the area under their smooth kernel-density-based nonparametric ROC curve by numerical integration. Since the smoothing does not affect the true SE of AUC, to first order, they use the U-statistic variance for the area under the smooth ROC, and calculate a large-sample normal theory confidence interval after applying a log transformation to reduce skewness. They also provide a variance estimate and transformation-based confidence interval for the numerically-integrated partial area under the smooth nonparametric ROC curve. Lloyd shows that for the smooth ROC curve based on standard-normal kernel density estimation, there is no advantage in terms of variance to determining the area under the curve by numerical integration rather than via the Mann–Whitney estimate.

Whether the AUC is estimated parametrically from the binormal model, or nonparametrically using the Mann–Whitney statistic, confidence intervals are usually obtained by relying on asymptotic normality. Using Monte Carlo simulation, Obuchowski and Lieber evaluated the coverage of the asymptotic confidence intervals and alternative ones based on the bootstrap and Student’s t-distribution for a single AUC and for the difference between two AUC. They generated both continuous and ordinal data with sample sizes between 10 and 70 for both diseased and nondiseased subjects and AUC of moderate (0.8) and high (0.95) accuracy. They found that for the difference in AUC, asymptotic methods provided adequate coverage even with very small sample sizes (20 per group). In contrast, for a single AUC, the asymptotic methods do not provide adequate coverage for small samples, and for highly accurate diagnostic tests, quite large sample sizes (> 200 patients) are required for the asymptotic methods to be adequate. Instead, they recommend using one of three bootstrap methods, depending on the estimation approach (parametric versus nonparametric) and AUC (moderate versus high).

**Partial AUC**

If only a particular range of specificity or sensitivities values is relevant, the partial AUC may be a more appropriate accuracy index than the AUC. McClish calculates the partial AUC or average sensitivity over a fixed range of FPR by direct numerical integration, with variances based on the binormal model; Thompson and Zucchini provide an alternative method based on integrating the bivariate normal probability density function. For settings in which high sensitivity is relevant, Jiang et al. extend McClish’s approach to obtain the average specificity over a range of high sensitivity.

**Sensitivity at fixed specificity**

Greenhouse and Mantel provide normal-theory and nonparametric tests of the hypothesis that a diagnostic test has at least a specified sensitivity (e.g. ≥90%) with FPR no higher than a specified value (e.g. ≤5%); however, their procedure does not
provide an explicit decision limit that has the desired sensitivity and specificity. Schafer\textsuperscript{49} discusses two methods for estimating an explicit decision limit with pre-specified specificity and/or sensitivity, one based on tolerance limits and the other an extension of the Greenhouse–Mantel method that applies when both sensitivity and specificity are pre-specified; he demonstrates that the extended Greenhouse–Mantel method can yield substantial gains in efficiency compared to the tolerance-limit method.

For the binormal model, McNeil and Hanley\textsuperscript{34} give pointwise confidence intervals for sensitivity at fixed FPR by transforming to probit coordinates \((\Phi^{-1}(\text{FPR}), \Phi^{-1}(\text{TPR}))\), constructing large sample confidence intervals, and transforming back to ROC coordinates; the simultaneous Working–Hotelling bands of Ma and Hall\textsuperscript{26} also have a pointwise interpretation (with a different confidence coefficient) as ‘vertical’ confidence bands for TPR at fixed FPR, or ‘horizontal’ bands for FPR at fixed TPR. Schafer\textsuperscript{45} gives a nonparametric large-sample pointwise confidence interval for sensitivity at the estimated decision limit needed to obtain a pre-specified specificity, by inverting the Greenhouse–Mantel\textsuperscript{43} test statistic. Zou \textit{et al.}\textsuperscript{21} suggested constructing a large-sample normal-theory confidence interval based on the kernel-estimated ROC curve in logit coordinates (logit(FPR), logit(TPR)), then transforming back to ROC coordinates; Lloyd\textsuperscript{25} suggested the analogous procedure based on a probit transformation.

