ROC Curves in Clinical Chemistry: Uses, Misuses, and Possible Solutions

Nancy A. Obuchowski,1* Michael L. Lieber,1 and Frank H. Wians, Jr.2

Background: ROC curves have become the standard for describing and comparing the accuracy of diagnostic tests. Not surprisingly, ROC curves are used often by clinical chemists. Our aims were to observe how the accuracy of clinical laboratory diagnostic tests are assessed, compared, and reported in the literature; to identify common problems with the use of ROC curves; and to offer some possible solutions.

Methods: We reviewed every original work using ROC curves and published in Clinical Chemistry in 2001 or 2002. For each article we recorded phase of the research, prospective or retrospective design, sample size, presence/absence of confidence intervals (CIs), nature of the statistical analysis, and major analysis problems.

Results: Of 58 articles, 31% were phase I (exploratory), 50% were phase II (challenge), and 19% were phase III (advanced) studies. The studies increased in sample size from phase I to III and showed a progression in the use of prospective designs. Most phase I studies were powered to assess diagnostic tests with ROC areas ≥0.70. Thirty-eight percent of studies failed to include CIs for diagnostic test accuracy or the CIs were constructed inappropriately. Thirty-three percent of studies provided insufficient analysis for comparing diagnostic tests. Other problems included dichotomization of the gold standard scale and inappropriate analysis of the equivalence of two diagnostic tests.

Conclusion: We identify available software and make some suggestions for sample size determination, testing for equivalence in diagnostic accuracy, and alternatives to a dichotomous classification of a continuous-scale gold standard. More methodologic research is needed in areas specific to clinical chemistry.

ROC curves have been widely accepted as the standard method for describing and comparing the accuracy of radiologic imaging and other medical diagnostic tests (1, 2). The accuracy of a diagnostic test is characterized by the sensitivity (detection of disease when disease is truly present) and specificity (recognition of disease absence when the disease of interest is truly absent). A ROC curve displays the sensitivity of a diagnostic test over all possible false-positive rates (FPRs); the FPR is the false detection of disease, or 1 – specificity. The ROC curve and the measures of accuracy derived from it have many advantages over other measures of accuracy, including the following: (a) they are independent of the prevalence of disease; (b) two or more diagnostic tests can be compared at any or all FPRs; and (c) summary measures of accuracy, such as the area under the ROC curve (AUCROC), incorporate both components of accuracy, i.e., sensitivity and specificity, into a single measure of accuracy.

Not surprisingly, ROC curves are used quite often by clinical chemists. We reviewed every original work that used ROC curves as part of the statistical analysis and was published in Clinical Chemistry in 2001 or 2002. We found 58 such articles. A review of these papers allows us to observe how the accuracy of clinical laboratory diagnostic tests are assessed, compared, and reported in the literature; and to identify common problems with the use of ROC curves and offer some possible solutions.

Methods

We searched online Clinical Chemistry issues from January 2001 to December 2002 for the following key phrases: “ROC curve”, “ROC analysis”, and “Receiver Operating
Characteristic”. We screened the results to rule out non-
original works (e.g., reviews) and papers that mentioned
ROC curves in the text or reference list but did not apply
them in the actual work. One of us read the papers
published in 2001 and another of us the papers published
in 2002, recording the following for each: phase of the
research, prospective or retrospective design, sample size,
presence/absence of confidence intervals (CIs) for mea-
sures of accuracy, nature of the statistical analysis (i.e.,
reporting the accuracy of a single test, comparison of two
or more tests to determine which test(s) is superior, or
assessment of the equivalence/noninferiority of a new
test to an established test), and any major problems with
the statistical analysis.

We defined the phase of the research as did Zhou et al.
(2). Phase I is the “exploratory phase”, in which a new test
is first evaluated in a clinical setting to determine whether
the test has any ability to discriminate diseased from
nondiseased patients (or, sometimes, to discriminate be-
tween two groups of diseased patients; e.g., iron defi-
ciency anemia vs anemia of chronic disease or liver
cirrhosis vs chronic hepatitis). ROC curves are often used
in these studies to test whether the AUCROC exceeds 0.5.
If not, then no further assessment of the diagnostic test
is warranted. Phase II in the clinical assessment of the
accuracy of a diagnostic test is called the “challenge
phase”, in which the accuracy of one or more tests is
estimated for difficult cases to determine for whom the
test may fail and to perhaps identify ways to improve it
before further assessment. Finally, phase III is the “ad-
vanced phase”, in which the accuracy of one or more tests
is estimated and compared for a well-defined and gener-
izeable clinical population. From these studies we can
estimate how the test will perform in clinical practice; in
contrast, phase I studies tend to overestimate accuracy
(because easy-to-diagnose patients are often selected for
the study sample), and phase II studies tend to underes-
timate accuracy (because difficult cases are often selected
for the study sample).

