24 August 1979, Volume 205, Number 4407 SCIENCE

Assessment of Diagnostic Technologies

Advanced measurement methods are illustrated in a study of computed tomography of the brain.

John A. Swets, Ronald M. Pickett, Susan F. Whitehead
David J. Getty, James A. Schnur, Joel B. Swets, Barbara A. Freeman

Many new diagnostic tests and many new, expensive imaging modalities are introduced into the health care system each year. Evaluating them can be difficult, particularly if alternative means exist to approximately the same diagnostic end. Yet, lest new techniques be introduced haphazardly, critical protocols and methods must be available so that they can be promptly assessed. We are currently completing for the National Cancer Institute a general protocol for the evaluation of diagnostic devices, with an emphasis on imaging modalities. The present study was undertaken both to refine and to illustrate this protocol (*1*).

Summary. A general protocol for rigorous evaluation of diagnostic systems in medicine was applied successfully in a comparative study of two radiologic techniques. Accuracies of computed tomography and radionuclide scanning in detecting, localizing, and diagnosing brain lesions were assessed with a sample of patients in whom tumor had been suspected. The principal means of analysis was the "relative operating characteristic," which is unique in providing a measure of accuracy that is largely independent of decision biases. Computed tomography was found to be substantially more accurate than radionuclide scanning.

To date, most comparative studies of imaging systems have taken one or the other of two inadequate approaches. One has been to measure fidelity—how well the system reveals the presence and detail of a standard test object, called a "phantom." The drawback of this approach is that what is measured is only the potential for mediating accurate detection and diagnosis of real lesions. The other approach, although based on using real cases and examining actual diagnostic performance, has been inadequate because of the simplistic accuracy indices that are usually obtained. Indices such as the proportion of true-positive responses, single true-positive false-positive pairs (and their ratio), single pairs of "sensitivity" and "specificity" values (and their sum), and agreement scores do not control for the influence of the reader's confidence threshold or "decision criterion," that is, the tendency to overcall or undercall disease, nor for the prevalence of disease in the study population at hand.

These two inadequacies are overcome in the approach of our general protocol. It specifies a procedure in which real cases and real diagnostic tasks are used and in which, moreover, performance is scored in relation to independent, external evidence. It further specifies a psychophysical method that generates performance data in the form of the relative (or receiver) operating characteristic (ROC). The ROC analysis provides an index of diagnostic accuracy that is independent of extra-image decision factors and of prior probabilities. It is borrowed from the general theory of signal detection (*2*), has been applied extensively in perceptual and cognitive studies in psychology (*3, 4*), and is now being increasingly applied in other fields (*5*), particularly in medicine (*6*).

