[Reprinted from RADIOLOGY, Vol. 149, No. 1, Pages 75–77, October, 1983.]

Barbara J. McNeil, M.D., Ph.D.
James A. Hanley, Ph.D.
H. Harris Funkenstein, M.D.
James Wallman, M.D.

Paired Receiver Operating Characteristic Curves and the Effect of History on Radiographic Interpretation

CT of the Head as a Case Study[1]

The use of a statistical technique for paired comparisons using receiver operating characteristic (ROC) curves is illustrated by studying the extent to which clinical history altered the interpretation of computed tomographic (CT) examinations of the head. Eighty-nine CT examinations of the head were presented in random order to four readers, first with minimum history (age and sex) and then several weeks later with complete neurological history as of the time the CT examination had been obtained. Using a paired ROC analysis, a small but significant ($p < .05$) improvement was detected for the interpretations in the presence of complete history; for readings without history the average area was 94.4% and for readings with history it was 97.7%.

Index terms: Diagnostic radiology, observer performance • Receiver operating characteristic curve (ROC)

Radiology 149: 75–77, October 1983

THE EXTENT to which clinical information is necessary for the accurate interpretation of radiographs is not well documented. *A priori* it could be argued that there might be a positive effect on accuracy, no change at all, or a negative effect caused by increased false-positive diagnoses. In this study we address this problem from two perspectives: methodological development and practical application. Methodologically, we have used the method of Swets and Pickett (1) as well as a recent refinement (2) to analyze paired data and illustrate here some of the considerations important in the use of that technique. Practically, we have studied the interpretation of computed tomographic (CT) examinations of the head as performed with and without extensive clinical history.

METHODS

Patient Sample

The patient population for this study was drawn from a previous data base consisting of over 2,000 patients who had CT studies of the head performed at the Dana Farber Cancer Institute, Boston (3). We first searched this data base for patients who had objective proof (biopsy, surgery, angiography, pneumoencephalography) of the final disease state or, alternatively, a diagnosis made on the basis of compelling clinical follow-up data at 18 months. In order to ensure that final clinical diagnoses were independent of presenting data, all eligible charts were reviewed by a neurologist to ensure that the CT examinations had not determined the final diagnoses. Then, a sample of 89 patients, 35 (39%) subsequently shown to have intracranial disease and 54 (61%) shown to be free of intracranial disease, was chosen. This proportion of patients with intracranial disease was similar to that in the original population of 2,000 patients.

TABLE I summarizes the disease processes and their means of diagnosis.

Readers

Four staff radiologists of the Division of Neuroradiology of the Department of Radiology at the Brigham and Women's Hospital were chosen for this study. Two had also had additional residency experience in internal medicine and two in neurology.

CT Technique

As indicated in a previous publication (3) all CT studies were performed on an EMI scanner (Model 1005) using a 160 × 160 matrix. Four to five pairs of sections (two sections per scan) were obtained on each patient at 8-mm intervals starting at the base of the skull.

[1] From the Departments of Radiology (B.J.M., J.W.) and Medicine (H.H.F.), Harvard Medical School and the Brigham and Women's Hospital, Boston, MA, USA, and the Department of Epidemiology and Health (J.A.H.), McGill University, Montreal, Canada.
Received Jan. 25, 1983; accepted and revision requested April 15, 1983; revision received May 26, 1983.

This work was supported by grants from the Hartford Foundation and the National Center for Health Care Technology.

See also the paper by Alderson *et al.* (pp. 225–230) in this issue.

TABLE I: Final Diagnoses and Means of Diagnosis

Diagnosis	Total	Biopsy, Surgery	Angiography	Followup
Normal	54	0	1	53
Abnormal	35	5	10	20
Tumor	10	3	2	5
Vascular disease	14	2	8	4
Miscellaneous	11	0	0	11

Contrast material (100 ml Conray 600 [meglumine iothalamate]) was generally used in patients suspected of having tumors, vascular malformations, or inflammation. It was generally not used in patients suspected of having atrophy, hydrocephalus, traumatic or hypertensive intracerebral hematoma, or infarction.