A difficulty with confidence intervals for sensitivity at fixed FPR is that the decision limit \(t\) on the test measurement scale that yields this FPR can only be estimated. An alternative is to construct a joint confidence region for the point (FPR, TPR) on the ROC curve corresponding to a given decision limit \(t\). Hilgers\textsuperscript{46} constructs a point estimator for \(t\) and joint confidence rectangle with confidence level \((1 - \alpha)^2\) based on separate distribution-free tolerance intervals for \(F\) and \(G\). Unlike the asymptotic methods, Hilgers’ method is valid for any sample size; on the other hand, Schafer\textsuperscript{45} found that substantial gains in efficiency (i.e. narrower confidence intervals) can be obtained with the Greenhouse–Mantel method versus Hilgers’ method when sample sizes are adequate for the Greenhouse–Mantel method to be valid. Zou \textit{et al.}\textsuperscript{21} suggest constructing joint confidence rectangles with coverage \((1 - \alpha)^2\) by forming large-sample confidence intervals for the kernel-based logit(\(p\)) and logit(\(q\)) and transforming them back to ROC coordinates. Campbell\textsuperscript{23} uses Kolmogorov–Smirnov confidence bands for \(F\) and \(G\) to provide equal-sized joint confidence rectangles for each observed point on the ROC curve that have simultaneous coverage \((1 - \alpha)^2\) for the true points (FPR, TPR).

\textit{Other indices}

Hsieh and Turnbull\textsuperscript{47} suggest using Youden’s index \((= \max_t[\text{sensitivity}(t) + \text{specificity}(t)])\) as a measure of diagnostic accuracy. They propose two nonparametric estimation methods for determining Youden’s index and the corresponding ‘optimal’ decision limit, based on estimating \(F\) and \(G\) with the empirical distribution functions and Gaussian kernel distribution estimates, respectively. However, this approach requires assumptions that limit its applicability: Youden’s index essentially dichotomizes the continuous diagnostic test results at the decision limit where the costs of a false positive and false negative are equal\textsuperscript{5} and it assumes \(F\) and \(G\) have monotone likelihood ratio, which often does not hold in practice (e.g. pairs of normal
and logistic distributions have monotone likelihood ratio only when the scale parameters of the diseased and nondiseased distributions are equal). Two other indices, the projected length of the ROC curve and the area swept out by the ROC curve, have been proposed recently as alternatives to the AUC for continuous diagnostic tests, but their statistical properties have not been worked out.

### 3.4 Incorporating covariate information

External factors can affect the performance of diagnostic tests by influencing the distributions of test measurements for diseased and/or nondiseased subjects. General approaches to incorporating covariate information into ROC analysis include stratification and modelling.

**Stratification**

Sukhatme and Beam use stratification of the diseased or nondiseased (or both) to assess the performance of a diagnostic marker against several types of diseased or nondiseased subjects. They develop methods for estimating and comparing stratum-specific empirical ROC curves and associated nonparametric areas, and for efficient estimation of the population ROC curve and associated area, when the diseased or nondiseased (or both) are stratified. One limitation, pointed out by Le, is that this approach can only be used with one binary or categorical covariate; it is less efficient with a continuous covariate and very impractical when several covariates need to be simultaneously included.

**Modelling**

An alternative approach is to pool data across strata and use a regression modelling approach to adjust for confounding effects. Pepe describes three general approaches to regression analysis of ROC curves with continuous test results: (1) modelling covariate effects on the test result, then calculating the induced covariate effect on the ROC curve; (2) modelling covariate effects on summary measures of accuracy, such as the area under the curve; and (3) directly modelling covariate effects on the ROC curve.

