jnm/ CONCISE COMMUNICATION

DECISION MAKING IN NUCLEAR MEDICINE: 1. BRAIN IMAGING

Raymond A. Berke*, Eugene L. Saenger, and Gustave K. Bahr

Nuclear Medicine Laboratory, Bureau of Radiological Health, Food and Drug Administration, Department of Health, Education and Welfare and University of Cincinnati College of Medicine, Cincinnati, Ohio

Brain images were read by a panel of nuclear medicine physicians, residents, and technicians to determine whether the image was normal or abnormal. Receiver Operating Characteristic curves were developed indicating that technologists scored somewhat better than physicians in this initial trial. This method is capable of wide use in improving the interpretations of a physician. Technologists may be able to aid by preliminary screening of cases for suspected abnormalities.

Efficacy as defined by Webster is the power to produce effects or intended results. There are many reasons why a clinician orders a particular test for his patient or orders one test in preference to another, but ultimately he wishes to rule in or rule out the presence of disease with the greatest possible accuracy and least amount of morbidity and expense. The object of this paper is to develop simple methods in nuclear medicine to evaluate our procedures with regard to efficacy. Our initial attempt has used brain imaging.

In considering efficacy of brain scanning as well as any other nuclear medicine or radiological procedure, not only must one take into account the accuracy of that procedure, but also its morbidity, ease of performance, and cost, as well as the relative merits of alternative examinations that are available. Accuracy includes both specificity and sensitivity, the first being the ability of a test to give a negative finding when the patient is free of the disease under study and the second being the ability to find the disease in patients who truly have it (*1*). An obvious alternative to brain scanning is cerebral angiography (*2*). The simplicity of performing a brain scan compared with the more elaborate arteriogram has an obvious appeal. No morbidity has been reported following the brain scan procedure. Scheinberg and Zunker reported a morbidity rate of 3.4% and mortality rate of 0.3% in 902 arteriographic procedures in 500 patients (*3*). In 1,000 patients, in whom 2,301 brachial and carotid arteriograms were done, Feild, et al (*4*) report a morbidity rate of 2.9% and a mortality rate of 0.21%.

The cost of the examination at our institution including the professional fee is $75 for the scan compared with $160 for a single carotid arteriogram. In addition, a scan can easily be performed on an outpatient basis, whereas arteriography invariably requires hospitalization. The total cost for the scan is therefore much less. Finally the accuracy of the examinations must be compared. Previous data have indicated that both examinations are sensitive; that is, they are not likely to miss the presence of disease (*5,6*). The brain scan is somewhat less specific than the arteriogram; that is, it is difficult to make the diagnosis of a specific disease from an abnormal brain scan. Since brain scanning appears to be sensitive, a negative examination would have a high probability of excluding cerebral pathology. This is, therefore, a good initial test in the evaluation of a patient with suspected neurological disease and because of the other factors such as cost, morbidity, and ease, a more easily applicable initial test than cerebral angiography. Plain skull roentgenography, also used as a screening procedure, is less expensive, has no morbidity, and is easy to perform but is relatively less sensitive and specific (*7*).

Many questions arise when one attempts to analyze how often brain scanning answers the specific question for which it was ordered. Since different

Received Apr. 24, 1973; original accepted June 27, 1973.

For reprints contact: Eugene Saenger, Nuclear Medicine Laboratory, Bureau of Radiological Health, FDA, Cincinnati General Hospital, Cincinnati, Ohio 45229.

* Present address: Nuclear Medicine Laboratory, University of California, Irvine, Calif.

physicians are interpreting a study, how do their attitudes and varying criteria for normality or abnormality affect the overall accuracy of the procedure? How much variation in interpretation of a particular examination is there? The matrix most often used (*8,9*) to answer these questions is as follows:

		Your diagnosis Postive	Negative	
Disease category	Pos	% true positive	% false negative	= 100% diseased patients tested
	Neg	% false positive	% true negative	= 100% negative patients tested

The following percents are used to calculate Receiver Operator Characteristics:

True-positive percentage =

$$\frac{\text{No. true positive read} \times 100}{\text{No. true positives read} + \text{No. false negatives read}}$$

