Alternative Approaches to Receiver Operating Characteristic Analyses

In this issue of Radiology, Berbaum et al (1) describe a new statistical technique for analyzing data from a receiver operating characteristic (ROC) study. This editorial is a tutorial on this new approach and an attempt to show where the approach fits in with conventional analysis and when it may be an oversimplification.

Traditional Approaches

The reported diagnostic performance under a certain experimental condition is usually obtained by averaging performance over a sample of observers. Usually all observers read the same sample of cases, generally only once in each condition. The uncertainty of the average is quantified with the use of a standard error (SE). Ideally, this SE should be a composite of all of the sources of variation introduced by sampling individual cases, observers, and reading occasions. The form of the SE and the meaning of each of the variance components are fully explained in the text by Snetsinger and Pickert (2). As expected, the average is more trustworthy (its SE is smaller) if it is calculated with more cases, more observers, and more occasions.

Inferences on the difference in performance in two conditional sets based on the SE of the observed difference in averages. If the conditions are tested on different sets of cases by different readers, the SE of the difference is obtained by combining the separate SEs in the usual way. If case s and/or readers are matched across the two conditions, the relevant components of the SE are reduced accordingly. The larger the case correlation across the two conditions (the larger the degree to which a set of cases that are above [below] average difficulty in one condition will be likewise in the other) and the larger the degree to which a set of observers who are above (below) average accuracy in one condition will likewise be in the other condition, the smaller the SE of the estimated difference in performance.

The computation of the composite SE is described in detail in reference 2. If case and/or reading variances are negligible, the "fixed" SE should be calculated. In such situations, estimation of the case variance and covariance can be difficult when the number of cases is small, because it involves splitting the dataset into subsets of cases and applying the Dorfman and All method outlined in a previous study. Similar, if reader and/or reading variances are negligible, the SEs can be based on those produced by the method of Dorfman and All (3).

When all three components are non-negligible, the "full"-SE approach should be calculated. In such situations, estimation of the case variance and covariance can be difficult when the number of cases is small, because it involves splitting the dataset into subsets of cases and applying the Dorfman and All method outlined in a previous study. Similarly, if reader and/or reading variances are negligible, the SEs can be based on those produced by the method of Dorfman and All (3).

In the jackknife method of Dorfman and Berbaum, one uses the same case sample to estimate the ROC curve, the area under the ROC curve, and the standard error of the area. This is done by systematically removing one case at a time and calculating the ROC curve for the remaining cases.

How to Calculate the SE?

The main implication of Dorfman and Berbaum's approach is that we no longer use jackknife per se, rather...
er, the main issue is that we base the analysis strictly on between-reader variation. In their users' guide (7), they advise that if two independent pools of observers are to be compared, pseudo
estimates for the different readers can be used to construct a test for inde-
pendent samples with the degrees of freedom based on the number of inde-
pendent readers. For the study with matched readers, they recommend that if two pools of data are obtained from a single group of observers presented with the same stimuli under two ex-
perimental conditions, a t test for paired observations can be performed on the paired pseudovariables of the two groups. Thus, in the matched-readers example, they test the average of the seven paired differences.

Ar=\(|A_{1}|-|A_{2}|-\text{level} 0.05\) using a
study-sample Student t test with (7 - 1 = 6 degrees of freedom) by computing the usual t critical ratio-0.05/6, where the SE is calculated from the individual d's in the same way as for any paired t

These suggestions differ from Swets and Picktel's approach, in that Dorf-
man and Berbaum effectively ignore case and randomizing variance when
calculating the SE and deal only with be-
tween-reader variation. (In principle, the other sources of variation could be included by extending the jockey-tiling to randomizing variance, as has been suggested by Hanley [9].)

I argue that Hanley's or any other common approach is not to be accepted
and will not cause serious under-estimation of the SE.