Several methods of the first type have been proposed. Pepe models the continuous test result as belonging to a location-scale family of unspecified form, with mean function depending on both disease status and covariates (including covariates specific to diseased or nondiseased), and variance depending only on disease status; the mean and variance parameters can be estimated by quasi-likelihood methods. In situations where it is reasonable to group the continuous test results into ordered categories, one could apply methods based on the proportional odds model or the continuation ratio model, although it is not clear whether either method can handle covariates that are specific to the diseased or nondiseased subjects; the continuation ratio-based approach has several advantages over the proportional odds-based approach, in that it belongs to the class of generalized linear models, and yields ROC curves that are concave but not necessarily symmetric, and fitted probabilities that are always between 0 and 1.

The second approach is appropriate for settings in which the experimental design permits calculation of an ROC curve and its summary measure \( \theta_k \) (e.g. AUC or partial area) at each observed covariate level \( k = 1, \ldots, K \). Pepe suggests using standard
linear regression techniques to model the expected value of \( H^{-1}_0(\hat{\theta}_k) \) as a linear function of the covariates at level \( k \), where \( H^{-1}_0 \) is a monotone transformation that gives a linear predictor with unrestricted range \((-\infty, \infty)\); e.g. since the range of AUC is restricted to \((0,1)\), the probit or logit transformation could be used to remove this restriction. Disadvantages of this approach are that it cannot incorporate continuous covariates or covariates that are specific to diseased subjects (e.g. disease severity), and it requires sufficient data at each covariate level (or combination of levels) to estimate an ROC curve with adequate precision.

The third approach differs from the first approach in that it models the relationship between test results in the diseased and nondiseased populations, rather than the test result distributions themselves. Pepe\(^{50}\) suggests modelling the ROC curve as \( \text{ROC}(t) = g\{\eta(t), cX\} \), where \( g \) is a known bivariate function in the range \((0,1)\), increasing in \( t \), and \( \eta(t) \) is a parametric function on the domain \((0,1)\). A necessary preliminary step is to estimate the test result distribution given relevant covariates in the population of nondiseased subjects, using parametric or nonparametric methods. Regression parameters are fitted using an estimating equations approach. Choosing \( g \) to be a survivor function and \( \eta \) to be a linear function of \( g^{-1} \) yields an ROC curve of the same form as one generated by assuming a location-scale model for the test results (Method 1 above). However, modelling the ROC curve directly gives much greater flexibility than modelling the test results in that it allows a much broader range of ROC curve models, it can allow the effect of covariates to vary with FPR (interaction), the distributions of test results in the diseased and nondiseased do not need to be from the same family, and the method can model and compare ROC curves for tests with completely different types of test results; on the other hand, currently the method is more difficult to implement, requires special programming, and lacks adequate model-checking procedures and knowledge of large and small sample properties.

Le\(^{34}\) proposed using the Cox proportional hazards regression model with the midrank-based empirical ROC curve to evaluate the effect of covariates on sensitivity in studies where certain covariates, such as disease severity, are available only for the diseased subjects and not the nondiseased subjects; the method can also be used to model specificity when certain covariates are only available for nondiseased subjects. The method can identify important covariates, but the regression coefficients do not have an interpretation similar to relative risk in the survival analysis setting, and the method does rely on proportional hazards and linearity assumptions, which may be strong but testable.

### 3.5 Dichotomous diagnostic tests

Most laboratory tests give a continuous measurement but some tests give a truly dichotomous result. For a single study in which \( n_0 \) nondiseased and \( n_1 \) diseased subjects are given the diagnostic test, sensitivity and specificity can be estimated as described in Section 2, and separate binomial confidence intervals calculated. Coughlin \textit{et al.}\(^{53}\) suggested using a logistic regression model with true disease status as a covariate to model sensitivity and specificity as a function of other covariates. Leisenring \textit{et al.}\(^{54}\) propose a more general approach using separate marginal regres-
sion models for sensitivity and specificity with robust sandwich variance estimators; the method permits incorporation of different covariates for diseased and nondiseased subjects, and can accommodate clustered and unbalanced data.

The literature on a particular dichotomous diagnostic test often contains several accuracy studies, each of which report different and possibly discrepant sensitivity and specificity estimates. Shapiro\textsuperscript{55} reviews quantitative and graphical methods for meta-analysis of such collections of diagnostic accuracy studies.