False-positive percentage =

$$\frac{\text{No. false positives read} \times 100}{\text{No. true negatives read} + \text{No. false positives read}}$$

Overall accuracy (%) =

$$\frac{\text{No. correct interpretations}}{\text{Total No. of studies}} \times 100.$$

In analyzing a study by Yerushalmy (*10*) involving a large series of chest photofluorograms for the presence of tuberculosis, Lusted noted a reciprocal relationship between percentage of true-positive and percentage of false-positive diagnosis. The curve that is generated is called the Receiver Operating Characteristic (ROC) curve. Such curves would also be applicable in brain scanning. Each observer operates on a particular point on the curve.

Establishing such a curve using a group of physicians interpreting a large series of scans in which accurate followup information is available will show the degree of individual variability in interpretation. Similarly it can be applied to establish criteria for interpretation of brain scans and possibly to enable other physicians to evaluate their own scan reading ability.

METHOD

Sixty-seven proven cases were selected by retrospective analysis of approximately 300 patients who had received brain scans in our laboratory. The distribution of these cases is shown in Table 1. Here, the true diagnoses are compared with the "official" interpretation—that is, the report of the scan that appeared in the patient's record. The official interpretation reflects the accuracy of group interpretation since each scan is first seen and initially interpreted by a resident and then discussed during the daily readout session where all the residents and at least one staff physician are present. As criteria for normality, at least 1-year followup without progression of symptoms was necessary. The diagnoses of the abnormal cases were proven by angiography, surgery, or postmortem examinations. All of the scintigraphs were made on a Nuclear-Chicago Pho/Gamma II or Pho/Gamma III camera. The usual adult intravenous doses of 10–15 mCi of ^{99m}Tc-pertechnetate were given after a blocking dose of 250 mg potassium perchlorate. At least four views were taken: anterior, posterior, right, and left laterals. The cases used in this series antedated routine cerebral flow studies and delayed scans. Some cases were chosen because of poor scan quality. This particular choice is believed in retrospect to have biased interpretation of the data.

TABLE 1. ABNORMALITIES IN BRAIN SCAN SERIES WITH RESULTS OF INTERPRETATIONS AT OFFICIAL READOUT

(Readings which appeared in patients' records)

Diagnosis	No. of cases	Percent correct interpretation	Percent correct
Normal	21	13	62
Abnormal	46	41	89
Primary brain tumor	7	7	100
Metastatic brain tumor	13	11	85
A-V malformation	4	4	100
Cerebrovascular accident	3	3	100
Brain abscess	4	4	100
Subdural hematoma	8	6	75
Cerebral contusion	2	2	100
Scalp or calvarial lesion	5	4	80
Total	67	54	80.6

The cases were coded and presented so that interpretation was either positive or negative. Five nuclear medicine staff physicians, three radiology residents assigned to nuclear medicine, and four technologists participated in the study. Each was given the set of scintigraphs three times with an interval of about 1 month separating each set of interpretations. The first and second set of readings were made without benefit of clinical history. A very brief clinical note was provided for the third set of interpretations. The radiology residents were first given the set during the first week of their 3-month rotation in nuclear

TABLE 2. COMPOSITE DATA

	True positive percent	False positive percent	Overall accuracy percent
Initial official interpretation:			
Not corrected	89.1	38.1	80.6
Corrected*	100	38.1	87.1
Staff:			
With history	67.8	9.5	74.9
Without history	63.7	12.1	71.3
Combined	65.2	11.2	72.6
Residents:			
With history	60.9	9.5	70.1
Without history	52.5	12.6	63.4
Combined	55.3	11.6	65.7
Technologists:			
With history	64.1	14.3	70.8
Without history	65.2	7.9	73.6
Combined	64.9	9.5	72.9

* Five cases deleted because abnormalities could not be shown on scan.

medicine. The observers had no knowledge as to the number of positive and negative cases that were in the series nor were they told of their results until after their third set of readings. True-positive, true-negative [TN/(TN + FP) × 100], false-positive, false-negative percentage [FN/(FN + TP) × 100], and overall accuracy (correct interpretations/total number of cases × 100) were calculated for each set of readings. Similar composite data for the three groups of readers were also calculated (Table 2).