Experimental Approach

All identifying data were first removed from the studies, and the studies were placed in random order. In the first part of the study the only information given to the reader was the age and sex of the patient. The reader was asked to spend no more than four minutes on each polaroid study and to record responses in one of five categories: definitely abnormal for intracranial disease, probably abnormal, possibly abnormal, probably normal, or definitely normal. The five categories allowed us to create an ROC curve. Two weeks later the studies were placed in a different random order, and this time the reader was given all clinical information that was available at the time the study was ordered. For example, in this phase of the study the reader might have been told the following: "The patient, a 59-year-old woman, with no history of hypertension, previous stroke, or known tumor, has apparently worsening papilledema, focal symptoms, and altered vision of several months' duration. Skull radiographs obtained prior to CT were normal." With the rating data provided by this second reading we were able to create a second ROC curve for each reader.

Statistical Methods

The data were analyzed according to the methods described in Swets and Pickett (1) and Hanley and McNeil (2). We first used the maximum likelihood technique described by Dorfman and Alf (4) to obtain areas under the ROC curve for each of the four readers in both of the two settings, and calculated the average difference (improvement) in the areas obtained when reading with history over those obtained without history. To conduct a test of significance on this improvement, we divided this difference by its standard error and referred the ratio to the table of normal (z) deviates. As described in Swets and Pickett (1), the overall standard error (SE) for this difference is a merging of three separate SEs corresponding to the patient or case (c) variability, between-reader (br) variability, and within-reader (wr) variability, along with the components of covariation between the two settings; we used the following equation for this purpose (this is Equation 5, Chapter 4 of Reference 1):

$$SE_{(diff)} = 2^{1/2}\left[S^2_{c+wr}\,(1 - r_{c-wr}) + \frac{S^2_{br+wr}}{l}\,(1 - r_{br-wr}) - S^2_{wr}\right]^{1/2} \quad (1)$$

r_{br-wr} = the observable correlation between the areas obtained when a set of readers reads the same cases in the two settings,

r_{c-wr} = the observable correlation between the areas obtained when a single reader reads a set of case samples in the two settings,

l = the number of readers,

S^2_{c+wr} = $S^2_c + S^2_{wr}$, the observable variance in area that would be found by having one reader read once each of a set of different case samples,

S^2_{b+wr} = $S^2_{br} + S^2_{wr}$, the observable variance in area that would be found by having one case sample read once by each of a set of different readers,

S^2_{wr} = the observable variance in area that would be found by having one reader read one case sample on two or more independent occasions.

The quantity S^2_{c+wr} was estimated by averaging the variances obtained from the Dorfman and Alf computer program for each reader and setting. The quantity r_{c-wr} was estimated from the correlations of the ratings given to individual patients in the two settings, according to the method of Hanley and McNeil (2). The quantities S^2_{br+wr} and r_{br-wr} were estimated according to the method of Swets and Pickett (1).

The only term remaining in order to use Equation 1 is S_{wr}. Although we have no direct measure of it, because we did not employ rereading in both experimental settings (with and without history), we can estimate it indirectly by using the approach of Swets and Pickett (5), which shows that the quantities r_c (the true correlation between areas that is induced by case-matching in the two settings), r_{c-wr} (the observed correlation between areas due to case-matching, which is attenuated by within-reader variation), S^2_{wr} (variability in area due to reader inconsistency), and S^2_c (true variation in area in different case samples) are related by the following formula (from Reference 5, Equation 8.4):

$$S^2_{wr}/S^2_c = (r_c/r_{c-wr}) - 1 \quad (2)$$

If we let $S^2_c = S^2_{c+wr} - S^2_{wr}$, then the following holds:

$$S^2_{wr} = S^2_{c+wr}\,[1 - (r_{c-wr}/r_c)] \quad (3)$$

For this type of experiment, it is reasonable to postulate a high true correlation (r_c) between areas induced by having the same cases. If we assume a conservative value of 0.75, we can calculate S^2_{wr} from

$$S^2_{wr} = S^2_{c+wr}\,[1 - (r_{c-wr}/r_c)]$$

(S^2_{c+wr} and r_{c-wr} will already have been obtained as described above.)