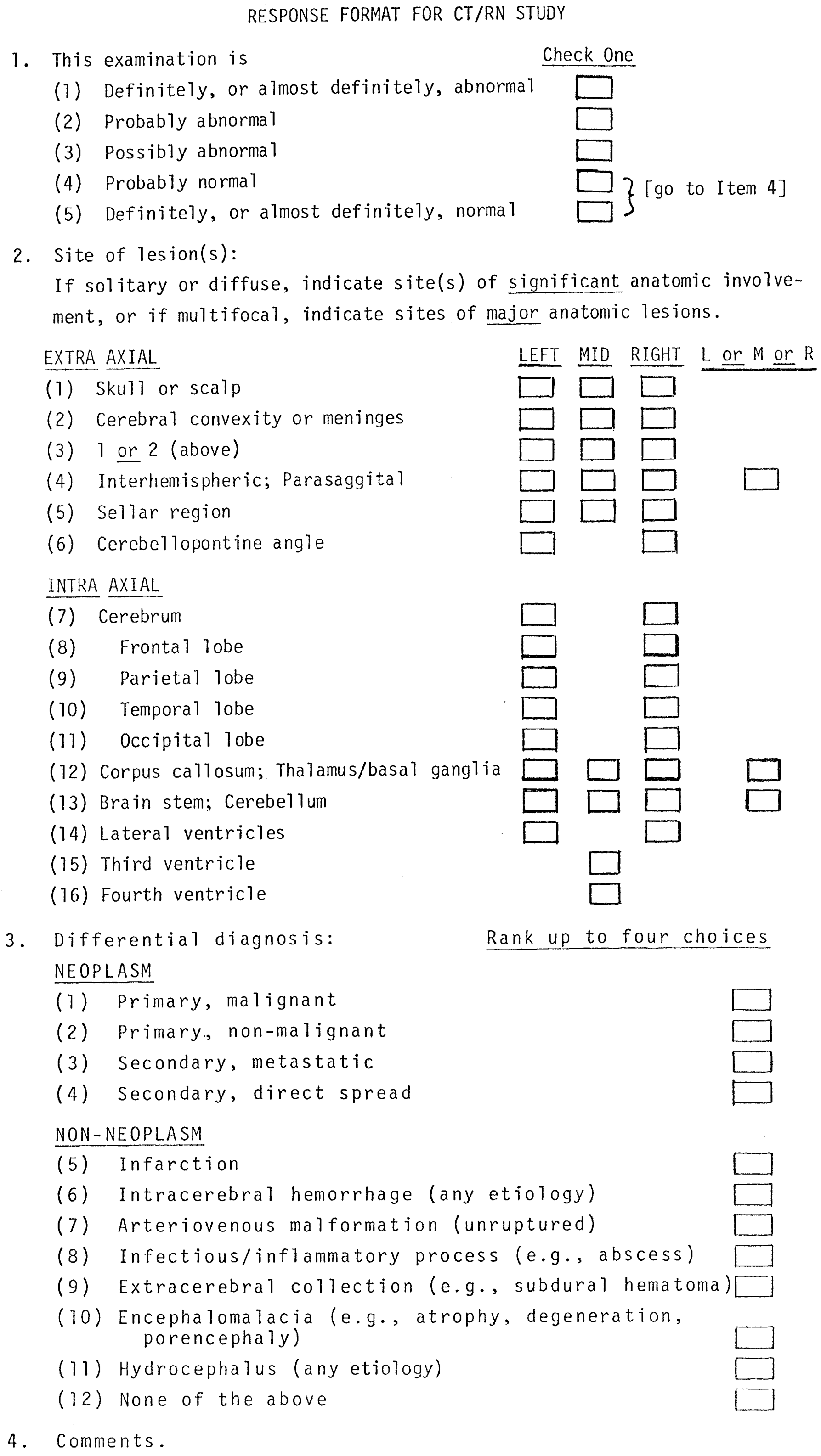
The ROC is a curve showing the various trade-offs existing between proportions of true-positive and false-positive responses, as the decision criterion is systematically varied, for a given capacity to discriminate between positive and negative cases. An ROC index of accuracy reflects the location of the entire curve rather than any particular operating point on the curve. An extension of the detection ROC treats also localization and classification of abnormalities. The recommended measurements, in addition, supply the appropriate starting point for assessing the usefulness of a diagnostic system in terms of medical efficacy, risk, and cost. The basic concepts of ROC analysis have recently been explicated for a general medical audience (*7*).

In the present study, we applied the ROC-centered psychophysical methods to measure the accuracy of computed tomography (CT) and radionuclide scanning (RN) in a sample of patients in whom brain tumor had been suspected. Collection of the test images and other case materials, over a 3-year period, was sponsored by the National Cancer Institute in a collaborative study at five major medical centers (*8*). Although the images were read at the several sites and various analyses of the data made at those sites as well as at a statistical coordinating center, we were commissioned to provide a central, retrospective analysis of the data by means of a reading test conducted in accordance with the new protocol, and the images were read anew in

The authors are staff members of or consultants to Bolt Beranek and Newman Inc., 50 Moulton Street, Cambridge, Massachusetts 02138.

0036-8075/79/0824-0753$01.75/0 Copyright © 1979 AAAS

our laboratory by radiologists recruited for that purpose. Because the protocol had not been applied before, we were asking in our study, in general, whether our psychophysical methods could be applied to complex diagnostic tasks to obtain reliable and valid estimates of accuracy or whether the complexity of real diagnostic tasks is so great as to preclude such estimates. In specific terms, we set out to measure the accuracy of CT in detecting, localizing, and diagnosing brain lesions, and also, as a point of reference, the accuracy of RN, which is CT's main competitor as a relatively noninvasive technique.



RESPONSE FORMAT FOR CT/RN STUDY

1. This examination is — Check One

(1)	Definitely, or almost definitely, abnormal	□
(2)	Probably abnormal	□
(3)	Possibly abnormal	□
(4)	Probably normal	□ } [go to Item 4]
(5)	Definitely, or almost definitely, normal	□ }

2. Site of lesion(s):

If solitary or diffuse, indicate site(s) of significant anatomic involvement, or if multifocal, indicate sites of major anatomic lesions.