We classified a study as retrospective if the patients in
the study sample were selected for the study based on
their known disease status. A prospective design, in
contrast, is one in which the patients are recruited based
on their signs and symptoms, before determining their
true disease status.

Results

Review of Literature

The findings of our 2-year review of Clinical Chemistry are
summarized in Table 1. We identified 18 (31%) phase I
studies in which ROC curves were used. These studies
were almost always retrospective in design (89%). Often
the phase I studies included patients with well-established
disease whose test results were contrasted with those from healthy volunteers. The phase I studies in
Clinical Chemistry tended to be moderate in size, with a
median of 78 diseased patients (range, 13–472) and 65
nondiseased patients (range, 11–940).

Phase II studies in which ROC curves were used were
the most common type of accuracy study in Clinical
Chemistry, with 29 such studies (50%), including 18 (62%)
retrospective in design and 11 (38%) prospective in de-
sign. These studies were larger than the phase I studies,
with a median of 88 diseased patients (range, 18–442) and
99 nondiseased patients (range, 8–730).

We found 11 phase III studies in which ROC curves
were used, with 64% being prospective. The median size
of these phase III studies was 140 diseased patients
(range, 15–721) and 174 nondiseased patients (range, 38–1366). Thus, the diagnostic accuracy studies in Clinical
Chemistry showed an expected increase in sample size

| Table 1. Results of 2-year review of Clinical Chemistry articles using ROC analysis. |
|---------------------------------|---------------------------------|---------------------------------|
| Total no. of studies | Phase I (Exploratory) | Phase II (Challenge) | Phase III (Advanced) |
| Design | | | |
| Retrospective, n | 16 | 18 | 4 |
| Prospective, n (%) | 2 (11%) | 11 (38%) | 7 (64%) |
| Sample size* | | | |
| Diseased | 78 (13–472) | 87.5 (18–442) | 140 (15–721) |
| Nondiseased | 64.5 (11–940) | 99 (8–730) | 173.5 (38–1366) |
| Reported CIs? | | | |
| Yes, n (%) | 8 (44%) | 23 (79%) | 8 (73%) |
| No, n | 10 | 6 | 3 |
| Nature of statistical analysis | | | |
| Single test, n (%) | 1 (6%) | 4 (14%) | 5 (45%) |
| Two or more tests, n | | | |
| Detect differences | 16 | 24 | 6 |
| Show equivalence | 1 | 1 | 0 |

* Median (minimum−maximum) sample sizes.
from phase I to III and a progression to more prospective designs.

COMMON PROBLEMS
The problems we identified with the use of ROC curves for comparing or reporting diagnostic accuracy in the above studies are summarized in Table 2. Most of the articles we reviewed (79%) included a comparison of two or more tests to determine which test(s) had superior diagnostic accuracy. In 40% of such articles, we found no statistical analysis of the comparison between tests. The authors of some articles reported the AUCROC of the tests and the CIs for the areas were but gave no statistical test for determining whether the ROC areas differed. In other studies, a statistical comparison of the accuracies of two tests was carried out, but the diagnostic tests had been performed on the same patients (paired sample design) and this pairing was not taken into account in the statistical analysis.

