James A. Hanley, PhD

Alternative Approaches to Receiver Operating Characteristic Analyses[1]

IN this issue of *Radiology*, Berbaum et al (1) describe what to many is a new statistical technique for analyzing data from a receiver operating characteristic (ROC) study. This editorial is a tutorial on this new approach and is an attempt to show where the approach fits in with conventional analyses 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 Swets and Pickett (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 conditions are 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 cases 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 rereading variances are negligible, the inferences are equivalent to analysis of the performance statistics of each observer with familiar statistical techniques. Similarly, if reader and rereading variances are negligible, the SEs can be based on those produced by the method of Dorfman and Alf (3). When all three components are nonnegligible, the "full-blown" 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 Alf estimation method to each subset; Hanley and McNeil (4) used a largely nonparametric approach to estimating the case covariances when one uses the "area" as the index of performance. They also illustrated the use of jackknifing of cases, coupled with the Dorfman and Alf procedure, to estimate directly the SE of a performance difference measured on the same cases (5). Metz et al (6) provide a bivariate extension of the binormal model, which allows one to estimate case variance and covariance parametrically.

Dorfman and Berbaum's Approach

Recently, Dorfman and Berbaum described a method and an accompanying computer program to compute jackknife estimates (and their SEs) of indexes extracted from a single ROC curve fitted to rating-method data that have been pooled over a group of observers (7). Such pooling might be the last resort if there are not sufficient cases to fit separate ROC curves for each observer. As will be discussed in a following section, the approach of Dorfman and Berbaum is more noteworthy for the form of the SE they use than for the use of jackknifing per se.

Until now, jackknifing has been used only to assess sampling variation due to cases (5). However, Dorfman and Berbaum use it to assess the variation due to readers. To understand their jackknife approach, it is best if we restrict attention to data gathered under a single experimental condition and to first review the four steps Dorfman and Berbaum might take (and that Berbaum et al did take in the second row of their Table 2 [1]) when there *are* enough data to estimate *directly* a separate curve for each observer. These steps are (*a*) calculate a separate value of the area for each of the seven observers, (*b*) calculate and report the average value of this index, (*c*) assume that case and rereading variances are negligible and base the SE of the reported average simply on the number (seven) of readers making up this average and the observed variation in the index between readers, and (*d*) base inferences on the *t* distribution.

In the jackknife method of Dorfman and Berbaum, one uses these same four steps, except that instead of the seven directly calculated areas, one uses indirectly calculated "pseudovalues" of the areas. These pseudovalues (which we will denote by asterisks) can be regarded as the contributions of the individual observers to the single ROC curve estimated from the pooled data. (See Fleiss and Davis [8] for a nontechnical exposition on jackknifing.) Thus, to obtain the pseudovalues, one must first calculate the area, which they denote Area_{all}, from the entire pooled rating-data. Then, to obtain the pseudovalue Area^*_i for observer *i*, one deletes the rating-data produced by the *i*th observer from the overall pool of data and uses the area, which they denote Area_{-i} and which is calculated from this reduced pool, to calculate $\text{Area}^*_i = (7 \times \text{Area}_{\text{all}}) - (6 \times \text{Area}_{-i})$. The analysis of these pseudovalues then proceeds in the same way as outlined in steps *a–d* above.

How to Calculate the SE?

The main implication of Dorfman and Berbaum's approach stems not from the use of jackknifing per se; rath-

Index terms: Editorials • Receiver operating characteristic curve (ROC) • Statistical analysis

Radiology 1988; 168:568–570

[1] From the Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave West, Montreal, Quebec, H3A 1A2, Canada. Received May 17, 1988; accepted May 18. Supported by an operating grant from the Natural Sciences and Engineering Council of Canada. **Address reprint requests to** the author.

See also the article by Berbaum et al (pp 507–511) in this issue.

er, the main issue is that they base the analysis strictly on between-reader variances. 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 t test for independent samples with the degrees of freedom based on the number of independent 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 experimental conditions, a t test for paired observations can be performed on the paired pseudovalues of the two groups. Thus, in the matched-readers example here, they test the average of the seven paired differences d_i = $\text{Area}_i(1) - \text{Area}_i(2)$ against zero using a one-sample Student t test (with $7 - 1 = 6$ degrees of freedom) by computing the usual critical ratio $\bar{d}/\text{SE}[\bar{d}]$, where the SE is calculated from the individual d_i's in the same way as for any paired t test.

These suggestions differ from Swets and Pickett's approach, in that Dorfman and Berbaum effectively ignore case and rereading variance when calculating the SE and deal only with between-reader variation. (In principle, the other sources of variation could be included by extending the jackknifing to both cases and readers, as has been suggested by Hanley [9].)

If one estimates the SE of the difference in an index between two conditions that use the same set of cases, and if the case-covariance is high, then the omitted variance component due to case selection will be small and will not cause serious underestimation of the SE.

The Dorfman-Berbaum approach of basing inferences on the number of and variation between readers correctly emphasizes that the believability of an observed difference depends on the number of observers in whom 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 objects that are 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 the smaller SEs it produces, should not be taken as a license to ignore the other components included in the Swets and Pickett formulation.