RESULTS

TABLE II presents the four pairs of ROC areas (and their estimated variances) obtained by maximum likelihood fits of the eight sets of rating data to Gaussian-based (binormal) ROC curves. All areas are extremely high, the lowest being 0.931 or 93.1%. Presumably, this is due to the high "accuracy" of read CT scans in detecting the presence or absence of lesions, although we cannot exclude the possibility that the criteria by which we selected the 89 cases, together with our insistence on "separate proof," produced a sample of cases of less than average diagnostic difficulty.

The average difference in the areas as a result of including history was 3.3% (TABLE II), bringing the average area to nearly 98%. Moreover, although there was a clear ranking of readers without history, no such ranking existed with history. Thus, the correlation r_{br-wr} is essentially zero.

In order to determine the standard error of the difference between areas, we used Equation 1 and the following estimates of the relevant variances and correlations:

S^2_{c+wr} = average of eight variances (columns 4 and 5 in TABLE II)
= 0.000572

r_{c-wr} = average of four between-area correlations, each correlation obtained from TABLE II of

Hanley and McNeil (2) *via* Kendall tau correlations of pairs of ratings
$= 0.37$

S^2_{br+wr} = average of two within-setting estimates of the between-reader variance in ROC areas, *i.e.*, average variance within columns 1 and 2 of TABLE II
$= 0.000154$

$S^2_{wr}/S^2_c = (r_c/r_{c-wr}) - 1$
$= (0.75/0.37) - 1 = 1.03$

i.e.
$S^2_{wr} = .00029$

Thus

$$SE_{DIFF} = 2^{1/2}[0.000572(0.63) + 0.000154(1.00)/4 - 0.000290]^{1/2} = 0.015$$

and
$z = 0.033/0.015 = 2.2 \quad (p < .05)$

History thus significantly improves the interpretation of CT studies of the head.

DISCUSSION

This study was a straightforward one serving primarily a methodological purpose but also a practical one. Methodologically, two issues need to be highlighted. The first is an obvious one relating to case selection. Because it is essential to avoid circular reasoning, care must be taken to include only those patients whose final diagnoses are made on the basis of pathologic specimens or follow-up data apart from the CT study itself. In order to ensure this we had to review many more cases than were ultimately chosen for the final rereading exercise. There was difficulty with the conclusions in one of the very early works in this field (6) because of this problem.

The second issue relates to analysis and emphasizes the need, especially in small samples such as this one, to consider all aspects of the experiment that might impinge on analysis. In this study the issue of using a paired rather than unpaired analysis was paramount. Thus, we had to consider in Equation 1 the correlation between areas present when a single reader reads the same set of cases in two settings, here with and without history (r_{c-wr}). By including this factor we were able to achieve a 37% reduction in case sampling variance of the differences in areas, brought about by the paired-cases design. We were able to obtain an estimate of r_{c-wr} by a previously discussed technique (2) that is useful for samples that are too small to subdivide into subsamples (1). Our set of 89 cases was too small to subdivide.