We found five reports in which the true disease status of the patient (i.e., the findings of the gold standard test) was not a simple binary outcome (e.g., celiac disease present or absent), but rather the true disease status was represented by a quantitative measurement. Some examples of gold standard tests that yield a continuous-scale outcome are insulin clearance to measure glomerular filtration rate and SPECT to measure left ventricular ejection fraction (LVEF). To perform a traditional ROC analysis, the true disease status of patients must be dichotomous. One approach is to choose a single cutpoint, for example, LVEF < 40%, such that patients with values less than the cutpoint are considered diseased, whereas others (who may only have a LVEF one percentage point higher) are considered nondiseased. Another approach is to choose two cutpoints. Patients with values less than the lower cutpoint or greater than the higher cutpoint are compared in the study, whereas patients with values in the gray zone (i.e., between the two cutpoints) are excluded from the study. An example is a study of indicators of iron status used to discriminate patients with iron deficiency anemia from anemia of chronic disease (3). The gold standard is ferritin concentrations measured by the ferrozine method. Concentrations < 20 μg/L are considered iron deficiency anemia, and concentrations > 240 μg/L (for women) and > 375 μg/L (for men) are considered anemia of chronic disease; patients with concentrations in the middle are excluded. The choice of cutpoints for dichotomizing the gold standard is critical, but the choice is often quite arbitrary. We show in a later section how the choice of cutpoints for the gold standard affects the AUCROC for the diagnostic test and how the two-cutpoint approach can overestimate accuracy.

Most checklists for studies reporting the diagnostic accuracy of medical tests (4–7) have noted the importance of reporting CIs for measures of test accuracy. Among the 58 articles we reviewed, 39 (67%) included CIs. Sometimes, however, the CIs were constructed inappropriately.

In one report, diagnostic accuracy was estimated from more than one observation from the same patient, so-called “clustered” data. An example of clustered data is measurements from the two ears of the same patient. Data from the same patient are inherently correlated to some degree. Often the correlation is small, but even a small amount of intracluster correlation will lead to incorrect P values (usually inappropriately small P values).

There were two articles that attempted to show that a new test was at least as accurate (and maybe more accurate) as an existing test, so-called “noninferiority” studies. In only one article, however, was a statistical test used that was appropriate for testing a hypothesis of equivalency. In the other, the authors concluded that the accuracies of the tests were equivalent because the difference between the ROC areas was not statistically significant at the 0.05 level. This approach is not valid for assessing equivalence or noninferiority because the risk is high for incorrectly concluding equivalence, particularly when the sample size is small.

Finally, from our review, it appeared that some studies were underpowered (too small of a sample size). In a later section we offer some direction in determining the appropriate sample size for different types of studies.

EXAMPLE OF TYPICAL STUDY WITH APPROPRIATE ROC ANALYSIS
Before addressing some possible solutions to these problems, we want to cite an example of a typical study seen in Clinical Chemistry with an appropriate ROC analysis. Martinez et al. (8) performed a phase II prospective study comparing three tests—total prostate-specific antigen (PSA), PSA complexed to α1-antichymotrypsin (PSA-α1-ACT), and the ratio of PSA-α1-ACT to total PSA (PSA-α1-ACT:PSA)—for the differential diagnosis of prostate cancer and benign prostatic hyperplasia. They recruited consecutive patients who had been referred for prostatic evaluation. A total of 146 patients met the eligibility criterion of a total PSA between 10 and 30 μg/L. All patients underwent biopsy. The authors estimated the AUCROC for each diagnostic test and compared the areas using nonparametric methods for paired data (because
the three diagnostic tests were performed on all study patients) through MedCalc (9). The authors reported 95% CIs for the ROC areas and cited the observed sensitivity and specificity at various cutoff points. The authors found that PSA-α1-ACT:PSA ratio had superior accuracy (i.e., statistically significant at the 0.05 level) for patients with total PSA values in both the ranges 10–20 μg/L and 20–30 μg/L. They concluded that additional prospective studies with large numbers of patients were needed to confirm their findings.

Possible Solutions to Common Problems with ROC Analysis

Constructing CIs for Sensitivity and Specificity

In constructing 95% CIs for the sensitivity and specificity of a diagnostic test, it is important to recognize that sensitivity and specificity are not independent of one another. Thus, the width of the CI for one is affected by the uncertainty in estimating the other. We illustrate the methods for CI construction with the following example. Suppose we are evaluating the diagnostic accuracy of a new commercial ELISA in detecting anti-cyclic citrullinated peptide antibodies for the diagnosis of rheumatoid arthritis. We find that, at a cutoff value of 50 units, sensitivity is 52% and specificity is 94%. We want to construct 95% CIs around each of these two parameters. However, rather than approaching this as if sensitivity and specificity (and their associated CIs) are independent, we have to use methods that correctly treat these two measures as being correlated with one another. Similarly, if we want to construct a CI for the sensitivity of this ELISA at a fixed FPR of 10%, we need to take into account the fact that the true-positive rate (TPR; i.e., sensitivity) and FPR are correlated.