	LEFT	MID	RIGHT	L or M or R
EXTRA AXIAL				
(1) Skull or scalp	□	□	□	
(2) Cerebral convexity or meninges	□	□	□	
(3) 1 or 2 (above)	□	□	□	
(4) Interhemispheric; Parasaggital	□	□	□	□
(5) Sellar region	□	□	□	
(6) Cerebellopontine angle	□		□	
INTRA AXIAL				
(7) Cerebrum	□		□	
(8) Frontal lobe	□		□	
(9) Parietal lobe	□		□	
(10) Temporal lobe	□		□	
(11) Occipital lobe	□		□	
(12) Corpus callosum; Thalamus/basal ganglia	□	□	□	□
(13) Brain stem; Cerebellum	□	□	□	□
(14) Lateral ventricles	□		□	
(15) Third ventricle		□		
(16) Fourth ventricle		□		

3. Differential diagnosis: — Rank up to four choices

NEOPLASM	
(1) Primary, malignant	□
(2) Primary, non-malignant	□
(3) Secondary, metastatic	□
(4) Secondary, direct spread	□
NON-NEOPLASM	
(5) Infarction	□
(6) Intracerebral hemorrhage (any etiology)	□
(7) Arteriovenous malformation (unruptured)	□
(8) Infectious/inflammatory process (e.g., abscess)	□
(9) Extracerebral collection (e.g., subdural hematoma)	□
(10) Encephalomalacia (e.g., atrophy, degeneration, porencephaly)	□
(11) Hydrocephalus (any etiology)	□
(12) None of the above	□

4. Comments.

Fig. 1. Response form for readings of CT and RN images.

Materials and Methods

The major characteristics of the cases, images, and other case materials were determined for the purposes of the collaborative study—a study with certain goals different from ours and, indeed, designed in advance of our study. To serve our goals, we had to impose certain restrictions which excluded from our sample a large number of the cases in the original sample. These restrictions have mainly to do with the availability of case data for establishing a reasonably credible diagnosis.

We were able to obtain imagery for 84 positive cases and 52 negative cases that met our requirements for (i) imaging in both the RN and the CT modalities, the latter with and without contrast enhancement, and (ii) credible "truth" data. For positives we required confirmation by autopsy or by histology based on cerebral biopsy or tissue from a craniotomy. The negatives were selected from a group of patients with extracerebral cancer (because the presumed normal subjects in the collaborative study were examined only by CT). They were accepted for our study if they were asymptomatic for cranial disease at the time of imaging, if the results of any other tests done then were negative, and if there was either a finding of no intracranial lesion at autopsy or live follow-up with no neurologic symptoms at least 8 months after the imaging. The diagnoses represented in our positive cases are shown in Table 1.

Twelve radiologists participated in our reading test, each spending 6 days in our laboratory over a period of 6 months. The six CT readers had an average of 3.3 years' experience in reading CT scans, about as much as was available. The six RN readers had an average of 11 years of experience. The response form, shown in slightly abbreviated form in Fig. 1, provided a basis for standardized scoring. The responses to item 1 are the basis for the conventional ROC curve. The five rating categories supply four points, representing four different decision criteria, on the ROC curve.

Answers for items 2 and 3 (site and type of lesion) were derived from available truth data by a neuroradiologist and a neuropathologist working independently, the latter having the final word on the few disagreements. The available data for positive cases included particulars about the type and locus of lesion in the codes of the Systematized Nomenclature of Pathology.

The CT display rested on a table which also provided space for writing on the response forms. A test administrator was seated to the right of the reader at a video terminal, from which he controlled the pace of the session. Both the display and the terminal were connected to a minicomputer that transferred images from magnetic tape to the display as required, stored the responses, and generally controlled the trial-by-trial sequence of events throughout a session.

The image-display system consisted of a COMTAL model 8000-SA image-display processor driven by the computer and in turn driving a CONRAC SNA 17-inch (43-centimeter) black-and-white video monitor. The layout of the control panel was essentially the same as that of the EMI Diagnostic Display Console, providing fast and slow controls for window level (WL) and a discrete control of window width (WW), including measure mode. An image on the monitor consisted of up to eight slices (images of the brain at different levels) presented simultaneously in a 3 × 3 matrix; the vacant cell in the lower right-hand corner contained information about current WL and WW. Each slice was approximately 70 by 70 millimeters. Each data element in the 160 by 160 EMI image matrix was represented by an independent pixel on the screen at one of 256 brightness levels.

The RN films had been taken by gamma scintillation cameras, with technetium-99m pertechnetate as the radiotracer. They were viewed on standard view boxes, two banks of four-panel GE Fluroline Illuminators. The test administrator was seated at a video terminal to the right of the view boxes. Readers were provided with both magnifying and minifying lenses. Films in our reading-test sample included a wide range of sizes and formats, namely, 11 by 14 inch individual standard views (anterior, posterior, left and right laterals, and vertex), 8 by 10 inch films containing multiple views, and 70-mm and 35-mm films of standard views. A portion of these cases included immediate views (30 minutes, minus the vertex), delayed views, and flow studies; another portion consisted of immediate and delayed views only; a third portion consisted of delayed views only. Occasional cases included special views (obliques, Waters').