The Dorfman-Berbaum approach of basing interferences on the number of
and variation between readers conveniently emphasizes that the believability of an observed difference depends on the number of cases and the size of the difference is observed, as well as on the number of cases used in the study. In other words, the unit of analysis is as
much the observer as it is the object that is being observed. Indeed, this is implicit in the terminology used by
Dorfman and Berbaum, who use the term "subject" to refer to an observer.
However, the simpler Dorfman and Berbaum approach, and its smaller SEs, could still be used, if not as a test to
ignore the other components included in the Swets and Picktel for-
tilation.

Implications for Sample Sizes

It is important if the early writings on SEs and with statistical inference in general—beginning with Dorfman and Alt and continuing with Hanley and McNeil and with Metz et al.—were fo-
cused solely on the numbers of cases and not at all on the numbers of raters.
SEs based only on the number of cases, do have a place, albeit in special excep-
tional-situations. They are appropriate for comparison of the performance of a specific (named) reader in one condition with the performance of the same (or another named) reader in a second (or perhaps in the same) condition. However, no matter how large the number of cases, one cannot usually make infer-
ces based on results of one reader to a whole class of readers (except, of

cause, when there is absolutely no be-
tween-reader variation). "Observer-
less" diagnostic systems, which invari-
antly give the same test results on a set of cases, are one such exception. Exam-

les of such systems might be (a) auto-

mated computer procedures that use objective features of images to detect abnormalities and (b) clinical predic-
tion techniques, such as discriminant

analysis, regression, and other patient-
sorting algorithms. In use unique-

cal clinical inductives to generate diag-

noses or prognoses. In such situations, where case and subject variance are both zero, the statistical conclusions are determined solely by means of the case-variance, the degree of case-

matching, and the size of the "case-
simple. The "case-based" SEs from the esti-
mation procedure of Dorfman and

Hanley (3), Hanley and McNeil (4), and Metz et

al (6) and the formulae and programs for calculating sample size developed

by Hanley and McNeil and by Metz et al are directly applicable in such situa-
tions.

The Correct Unit of Analysis: An Example

Because it is often difficult to know
which is the correct unit of analysis (whether source of variance or which is "n" to use in the numerator and de-

nominator of the SE [1]), it is worth considering an illustration. Imagine that we wish to test whether a particular method of academic training leads to better performance than another. Performance of training is to be as-

sessed (estimated) on a sample of exam-
inations. If we compute the results of the methods using one

traineepor method, the only effect of increasing the number of questions

used is to make us more convinced about which of the two methods performs

better in these two tests. Indeed, unless we were sure that both these

performers would stand the same relations to each other on another day,

we might need to augment the number of exami-
nations or sessions to take into account within "candidate" variability. Either way, no matter how much we increase these two n's, we still cannot infer how

well trainers in general will perform in each of the two conditions. This can only be achieved by increasing the n of

trainees assessed. Of course, there should be enough questions to avoid the situation in which, somehow by

chance, the limited number of ques-
tions used favored one condition over the other; one hopes that any observed difference between two observers (ei-

ther in the same condition or in differ-

ent conditions) is not due to the ques-
tions selected.

Since case and readers in an observ-

er performance study are analogous to questions and trainers in a test sample,

use of a sufficient number of the same (or matched) cases in both condi-

tions should allow one to control that the contribution of case variance to the composite SE is minimal and that with that of between-reader variance. Then, in planning the statistical power of an observer performance study, one can be guided by the same calculations used for simple comparisons of means taken over observers. One can simply consult nomograms or tables for two-

sample t tests showing the number of subjects (observers) required to have a specified probability of detecting a differ-

ence of 0.05 when the projected be-
tween-reader standard deviation is 0.11. For matched readers, one can consult the table for the one-sample t test, for the difference of the pair differences. The tables are labeled in terms of the "signal-to-

noise" ratio, n/\.