Zhou et al. (2) discuss several approaches to CI construction. We present here a parametric approach to constructing CIs for sensitivity at a particular FPR [or analogously, for constructing CIs for specificity at a particular false-negative rate (FNR)]. This method assumes that the test results, or a transformation of them, follow a binormal distribution (that is, one normal distribution for test results of patients with disease and another normal distribution for test results of patients without disease). We use the following notation to describe the binormal distribution:

\[ a = (\mu_1 - \mu_0) / \sigma_1 \text{ and } b = \sigma_0 / \sigma_1 \]

where \( \mu_1 \) and \( \mu_0 \) are the means and \( \sigma_1 \) and \( \sigma_0 \) are the standard deviations of the normal distributions for patients with and without disease, respectively.

The sensitivity at a particular FPR = \( \epsilon \) can be estimated from \( \hat{Se}_{FPR} = \epsilon = 1 - \Phi(bZ_e - \hat{a}) \), where \( \Phi \) is the cumulative normal distribution function, and \( Z_e \) is such that \( \Phi(Z_e) = \epsilon \). For example, if we are interested in the sensitivity at a FPR of 10% (i.e., \( \epsilon = 0.10 \)), then \( Z_e = 1.28 \). If we obtain estimates of \( \hat{a} = 0.8 \) and \( \hat{b} = 1.2 \), then \( \hat{Se}_{FPR = 0.10} = 1 - \Phi(1.2 \times 1.28 - 0.8) = 0.23 \).

To get the variance of \( \hat{Se}_{FPR = \epsilon} \) and its CI, we transform \( \hat{Se}_{FPR = \epsilon} \) to \( Z_e \) as follows:

\[ \hat{Z}_e = \hat{b}Z_e - \hat{a} \]

Then

\[ \text{Var}(\hat{Z}_e) = \hat{b}^2 \text{Var}(Z_e) + \text{Var}(\hat{a}) - 2\hat{b}\text{Cov}(\hat{a}, \hat{b}) \]

Estimates of the variance of parameters \( \hat{a} \) and \( \hat{b} \) and their covariance are available from programs such as ROCKIT (10).

Assuming asymptotic normality, the 100(1 − \( \alpha \))% CI for the transformed sensitivity corresponding to a particular FPR of \( \epsilon \) is:

\[ \text{LL}(\hat{Z}_e) = \hat{Z}_e - z_{(\alpha/2)} \sqrt{\text{Var}(\hat{Z}_e)} \]

and

\[ \text{UL}(\hat{Z}_e) = \hat{Z}_e + z_{(\alpha/2)} \sqrt{\text{Var}(\hat{Z}_e)} \]

where \( z_{(\alpha/2)} \) is the upper \( \alpha/2 \) percentile of the standard normal distribution. The lower and upper “transformed” confidence limits, LL and UL, for sensitivity can then be calculated by:

\[ 1 - \Phi[\text{LL}(\hat{Z}_e)] \]

and

\[ 1 - \Phi[\text{UL}(\hat{Z}_e)] \]

Continuing with our example, \( \hat{Z}_e = 0.736 \). If we obtain estimates of \( \text{Var}(\hat{a}) = 0.0214 \) and \( \text{Var}(\hat{b}) = 0.0091 \), with \( \text{Cov}(\hat{a}, \hat{b}) = 0.0068 \), then:

\[ \text{Var}(\hat{Z}_e) = \hat{b}^2 \text{Var}(Z_e) + \text{Var}(\hat{a}) - 2\hat{b}\text{Cov}(\hat{a}, \hat{b}) = 0.0189. \]

Subsequently

\[ \text{LL}(\hat{Z}_e) = 0.736 - z_{(0.025)} \sqrt{0.0189} = 0.4666 \]

and

\[ \text{UL}(\hat{Z}_e) = 0.736 + z_{(0.025)} \sqrt{0.0189} = 1.005. \]

Finally, transforming the confidence limits into a 95% CI for sensitivity at a fixed FPR of 10%, we obtain:

\[ 1 - \Phi[\text{LL}(\hat{Z}_e)] = 1 - \Phi(0.4666) = 0.32 \]

and

\[ 1 - \Phi[\text{UL}(\hat{Z}_e)] = 1 - \Phi(1.005) = 0.16 \]

The 95% CI for sensitivity is therefore (0.16–0.32).