Implications for Sample Sizes

It is unfortunate if the early writings on SEs and with statistical inference in general—beginning with Dorfman and Alf and continuing with Hanley and McNeil and with Metz et al—were focused solely on the numbers of cases and not at all on the number of readers. SEs based only on the number of cases do have a place, albeit in special exceptional 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 inferences based on results of one reader to a whole class of readers (except, of course, where there is absolutely no between-reader variation). "Operatorless" diagnostic systems, which invariably give the same test results on a set of cases, are one such exception. Examples of such systems might be (*a*) automated computer procedures that use objective features of images to detect abnormalities and (*b*) clinical prediction techniques, such as discriminant analysis, regression, and other patient-sorting algorithms, that use unequivocal clinical indicants to generate diagnoses or prognoses. In such situations, where case and rereading 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 sample. The "case-based" SEs from the estimation procedures of Dorfman and Alf (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 situations.

The Correct Unit of Analysis: An Example

Because it is often difficult to know which is the correct unit of analysis (ie, which source of variation and which "n" to use in the numerator and denominator of the standard error [10]), 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 trainees is to be assessed (estimated) on a sample of examination questions. If we compare the results of the methods using one trainee per method, the only effect of increasing the number of questions used is to make us more convinced about which of the methods performs better *in these two trainees*. Indeed, unless we were sure that these two persons would stand in the same relation to each other on another day, we might need to augment the number of examinations 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 trainees *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 questions used favored one condition over the other; one hopes that any observed difference between two observers (either in the same condition or in different conditions) is not due to the questions selected.

Since cases and readers in an observer performance study are analogous to questions and trainees in this example, the use of a sufficient number of the same (or matched) cases in both conditions should allow one to consider that the contribution of case variance to the composite SE is minimal compared 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 difference of δ when the projected between-reader standard deviation is σ (11). For matched readers, one consults the table for the one-sample t test, where σ is the projected standard deviation of the pair differences. The tables are tabulated in terms of the "signal-to-noise" ratio, δ/σ.

Nonparametric Tests: Being Convinced by Consistent Differences

If one is uncomfortable performing parametric tests on such few numbers, the nonparametric analogs of the t tests (rank tests) are an attractive alternative. Indeed, they illustrate the minimum number of readers needed to reach a "significant" difference: If, in a study that uses three readers in one condition and three (unmatched) readers in the other, the performances for the three readers in one condition all rank higher than those of the three in the other condition, and if this was the hypothesized direction, such a pattern is associated with a P value (one-sided) of 1/20 or .05 with the rank sum test. If a study with five matched readers produces five intercondition d's that are consistently in the hypothesized direction, this pattern is associated with a P value (one-sided) of 1/32 or .03 with the sign test (in the study of Berbaum et al [1], the difference between location prompted and unprompted detection accuracy was present in all seven observers). Many investigators have reported such patterns without formal statistical tests, knowing instinctively that the differences must be "real." In

fact, if they are achieved despite the statistical noise caused by low numbers of cases and no rereadings, 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 requirements, a reading test should as a rule have "at least several" readers; in another 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 expected to read limits the number of readers willing to participate in an observer performance study. Fortunately, the increased emphasis on and understanding of the value of a larger selection of readers will make it easier to reduce somewhat the case numbers and thereby allow more readers to participate (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 pseudovalues can overcome the practical difficulty of fitting reader-specific ROC curves from such a small number of cases, although the pooling can produce larger estimates of between-reader variance than is seen in the data from individual readers. However, even if these pooling artifacts could be avoided, the number of cases still cannot 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, El-Khoury GY, Franken EA Jr, Kathol M, Montgomery WJ, Hesson W. Impact of clinical history on radiographic fracture detection. Radiology 1988; 168:507–511.
2. Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic, 1982.
3. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. J Math Psychol 1969; 6:487–496.
4. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148:839–843.
5. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 1984; 4:137–150.
6. Metz CE, Wang P-L, Kronman HB. A new approach for testing the significance of differences between ROC curves from correlated data. In: Deconink F, ed. Information processing in medical imaging. The Hague: Nijhoff, 1984; 432–445.
7. Dorfman DD, Berbaum KS. RSCORE-J: pooled rating-method data—a computer program for analyzing pooled ROC curves. Behav Res Methods Instrum Comput 1986; 18:452–462.
8. Fleiss J, Davis M. Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. Am J Epidemiol 1982; 115:841–845.
9. Hanley JA. Observer performance statistics. Presented at Chest Imaging Conference '87, University of Wisconsin, Madison, August 31–September 2, 1987.
10. Whiting-O'Keefe QE. Choosing the correct unit of analysis in medical care experiments. Med Care 1984; 22:1101–1114.
11. Beyer WH. Handbook of tables for probability and statistics. 2d ed. Cleveland: CRC, 1968.
12. Berbaum KS, Franken EA, Dorfman DD, et al. Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. Invest Radiol 1986; 21:532–539.