4 Comparing the accuracy of diagnostic tests

This section reviews methods for comparing the discriminatory accuracy of two or more diagnostic tests; its organization mirrors that of Section 3. The choice of method depends on whether the study used a parallel-group or ‘unpaired’ design, in which each diagnostic test is applied to a different group of subjects, or a paired design, in which the diagnostic tests are applied to the same subjects; the paired design is more efficient, but requires analysis methods that properly handle the within-subject correlation in the data. Many of the methods for correlated data can also be used with unpaired data. Beam\textsuperscript{56} recently reviewed methods for analysing correlated data in ROC studies, with an emphasis on clustered ordinal data; however, many of the conceptual considerations and the nonparametric methodological strategies in his paper are also applicable to continuous data.

4.1 Comparing ROC curves

Parametric and semiparametric ROC curves can be compared using hypothesis tests and confidence intervals for differences in the ROC curve parameters. Metz and Kronman\textsuperscript{57} provided normal-theory methods for comparing unpaired binormal ROC curves. For paired data, Metz et al.\textsuperscript{58} assume the latent test result distributions are jointly bivariate normal, and derive hypothesis tests for equality of the binormal ROC curve parameters; Metz et al.\textsuperscript{59} extend this methodology to partially-paired designs, in which some but not all subjects were tested with both diagnostic tests. The method assumes that the paired and unpaired portions of the sample are statistically independent and are representative of the same population, and that failure to obtain data from both tests does not bias the case sample.

Three different nonparametric approaches to comparing ROC curves have been proposed that would be useful in cases where the ROC curves cross or where the AUC does not seem to adequately reflect the situation. Campbell\textsuperscript{23} provides a test of the hypothesis of no difference between two correlated ROC curves based on the sup norm and studies the distribution of the test statistic using the bootstrap. Venkatraman and Begg\textsuperscript{60} provide a distribution-free permutation test procedure for testing whether two ROC curves based on paired continuous test results are identical at all operating points; further work is needed to extend this method to unpaired data. Moise et al.\textsuperscript{61} suggest a hypothesis test based on the integrated absolute difference between two correlated ROC curves (i.e. the area between the two ROC curves), with standard error
estimated by bootstrapping. Simulation studies conducted by each author found that each of the three methods had superior power to the usual AUC-based test when the ROC curves crossed, but the three methods have not been compared with each other.

4.2 Comparing summary accuracy indices

**AUC and partial AUC**

Wieand et al. gave a parametric test for difference in AUC for two paired tests with normally-distributed test results; the asymptotic variance of the test statistic was obtained using the delta method. Hanley and McNeil provided a nonparametric test of difference in trapezoidal AUC for the unpaired design. Hanley and McNeil developed a semiparametric test for comparing nonparametric areas under correlated ROC curves that accounts for the correlations induced by the paired design assuming an underlying binormal model holds. DeLong et al. developed a fully nonparametric approach to comparing correlated AUC, based on the theory of U-statistics, in which all of the covariance terms are estimated nonparametrically, leading to an asymptotically normal test statistic. Mossman suggests using resampling methods such as the bootstrap and jackknife for hypothesis testing to compare independent or correlated AUCs (or other summary indices).

Parametric normal-theory methods for comparing diagnostic tests with respect to partial AUC are also available. Wied et al. gave a class of nonparametric statistics for comparing two correlated or independent ROC curves with respect to average sensitivity over a restricted range of specificity (partial area), with test procedures and confidence intervals based on asymptotic normality; this class includes the full AUC and the sensitivity at a common specificity as special cases.