Note that there is software for constructing a CI for sensitivity at a fixed FPR (or specificity at a fixed FNR), as well as for testing for a difference between two diagnostic tests at a fixed FPR (or FNR) (10).
curves and their areas (AUCs) for paired (same patients) or unpaired (different patients) designs, from clustered data or one observation/patient data set, using parametric or nonparametric methods. There is also software for estimating and comparing the partial areas under ROC curves (i.e., ROC area in the specified FPR range of \( c_1 \) to \( c_2 \)) and for estimating sample size or power. See the recent review by Stephan et al. (23) for a comparison of some of the available ROC software.

**COMPARING NONINFERIORITY OF THE ACCURACY OF A NEW TEST WITH THAT OF A STANDARD TEST**

When developing a new diagnostic test, there are situations in which the new test needs to have accuracy as good as, but not necessarily better than, the accuracy of an existing, or standard, test for the new test to replace the standard test. For example, the new test might be safer, easier, or quicker to perform, or it may be less expensive than the standard test.

For these types of studies, we want to make sure that the accuracy of the new test is not inferior to the accuracy of the standard test before replacing the standard test. We need to test the hypotheses:

**Null hypothesis:** \( \theta_s - \theta_N \geq \Delta_M \)

vs

**Alternative hypothesis:** \( \theta_s - \theta_N < \Delta_M \)

where \( \theta_s \) is the accuracy of the standard test, \( \theta_N \) is the accuracy of the new test, and \( \Delta_M \) is the smallest difference in accuracy that is unacceptable. Note that \( \Delta_M \) should be specified in the planning phase of the study (i.e., not after examining the data; this can lead to bias).

Consider an example. Suppose a standard test has an AUC of 0.90. A new, quicker, and less expensive test has been developed; we determine that the new test must have an AUC of 0.85 or greater to replace the standard test. Then, \( \Delta_M = 0.06 \).

An appropriate statistical test for assessing noninferiority is:

\[
z = \left( \frac{\hat{\theta}_s - \hat{\theta}_N - \Delta_M}{\sqrt{\text{var}(\hat{\theta}_s - \hat{\theta}_N)}} \right)
\]

where \( \hat{\theta}_s \) and \( \hat{\theta}_N \) are the estimates of accuracy for the standard and new test, respectively, and \( \text{var}(\hat{\theta}_s - \hat{\theta}_N) \) is an estimate of the variance of the difference between accuracies of the two tests. Any measure of accuracy can be used (e.g., AUC or partial area under the ROC curve), and software (9–22) is available for estimating the accuracies of the tests and the variance of their difference. For a type I error rate of 5%, we would reject the null hypothesis if \( z \) is less than \(-1.645\); otherwise, we do not reject the null hypothesis and thus have insufficient evidence to replace the standard test.

ROC ANALYSIS WHEN THE GOLD STANDARD DOES NOT YIELD A DICHTOMOUS OUTCOME

In this section we show through a simulation study the effects on the AUC of dichotomizing the results of a gold standard when that gold standard does not produce binary results (i.e., disease is present or absent) but rather the results are on a continuous scale. We then offer a new method for estimating the AUC that does not require dichotomizing the gold standard.

In a simulation study, we first investigated how the AUC is affected when a single cutpoint for separating diseased and nondiseased patients is varied. We generated 1000 samples of 1000 observations each from a bivariate normal distribution (i.e., we used a large sample size so that differences could be attributed to systematic bias and not to random variability). One of the random variables from this bivariate distribution was designated as the gold standard, and the other random variable was considered the diagnostic test. We set the correlation between these two variables equal to 0.75. We then varied the cutpoint of the first random variable, such that observations with values above the cutpoint were considered “diseased” and observations below the cutpoint were considered “nondiseased”. As shown in Table 3, the nonparametric estimates (24) of the ROC area varied as the cutpoint varied; the differences were small but were statistically significant. In other words, although the underlying relationship between the gold standard and diagnostic test did not change, the estimates of the AUC did change because of the unnatural dichotomization of the outcomes.