RESULTS AND DISCUSSION

The difference in the Receiver Operator Characteristics (ROC) between the composite staff readings and that of the initial official interpretation is quite striking (Table 2). The staff has a significantly lower overall accuracy, 72.6 compared with 80% for the official interpretation but also a much lower false positive rate, 11.2 compared with 38.1%. In attempting to explain this difference, several factors appear to be involved. First, the individual reader had no knowledge of how many positive and negative scans there were. The usual rate of abnormal brain scans in our laboratory is about 14%. Since the proportion of positive scans in this series (46/67) is much higher than is encountered in routine daily practice, without being forewarned, the observer's attitude may have prevented him from calling such a high percentage of the studies abnormal. Second, the individuals interpreted the scans alone, without a group readout, the usual method of scan interpretation in our laboratory. This would seem to indicate that group interpretations are more accurate. Third, it seems quite likely that the false-positive rate of 38% for the official reading is due to the method of selection of normal cases in this study, i.e. many cases were chosen from cases initially read as abnormal which on long-term followup appeared normal. In searching a large number of abnormal reports, it is only possible to find either true-positive or false-positive results without the true and false-negative possibilities. Therefore, the method used to select the 67 cases can explain the high number of false-positive results in the official hospital reports.

There is wide variation between individual staff members which shows consistently that some tend to under-read while others tend to over-read although each individual tended to be consistent within himself. There was slight improvement in interpretation when the clinical history was presented. The residents did not do as well as the other groups; however, their final set of interpretations does very nearly equal that of the other groups, indicating most of this difference to be due to the improvement in ability of interpretation since the first set of readings were made when the resident was on the service a very short time.

The technologists showed wide variation of ability. Their results with history were slightly worse than without. The composite results of the technologists were about equal to that of the staff.

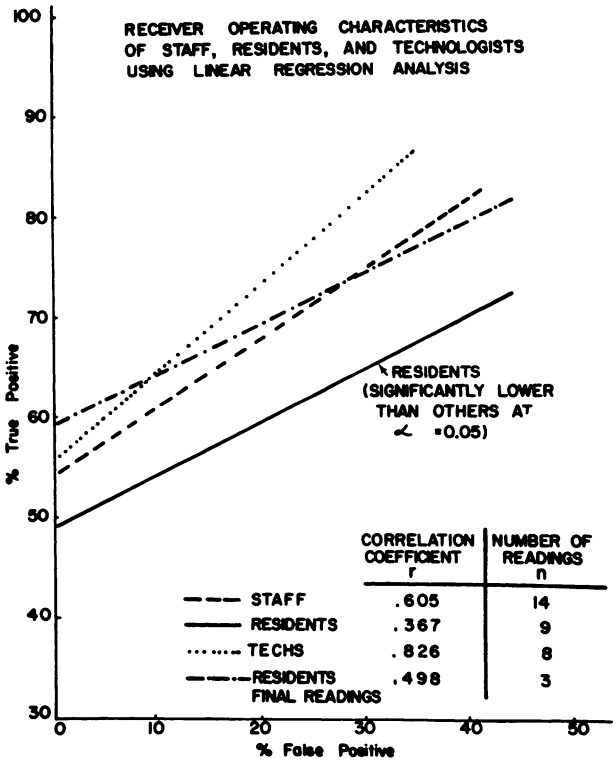

FIG. 1. Receiver Operating Characteristics of staff, residents, and technologists using linear regression analysis.

Figure 1 shows the composite Receiver Operator Characteristics of all three groups assuming a linear relationship for each group. It can be seen that the correlation coefficients for the lines are poor; however, when the slopes and intercepts are compared using the t-test, the line representing the residents ROC is significantly different from the others.

One explanation for the rather low overall accuracy of the groups was that 5 of the 67 scans were inaccurately read by everyone on every occasion. These five scans were technically of the poorest quality in the group. In all of the cases, the patients were found to have disease, but the scans were interpreted as negative. By subtracting the 5 cases from the total of 67 and using 62 in the denomination to calculate overall accuracy, the accuracy is increased to about 78%, more nearly approximating data given by others for brain scan accuracy.