**Sensitivity at common specificity**

Greenhouse and Mantel give parametric and nonparametric hypothesis tests for comparing sensitivities of continuous tests at a prespecified common specificity; methods are given for both independent and paired data, and all methods correctly take into account the uncertainty in the estimated FPR, which adds an additional variance component to the uncertainty in the estimated sensitivity. Linnet modifies their methods to produce a combined parametric/nonparametric procedure that is appropriate when the test measurements are approximately normally distributed for nondiseased subjects but non-normally distributed for diseased subjects. He also points out that in the medical literature it is common to use the simple binomial test for difference in proportions, rather than the correct Greenhouse–Mantel procedures, to compare diagnostic tests with respect to sensitivity at an estimated specificity of 2.5%, i.e. at the upper limit of each test’s reference range (the central interval containing 95% of the test results of the nondiseased population, discussed elsewhere in this issue). Linnet shows that using the simple binomial test in this way, which ignores the uncertainty in the estimated FPR, can increase its type I error to about seven times the nominal value of 0.05. Li et al. provide tests for equality of two independent ROC curves at a fixed FPR and over a range of FPR values based on a vertical shift quantile comparison function.
Beam and Wieand\textsuperscript{65} provide a method for comparing one or more continuous diagnostic tests to one discrete diagnostic test. The comparison of areas or partial areas under the curve in this setting may be inappropriate, because even if the sensitivities of a continuous and a discrete test are comparable at all decision limits available to the discrete test, the concavity of the ROC curve of the continuous test ensures it will have larger area under the curve than the discrete test. They avoid this problem by comparing the tests only at one or more decision limits available to the discrete test, i.e. comparing with sensitivity at a common specificity (or average sensitivity across a set of common specificities). Their method is related to the Greenhouse–Mantel method, but it differs in that the common specificity cannot be arbitrarily chosen but rather must be estimated due to the discreteness of the test; on the other hand, the Beam-Wieand method permits comparison of more than two diagnostic tests, provided only one is discrete, and it allows comparison of sensitivity at more than one specificity of interest.

### 4.3 Incorporating covariate information

Sukhatme and Beam\textsuperscript{49} show how to compare two diagnostic tests when the diseased or nondiseased (or both) are stratified. Pepe\textsuperscript{50} points out that diagnostic tests can be compared using regression approaches that model the ROC curve or a summary index (e.g. AUC) by including a test indicator as one of the covariates in the model; however, regression approaches that model the test results themselves can only be used for comparison if the results of the two tests can be modelled sensibly in the same regression model.

### 4.4 Dichotomous diagnostic tests

Comparison of sensitivity and specificity of two dichotomous diagnostic tests based on a paired design can be done using McNemar’s test for paired binary data.\textsuperscript{54} The marginal regression formulation of Leisenring \textit{et al.}\textsuperscript{54} described in Section 3 can be used to compare sensitivities and specificities of two or more diagnostic tests, even if all tests are not carried out on all subjects, and it can accommodate clustered data and adjust for covariates; this method corresponds to McNemar’s test when two diagnostic tests are given to all subjects in a simple paired design. Lachenbruch and Lynch\textsuperscript{66} generalize McNemar’s test to assess equivalence of sensitivity and specificity of two-paired dichotomous tests.

### 5 Methods for context-specific interpretation

This section discusses three aspects of context-specific interpretation of diagnostic test results: selection of an optimal decision limit to define a positive test result, estimation of the likelihood ratio needed to convert an individual test result into a posterior probability of disease, and context-specific assessment of diagnostic test performance.

#### 5.1 Selection of an optimal decision limit

In screening situations, it is often desired to use the diagnostic test to determine whether or not a subject needs further work-up or treatment. Selection of a decision limit at which to dichotomize a continuous diagnostic test results as positive (further
work-up or treatment required) or negative (no further work-up required) is an optimization problem; the appropriate method depends on the specific objective and data available and, clearly, many different optimization criteria could be used. Two examples are described here.