Now consider the effect of using two cutpoints to identify two populations—one with values below the first cutpoint and the other above the second cutpoint—omitting from the analysis the patients with values between the two cutpoints. We used simulated data as described previously. As shown at the bottom of Table 3, the estimated AUC increased greatly when the cases in the gray zone were omitted, and the wider the gray zone, the greater the increase in the ROC area. Clearly, dichotomizing a continuous-scale gold standard introduces bias into the estimates of the accuracy of a diagnostic test.

A nonparametric estimate of accuracy, analogous to

<table>
<thead>
<tr>
<th>Cutpoint(s)</th>
<th>Average ROC area*</th>
<th>SE of ROC area*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.8559</td>
<td>0.0116</td>
</tr>
<tr>
<td>0.50</td>
<td>0.8623</td>
<td>0.0120</td>
</tr>
<tr>
<td>0.84</td>
<td>0.8737</td>
<td>0.0128</td>
</tr>
<tr>
<td>1.65</td>
<td>0.9172</td>
<td>0.0167</td>
</tr>
<tr>
<td>−0.5 and +0.5</td>
<td>0.9460</td>
<td>0.0083</td>
</tr>
<tr>
<td>−1.0 and 1.0</td>
<td>0.9874</td>
<td>0.0046</td>
</tr>
</tbody>
</table>

* Average over 1000 datasets.

a Average SE over 1000 datasets; for each dataset we estimated the SE using the method of DeLong et al. (10).
the nonparametric ROC area estimate, can be constructed without any cutpoints. The formula is given in Eq. 1.

In words, we compare the test scores of each patient with the test scores of all other patients, assigning a weight of 1 if the test result of the patient with a higher (more serious) outcome value exceeds the test result of the patient with the lower (less serious) outcome value. If two patients have the same test result or the same gold standard outcome, a weight of 0.5 is assigned. Otherwise, a weight of 0 is given. The sum of these weights divided by the number of pairs is the nonparametric estimate of accuracy. The interpretation is similar to that for the ROC area estimate, can be constructed without any cutpoints. The formula is given in Eq. 1.

\[
\hat{\theta} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi(x_{it}, x_{jt})
\]

where \( i \neq j \), \( n \) is the total number of patients in the study sample, \( x_{it} \) is the test result of the \( i \)th patient with gold standard outcome \( t \), \( x_{jt} \) is the test result of the \( j \)th patient with gold standard outcome \( s \), and:

\[
\psi = 1 \text{ if } t > s \text{ and } x_{it} > x_{jt} \text{ or } s > t \text{ and } x_{it} > x_{jt}
\]

\[
\psi = 0.5 \text{ if } t = s \text{ or } x_{it} = x_{jt}
\]

\[
\psi = 0 \text{ otherwise.}
\]

Note that the definition of \( \psi \) can be modified if the scales of the gold standard and diagnostic test are inversely related.

This type of accuracy estimate has been used previously to assess the prognostic ability of models for predicting survival time (25). A CI for this accuracy measure or a CI for the difference between two estimates of accuracy can be obtained by bootstrapping (26).

**COMPUTING THE REQUIRED SAMPLE SIZE FOR PHASE I, II, AND III STUDIES**

In this section we discuss some possible approaches to sample size calculation for studies of diagnostic test accuracy. We discuss sample size according to the phase of the study because the goals of the phases differ, and thus the sample size requirements differ. The methods for sample size calculation that we present here are based on large-sample theory for normally distributed data (27);

the reader should be mindful of this when using these formulae.

In phase I studies, the usual goal is to determine whether the diagnostic test has any ability to discriminate diseased patients from controls. A useful hypothesis on which to base sample size estimation is whether the AUC exceeds 0.5. The null hypothesis is that the AUC equals 0.5, vs the alternative hypothesis that the AUC is >0.5 (one-sided test):

null hypothesis: \( \theta \leq 0.5 \)

vs

alternative hypothesis: \( \theta > 0.5 \)