Non-parametric Tests Being

Conceived of Consistent Differences

If one is uncomfortable performing parametric tests on so few numbers, the non-parametric analogs of the t tests (rank tests) are an effective alternative. Instead, they illustrate the minimum number of readers needed to reach a "significant" difference: It, in a study with two (or more) • readers and three (or more) readers in the other condition, the performance for the two

readers in one condition all rank high-

er than those of the three in the other condition, and if the data were hypo-

thesized in the way such a pattern is associ-

ated with p value (one-sided) of 0.12 or 0.85 with the rank sum test. If a study with five matched readers produces five intercorrelation t's that are consis-

tently in the hypothesized direction, this pattern is associated with a p value (one-sided) of 0.12 or 0.85 with the sign test in the study by Berbaum et al [1]. The distribution of the

permuted and unpermuted detection accuracy was presented to all observers.

Many investigators have re-

ported such patterns without formal statistical tests, knowing intuitively that the differences must be "real."
fact, if they are achieved despite the
statistical noise caused by low numbers
of cases and no remedies, one could
argue that the patterns are all the more
remarkable. Although they leave the
choice of the number of readers up to
an investigator's scientific judgment,
Swets and Pickett (2) suggest that, even
apart from issues of power require-
ments, a reading too should as a rule
have "at least several" readers; in an-
other chapter, they "emphasize again
that one should strive to work with a
reasonably large sample of readers,"
since small samples can easily give rise
to "sampling oddities."

More Readers, Fewer Cases?
The number of cases a reader is ex-
pected to read limits the number of
readers willing to participate in an ob-
server performance study. Fortunately,
the increased emphasis on and under-
standing of the value of a larger selec-
tion of readers will make it easier to re-
duce somewhat the case numbers and
thereby allow more readers to partici-
ate (the number of cases in this and
other studies by Berbaum and Dorfman
was small enough that they could list
the individual characteristics of the
cases [12]). The use of pooling and
pseudocases can overcome the practi-
cal difficulty of fitting reader-specific
ROC curves from such a small number
of cases, although the pooling can pro-
duce larger estimates of between-read-
er variance than is seen in the data
from individual readers. However,
even if these pooling artifacts could be
avoided, the number of cases still can
not be allowed to be so small that it is
impossible to generalize from them.
Even if one were to employ jackknifing
of readers, a study with two cases and
140 readers cannot equal one with 40
cases and seven readers.

References
1. Berbaum KS, R-Khouri CY, Franken KA,
   Jr, Kettler M, Montgomery NJ, Hannan W.
   Impact of clinical history on radiographic
2. Swets JA, Pickett RM. Evaluation of diag-
nostic systems: methods from signal detec-
tion theory. New York: Academic,
   1982.
3. Dorfman DD, Ah P. Maximum likeli-
hood estimation of γ-powers of signal
detection theory and determination of
   confidence interval using monte carlo data.
4. Hanley JA, McNeil BJ. A method of com-
paring the areas under receiver operating
   characteristic curves derived from the
5. McNeil BJ, Hanley JA. Statistical ap-
   proaches to the analysis of receiver operat-
   ing characteristic (ROC) curves. Med
6. Mets CE, Wang P-L, Keimann HB. A
   new approach for testing the significance of
differences between ROC curves from
   correlated data. In: Denison S, ed. Infor-
mation processing in medical imaging.
7. Dorfman DD, Berbaum KS, BSRCORE;
   pooled rating method data—a computer
   program for analyzing pooled ROC
curves. Behav Res Methods Instrum
8. Fino J, Davis M. Jackknifing functions
   of multichannel frequency, with an appli-
cation to a measure of concordance. Am J
   Epidemiol 1982; 115:443-450.
9. Hanley JA. Observer performance statis-
tics. Presented at Chest Imaging Confer-
ence '87, University of Wisconsin, Madi-
10. Whiting-O'Keefe KE. Choosing the cor-
    rect unit of analysis in medical care ex-
11. Bayer WH. Handbook of tables for prob-
    ability and statistics. 2d ed. Cleveland:
    CRC, 1968.
12. Berbaum KS, Franken KA, Dorfman DD, et
    al. Tentative diagnoses facilitate the de-
    tection of диверс lesions in chest radio-