One possible objective is to minimize the expected costs of misclassification of diseased and nondiseased subjects, in which case the Bayes minimum cost decision rule is optimal. This decision rule classifies a subject as diseased or nondiseased according to whether or not the likelihood ratio $f(X|D^+)/f(X|D^-)$ of the test result $X$ exceeds a fixed decision limit

$$Q = \frac{(1 - P(D+))R_-}{P(D+)}R_+$$

where $P(D+)$ is the prior probability of disease, and $R_-$ and $R_+$ are the regrets (difference between cost of correct and incorrect decision) associated with classifying a subject as nondiseased and diseased, respectively. Strike describes two methods for deriving the optimal decision limit $X_Q$: if the test result distributions for both diseased and nondiseased are assumed normally distributed, the decision limit $X_Q$ can be calculated algebraically by solving the equality $f(X|D^+)/f(X|D^-) = Q$ for $X$. Alternatively, the decision limit can be determined graphically from the ROC curve, using the fact that the slope of the ROC curve at any decision limit $X$ equals the likelihood ratio at that decision limit; $X_Q$ is the decision limit corresponding to the point where the line with slope $Q$ is tangent to the ROC curve.

As an alternative to the cost-minimizing approach, Somoza and Mossman suggest selecting a prevalence-specific decision limit that maximizes diagnostic information (in the information theoretic sense) in the particular clinical situation. This method is appropriate when the objective of testing is to reduce overall uncertainty about the patient’s disease status, rather than to balance the error rates based on their respective costs, and in situations where misclassification costs are open to dispute or are difficult to estimate precisely.

5.2 Estimation of the likelihood ratio

Several methods have been proposed for estimating the likelihood ratio $L(X)$ corresponding to an individual test result $X$. If the ROC curve of the diagnostic test is available for a population similar to the population of interest, the likelihood ratio can be estimated graphically as the slope of the tangent to the ROC curve at the point corresponding to the test result $X$; note that this method requires knowledge of the actual test measurements corresponding to each point on the ROC curve, which may not be included in published papers.

For the case where test results for both diseased and nondiseased subjects are normally distributed, Strike directly calculates the likelihood ratio corresponding to an individual test result and gives a normal-theory standard error and large-sample confidence interval for the posterior probability of disease. If one or both test result distributions are non-normal and cannot both be transformed to normality by a single transformation, he suggests using kernel density estimation to obtain separate smooth estimates of the nondiseased and diseased test result distributions, then calculating the likelihood ratio at the test result $X$ as the ratio of the kernel density estimates.
and \( f(X|D=) \) and \( f(X|D=) \); however, he does not give a variance estimate or confidence interval for the resulting estimated posterior probability of disease. Linnet\(^{68}\) suggested a different nonparametric method based on estimating relative frequency polygons of the test result distributions.

Albert\(^{69}\) suggests estimating the likelihood ratio \( L(X) \) using logistic regression of true disease status on the test result \( X \). This approach assumes that the likelihood ratio can be well-approximated by an exponential function; in particular, \[ L(X) = \exp \left( \alpha_0 + \alpha_1 X + \log \left( \frac{n_{D=}/n_{D=}}{n_{D=}/n_{D=}} \right) \right), \]

where \( \alpha_0 \) and \( \alpha_1 \) are the usual logistic regression coefficients and \( n_{D=} \) and \( n_{D=} \) are the numbers of nondiseased and diseased subjects, respectively. This method can handle dichotomous or continuous diagnostic tests, and can easily incorporate covariate information. Albert and Harris\(^{70}\) and Simel \textit{et al.}\(^{71}\) discuss confidence intervals for likelihood ratios.

### 5.3 Context-specific assessment of diagnostic test performance

The methods for assessing discriminatory accuracy of a diagnostic test reviewed in Sections 3 and 4 only provide information about the ability of the test to distinguish between diseased and nondiseased patients; however, they do not directly indicate how well the test result for an individual patient predicts the probability of disease in that patient, which will vary depending on the prevalence of disease. This section reviews methods for assessing context-specific diagnostic test performance. The results of all these methods are prevalence-dependent and may not be directly generalizable to other settings.