A formula for computing sample size to test these hypotheses is:

\[
n_D = \left[ z_{\alpha} \sqrt{0.0792 \times (1 + 1/\kappa)} + z_{\beta} \sqrt{V(\hat{\theta})} \right]^2 / (\theta - 0.5)^2
\]

where \( V(\hat{\theta}) \) is the variance function of \( \hat{\theta} \), given by:

\[
V(\hat{\theta}) = (0.0099 \times e^{-A^2/2}) \times (5A^2 + 8 + [A^2 + 8]/\kappa)
\]

i.e., the variance of \( \hat{\theta} \) is equal to \( V(\hat{\theta})/n_D \); \( A = \Phi^{-1}(\theta) \times 1.414; \Phi^{-1} \) is the inverse of the cumulative normal distribution function; \( \kappa \) is the ratio of the number of control patients (\( n_C \)) to the number of diseased patients (\( n_D \)) in the study sample (i.e., \( \kappa = n_C/n_D \)); \( \theta \) is the conjectured area under the ROC curve (under the alternative hypothesis); \( z_{\alpha} \) is the upper \( \alpha \)th percentile of the standard normal distribution, where \( \alpha \) is the type I error rate (usually \( \alpha = 0.05 \)); and \( z_{\beta} \) is the upper \( \beta \)th percentile of the standard normal distribution, where \( \beta \) is the type II error rate (often \( \beta = 0.10 \) or 0.20).

In Table 4 we have computed the sample size for a range of values for \( \kappa \) (e.g., \( \kappa = 1 \) means equal numbers of patients with and without the disease in the study sample; \( \kappa = 0.5 \) means twice as many diseased patients as control patients in the study sample; and \( \kappa = 2.0 \) means twice as many control patients as diseased patients in the study sample) and for a range of values for the conjectured AUC. The type I error rate has been set at 0.05; the type II error rate is \( \leq 0.10 \) (power \( \geq 0.90 \)).

**Table 4. Sample size (\( n_C/n_D \)) needed for phase I study.**

<table>
<thead>
<tr>
<th>Conjectured accuracy</th>
<th>Poor (( \theta = 0.60 ))</th>
<th>Fair (( \theta = 0.70 ))</th>
<th>Moderate (( \theta = 0.80 ))</th>
<th>High (( \theta = 0.90 ))</th>
<th>Near perfect (( \theta = 0.95 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \kappa = 0.25 )</td>
<td>84/334</td>
<td>20/80</td>
<td>10/40</td>
<td>10/40</td>
<td>10/40</td>
</tr>
<tr>
<td>( \kappa = 0.50 )</td>
<td>101/201</td>
<td>25/49</td>
<td>10/20</td>
<td>10/20</td>
<td>10/20</td>
</tr>
<tr>
<td>( \kappa = 1.0 )</td>
<td>135/135</td>
<td>33/33</td>
<td>14/14</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>( \kappa = 2.0 )</td>
<td>203/102</td>
<td>50/25</td>
<td>21/11</td>
<td>20/10</td>
<td>20/10</td>
</tr>
<tr>
<td>( \kappa = 4.0 )</td>
<td>339/85</td>
<td>84/21</td>
<td>40/10</td>
<td>40/10</td>
<td>40/10</td>
</tr>
</tbody>
</table>

*When the computed sample size is <10, we recommend that at least 10 diseased patients and 10 controls be included in the study. The sample sizes in this table reflect this recommendation.*
For example, for a balanced design (i.e., equal numbers of patients with and without disease), if the accuracy of the diagnostic test is expected to be fair (e.g., 0.70), then 33 control patients and 33 diseased patients (total of 66 patients) are needed. From Table 1 it looks like the phase I studies in Clinical Chemistry are, in general, reasonably powered for assessing diagnostic tests with AUC ≥ 0.70.

In phase II studies, we often compare the accuracies of two or more diagnostic tests. The study sample usually represents patients difficult to diagnose; the study patients might have early or atypical disease and/or other conditions that might interfere with the test, and the controls will likely have other conditions that might mimic the disease of interest. A common measure of accuracy for comparing tests at this phase is again the AUC, but other measures of accuracy, such as the partial area under the ROC curve, are also applicable, and sample size determination is similar.