Different levels of competence were found between the three groups involved in the study. As predicted by Lusted, Receiver Operator Characteristics do show that certain individuals tend to consistently over-read or under-read. Obviously, the goal of any series of interpretations is to approach TP% = 100 and FP% = 0. What can be done to achieve this, since at a given level of competence it is predicted that one can only move along an existing ROC curve? First, the state of the art must be improved by developing more advanced imaging equipment and more specific radiopharmaceuticals. Second, with studies such as these we may be able to analyze the cause for certain observer errors and possibly correct them, thus increasing observer accuracy. Third, multiple interpretations (group reading) can also improve Receiver Operating Characteristics.

The ROC curves generated in this study do not follow closely the shape predicted by Lusted. Perhaps the number of observations or the size of the sample was too small, thus introducing much statistical "noise."

The method of sample selection has an obvious effect on the ROC parameters. In a recent article by Krishnamurthy, et al (*11*), analysis of autopsy-correlated brain scans was made. In this sample, 84 cases were selected with the elimination of over 2,000 others. Obviously, the percentage of significant disease will be much higher in a population that is soon to have an autopsy than in one that is not. In examining a population in which almost all have disease, the result of a test can only be either truly positive or falsely negative since these parameters are calculated using the total number of patients with disease. On the other hand, true-negative and false-positive interpretations must be low since these are the patients without disease. These conclusions are corroborated by a very high (33%) false-negative rate and a very low (1.2%) false-positive rate.

Our study suggests that certain technologists are quite skilled in separating normal from abnormal results. Preliminary screening of examinations by these qualified technologists might also help in increasing overall accuracy.

Finally, these techniques can advise the interpreting physician concerning his own likelihood of errors in diagnosis and eventually will contribute to more accurate diagnosis with lessening of radiation exposure and expense to the patient.

ACKNOWLEDGMENTS

This work was originally presented at the 19th Annual Meeting of the Society of Nuclear Medicine, July 1972.

This work was supported in part by the following: U.S. Public Health Service Contract No. PH 86-67-212; the General Research Support Grant, RR 5408, National Institutes of Health; and the Albertine O. Schoepf Research Fund.

Representative products and manufacturers are named for identification only, and listing does not imply endorsement by the U.S. Department of Health, Education, and Welfare.

REFERENCES

1. VECCHIO TJ: Predictive value of a single diagnostic test in unselected population. *N Engl J Med* 274: 1171–1173, 1966

2. TAVERAS JM, WOOD EH: *Diagnostic Neuroradiology.* Baltimore, Williams & Wilkins, 1964

3. SCHEINBERG P, ZUNKER E: Complications of direct percutaneous carotid arteriography. *Arch Neurol* 8: 676–684, 1963

4. FEILD JR, LEE L, MCBURNEY RF: Complications of 1,000 brachial arteriograms. *J Neurosurg* 36: 324–332, 1972

5. MCAFEE JG, FUEGER GF: The value and limitations of scintillation scanning in the diagnosis of intracranial tumors. In *Scintillation Scanning in Nuclear Medicine,* Quinn JL, ed, Philadelphia, WB Saunders Co, 1964, pp 183–217

6. WITCOFSKI RL, MAYNARD CD, ROPER TJ: A composite analysis of the accuracy of the technetium-99m pertechnetate brain scan: Followup of 1,000 patients. *J Nucl Med* 8: 187–196, 1967

7. BELL RS, LOOP JW: Utility and futility of radiographic skull examination for trauma. *N Engl J Med* 284: 236–239, 1971

8. LUSTED LB: Decision-making studies in patient management. *N Engl J Med* 284: 416–424, 1971

9. LUSTED LB: Signal detectability and medical decision-making. *Science* 171: 1217–1219, 1971

10. YERUSHALMY J: Reliability of chest radiography in the diagnosis of pulmonary lesions. *Am J Surg* 89: 231–240, 1955

11. KRISHNAMURTHY GT, MEHTA A, TOMIYASU U, et al: Clinical value and limitations of ^{99m}Tc brain scan: An autopsy correlation. *J Nucl Med* 13: 373–378, 1972