Linnet\(^{72}\) suggests assessing diagnostic test performance based on the accuracy of estimated posterior probabilities. He proposes using a strictly proper scoring rule, which is a function of the posterior probabilities that has maximum expected value at the underlying true probabilities. His approach is to estimate the posterior probability as a function of diagnostic test result based on a training set of observations, then subsequently to assess the score on a new set of test observations. Bootstrap standard error estimation and bias correction allows the same subjects to serve as training and test sets. Comparison of two diagnostic tests can be done using the bootstrap.

Somoza and Mossman\(^{67}\) suggest comparing diagnostic tests by determining the prevalence-specific decision limit for each test that maximizes diagnostic information (i.e. maximizes the reduction in uncertainty about disease status), then comparing the tests at their optimal decision limits, which may occur at different FPR. This method may be more useful than comparing partial AUC or sensitivity at fixed specificity when ROC curves cross; in this situation, comparison of partial areas or TPR at fixed FPR may be difficult if the ranges of FPR where the tests perform optimally do not coincide. Many other related approaches have been proposed in the medical decision making literature to incorporate the consequences (in terms of risks and benefits) of subsequent diagnostic decisions based on the test result into the evaluation of diagnostic test performance. A review of this literature is beyond the scope of this paper; a good starting place is a recent paper by Moons \textit{et al.}\(^{73}\)

Bloch\(^{74}\) provides two alternative methods for comparing two diagnostic tests in a paired design where the decision limit for each test and the costs of a false positive and false negative are specified \textit{a priori}. One method is based on comparing risks of the two tests, and the other on comparing kappa coefficients.
6 Concluding remarks

The assessment and interpretation of laboratory diagnostic tests has received increasing attention in the statistical literature in recent years. Many new methods have been proposed, but the relative advantages and disadvantages of different methods have not yet been assessed. More work is needed in identifying the most promising methods and incorporating them into widely available statistical software.

This paper has reviewed methods for assessing, comparing, and interpreting individual diagnostic tests. In actual medical practice, however, individual diagnostic test results are not interpreted in a vacuum, but are combined by the physician with all other available information to arrive at a diagnosis. A large and growing literature exists on computer-based methods for diagnosis and prognosis, including both statistical methods such as logistic regression and recursive partitioning, and nonstatistical methods such as expert systems and neural networks. Strike\textsuperscript{1} and Begg\textsuperscript{4} give brief overviews of this area and include several references.

Despite the availability of methodology for interpreting diagnostic test results, relatively few physicians actually use formal methods such as sensitivity, specificity, likelihood ratios, or Bayes theorem when interpreting diagnostic test results. A recent telephone survey of 300 practicing physicians found that only 3% reported calculating posterior probability of disease using Bayes theorem, 1% reported using ROC curves, and 1% reported using likelihood ratios when evaluating diagnostic test results.\textsuperscript{75} On the other hand, 84% reported using a test’s sensitivity and specificity when conducting diagnostic evaluations, primarily when interpreting test results; however, when asked how they used these indexes, 95% described an informal method that amounted to estimating the test’s posterior error probabilities (i.e. proportion of patients who test positive that subsequently are found to be nondiseased, and proportion of patients who test negative that subsequently are found to be diseased) based on patients in their own practice, rather than using published sensitivity and specificity data. Additional efforts to make the formal diagnostic test interpretation methods more user-friendly and explain their use, like the two recent papers by Jaeschke et al.\textsuperscript{76,77} are clearly needed.

Acknowledgement

Supported in part by National Institute of Allergy and Infectious Diseases, cooperative agreement number U01-AI411.

References


Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 1975; 12: 387–415.


McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating


38 Walsh S. Limitations to the robustness of binormal ROC curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power. *Statistics in Medicine* 1997; 16: 669–79.


74 Bloch DA. Comparing two diagnostic tests against the same ‘gold standard’ in the same sample. *Biometrics* 1997; 53: 73–85.
Copyright of Statistical Methods in Medical Research is the property of Arnold Publishers and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.