We were not able to estimate S^2_{wr} directly because we had no rereadings in a single setting. Ideally, rereadings would have been preferred to our indirect estimation technique but logistically they are frequently a problem in experiments of this type. One other comment must be made regarding S^2_{wr}. Our indirect estimation technique has led to a value of S^2_{wr} greater than S^2_{br+wr}. Obviously, this is impossible for a population (infinite sample) and in our case probably represents a sampling problem. This indirect technique could overestimate r_c. This is impossible to verify in this case because of the zero value of r_{br-wr}. Were we to have a measurable correlation (as we would expect if the areas were not so high) we could use the following equation

$$S^2_{wr} = S^2_{br+wr}[1 - r_{br-wr}/r_{br})]$$

along with Equation 3 in the following way. By measuring S^2_{c+wr}, S^2_{br+wr}, r_{c-wr}, r_{br-wr} and by assuming the between-reader correlation r_{br} as high (near 1), we would have two equations in two unknowns (S^2_{wr}, r_c) and would thereby be able to check our estimate of r_c.

Finally, the particular CT example chosen here was a good one for our methodologic purposes, since it involved a clinical situation in which both true- and false-positive ratios could be calculated and hence ROC curves generated. This contrasts with other more complex experimental situations where it has not been possible to calculate false-positive ratios and hence not possible to obtain a true and unbiased measure of the impact of history on interpretation (6, 7). This example was also useful because we were able to use the same sets of observers for both parts of the experiment (*i.e.*, history *vs.* no history), leading to greater control over different diagnostic criteria among different radiologists (8).

The practical results of this study indicate a small but significant increase in performance in the presence of as complete a history as might be available on a carefully completed requisition. The increase in performance translates into 3.3% fewer errors for a population in which 40% of patients with abnormal CT studies of the head are diseased. A larger percentage increase might have been expected if the interpretations of the CT head studies had not been so accurate without the benefit of a complete history. This conclusion should enhance the reputation of the diagnostic radiologist as not only an essential person in the choice of procedures to be performed but also as an integrator of clinical and diagnostic information (9).

TABLE II: Areas Under ROC Curves Fitted to Rating Data*

Reader	Area Under ROC Curve: Without History	Area Under ROC Curve: With History	Area Under ROC Curve: Difference	Estimated Variances of ROC Curves†: Without History	Estimated Variances of ROC Curves†: With History
1	0.931	0.988	0.057	0.001069	0.000074
2	0.947	0.966	0.019	0.000870	0.000437
3	0.934	0.973	0.039	0.001289	0.000234
4	0.963	0.982	0.019	0.000471	0.000128
Average	0.944	0.977	0.033		

* Data from four radiologists who read 89 CT head scans with and without patient history.
† Estimates produced by Dorfman and Alf computer program; these correspond to the quantity S^2_{c-wr} in Equation 1.

Acknowledgments: We are grateful to three of our other colleagues for their help in these interpretations—Steven Hammerschlag, Mohammed Naheedy, and Gerald O'Reilly—and to JoAnne Polak for research assistance. We are also grateful to Dr. Charles Metz for his comments on this manuscript.

Department of Radiology
Harvard Medical School
25 Shattuck Street
Boston, MA 02115

References

1. Swets JA, Pickett RM. Evaluation of diagnostic systems, methods from signal detection theory. New York: Academic Press, 1982, Chapters 3 and 4.
2. 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.
3. McNeil BJ, Hanley J, Funkenstein HH, Rumbaugh C. Utilization of computed tomography of the head in a tertiary care hospital. Radiology 1981; 139:113–118.
4. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence interval-rating method data. J Math Psychol 1969; 6:487–496.
5. Swets JA, Pickett RM. Evaluation of diagnostic devices in clinical medicine. Cambridge: Bolt, Beraneck and Newman, Report #3819 prepared for the National Cancer Institute, 1979.
6. Schreiber MH. The clinical history as a factor in roentgenogram interpretation. JAMA 1963; 185:399–401.
7. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. AJR 1981; 137:1055–1058.
8. Eldevik OP, Dugstad G, Orrison WW, Haughton VM. The effect of clinical bias on the interpretation of myelography and spinal computed tomography. Radiology 1982; 145:85–89.
9. Heilman RS. What's wrong with radiology. N Engl J Med, 1982; 306:477–479.