We consider the null hypothesis, that the accuracies of two tests are equal, vs the alternative hypothesis, that the accuracies differ (two-sided test). (Later, we consider sample size determination for noninferiority studies.) A formula for computing sample size to test these hypotheses is given in Eq. 4:

$$n_D = \frac{\left( z_a/2 \sqrt{V_0(\hat{\theta}_1 - \hat{\theta}_2)} + z_\beta \sqrt{V_A(\hat{\theta}_1 - \hat{\theta}_2)} \right)^2}{(\theta_1 - \theta_2)^2} \quad (4)$$

where $z_a$ and $z_\beta$ are the same as in Eq. 2 (but here we have a two-sided hypothesis, so we use $\alpha/2$); $V_0$ and $V_A$ denote the variance functions under the null and alternative hypotheses, respectively; $\theta_1$ and $\theta_2$ denote the conjectured accuracies of diagnostic test 1 and test 2, respectively; and $V(\hat{\theta}_1 - \hat{\theta}_2) = V(\hat{\theta}_1) + V(\hat{\theta}_2) - 2C(\hat{\theta}_1, \hat{\theta}_2)$. For paired designs (i.e., the same study participants undergo both diagnostic tests), the results from the two tests will be correlated, i.e., $C(\hat{\theta}_1, \hat{\theta}_2)$, the covariance function, will be nonzero (usually taking on a positive value); $n_D$ in Eq. 4 is the number of patients with disease needed for the study. For unpaired designs (i.e., different study participants undergo the two tests), $C(\hat{\theta}_1, \hat{\theta}_2)$ is zero, and $n_D$ is the number of patients with disease needed for each diagnostic test. There is useful software for sample size determination for comparing two AUC (10, 11, 21) or two partial areas (11).

We now consider testing noninferiority of a new test to a standard test. The sample size calculation is similar to that in Eq. 4, but we need to take the value of $\Delta_{H}$ into account. The formula for sample size determination for testing noninferiority is given in Eq. 5 (28).

$$n_D = \frac{(z_a + z_\beta)^2V_A(\hat{\theta}_2 - \hat{\theta}_N)}{(\theta_2 - \theta_N - \Delta_{H})^2} \quad (5)$$

In phase III studies we usually assess and compare the accuracy of diagnostic tests for a well-defined and generalizable population. We want to report CIs for test accuracies, and we want these CIs to be narrow so that clinicians using the tests in practice have a good sense of the abilities of the tests and can interpret the results appropriately for their patients. The AUC is a global measure of the accuracy of a test, i.e., it is the average sensitivity over all possible values of specificity, or the average specificity over all possible values of sensitivity (1, 29). It is not very useful for a phase III study (2, 30). One alternative to the AUC is to find the optimal point on the ROC curve (based on the prevalence of disease among patients undergoing the test and the costs, e.g., patient morbidity or monetary, of false positives and false negatives) and report the corresponding sensitivity and FPR and their CIs (31–35). Another approach is to estimate the average sensitivity for the range of FPRs that is useful clinically (e.g., average sensitivity when the FPR is <0.10) or the average specificity for the range of sensitivities useful clinically (e.g., average specificity when sensitivity is >0.90), i.e., the partial area index (36, 37). Determination of the required sample size for phase III studies using these indices of accuracy is often more complex because we need to know the shape of the ROC curve, which can be estimated from previous phase II studies. For further discussion of sample size issues, see Zhou et al. (2) or Obuchowski (38).

**Discussion**

Clinical chemists have been using ROC curves to characterize diagnostic test accuracy for many years. An excellent primer on ROC analysis was published in Clinical Chemistry by Zweig and Campbell in 1993 (35) and remains a key reference in this field. There have also been several articles on the quality of reporting of diagnostic accuracy studies in Clinical Chemistry (5, 6), the latest (6) recommending ~40 items for inclusion in published papers on diagnostic accuracy.

We focused here on how ROC curves are currently being used by clinical chemists and what shortcomings exist in the ROC analyses being performed. Although there were many reports published on the accuracy of new and old diagnostic tests, we reviewed only the 58 articles using ROC curves. These 58 articles represented a spectrum from early, exploratory studies to large studies of mature tests. They included a mixture of retrospective and prospective studies, with sample sizes ranging from 30 to >0000 patients.

Some of the problems we saw with the use of ROC curves are common in other disciplines; for example, a lack of reporting of CIs for measures of accuracy is a common problem in diagnostic radiology as well. We were surprised by the large number of articles that failed to properly compare the accuracies of the two diagnostic tests. Identification of available software should resolve this problem. Other problems, in particular, the dichotomization of a continuous-scale gold standard outcome, are specific to the kinds of diagnostic tests often evaluated by clinical chemists; further methodologic research is needed in these areas.
References