Each day of the reading test was divided into four sessions of 1 to $1^1/_2$ hours each. In addition, there was a preliminary session of four practice cases on the first and second visits to familiarize the readers with the procedure and equipment. For RN, the test administrator mounted all the available films for each case on the view boxes in a standard order prescribed by the reader (the readers did not handle the films). For CT, the computer first displayed the set of slices associated with one mode (either contrast or noncontrast); both modes were then available for viewing as the reader desired.

Table 1. Diagnoses represented among the abnormal cases in our sample.

Diagnosis	Number
Primary tumors	
Glioma	31
Meningioma	13
Neurilemoma	5
Chromophobe adenocarcinoma	3
Craniopharyngioma	1
Pinealoma	1
Colloid cyst	1
Benign teratoma	1
Secondary tumors	
Metastatic	18
Direct spread/metastatic	1
Nonneoplasms	
Hematoma	3
Hemorrhage	1
Infarction	2
Arteriovenous malformation	1
Aneurysm	1
Chronic diffuse inflammation	1

At the time the images were presented, the test administrator provided the reader with limited background information about the case—the patient's age, time of imaging relative to isotope injection (RN), and presence or absence of a contrast agent (CT). The reader was asked to vocalize his choices as he marked them on the response form to enable the test administrator to enter them into the computer disk storage. To ensure a reasonable rate of progress, a time limit of 7 minutes was set for each case. This time limit was meant to be liberal, and was, indeed, rarely approached; when it was, the test administrator provided a warning about 1 minute before the deadline.

Almost all of the 136 cases comprising our sample were viewed twice by each CT and RN reader, once on each of two different days separated generally by more than 1 month. The replication of readings was undertaken to obtain an estimate of interreader reliability, and potentially to increase the reliability of our accuracy indices, but in the present analyses we focus on the first viewing of the cases. Every reader viewed the same set of cases on a given day in the sequence of 6 days but had a different random ordering of cases, with new and formerly viewed cases randomly intermixed. In order to minimize the effect of the random sequential ordering of cases upon the CT-RN comparison, each of the six RN readers was randomly paired with one of the six CT readers, and a particular RN reader received the same case sequence on each of the 6 days as the paired CT reader did.

Our test readers judged about 45 cases per day to be a reasonable load in our setting. Relevant aspects of our setting were that interruptions were not permitted, everything we could think of to facilitate the reading process was done, only limited case background had to be considered, and treatment of the patient did not follow upon the diagnosis rendered.

Data Analyses

The axes of the ROC plot—proportions of false-positive and true-positive responses—are symbolized as $P(FP)$ and $P(TP)$. We obtain four points on an ROC curve for each reader by considering the boundaries of the five categories of the rating scale (item 1, Fig. 1) as different decision criteria, or confidence thresholds, for a positive response. One point, representing a strict criterion, corresponds to $P(FP)$ and $P(TP)$ for only those cases that are placed in the highest confidence category (category 1). A second point is based on the cases placed in category 2 together with those placed in category 1, and represents a less strict criterion; both $P(TP)$ and $P(FP)$ are higher than for the first point. And so on through the fourth category, to a very lenient criterion. Four points are obtained from five categories because $P(TP) = P(FP) = 1.0$ when all five categories are combined.

Our indices of detection accuracy are based on the assumption that the empirical ROC can be viewed in terms of the normal, or Gaussian, probability distribution. Consistent with this assumption, we plot ROC data on double-probability (binormal) coordinates, on which the normal deviate is linearly spaced, and we fit the ROC points by a straight line. Our fundamental index, termed A_z, is the area beneath the fitted, binormal ROC, and ranges from a minimum of 0.50—representing chance behavior, for an ROC along the major diagonal, where $P(TP) = P(FP)$—to a maximum of 1.0—representing perfect discrimination, for an ROC showing $P(TP) = 1.0$ for all values of $P(FP)$. We report also the intercept and slope of the ROC, which provide a basis for reconstructing the ROC.

Though A_z is a quantity derived from the curve that does not correspond to any of the commonly observed response proportions, it can be related to a particular response proportion and so gain additional meaning. In a so-called forced-choice test, in which each trial presents one image from the set of positive cases and one image from the set of negative cases, and the reader is asked to say which is which, A_z is theoretically equal to the probability of a correct response (*3*, p. 47). In other words, A_z ranges from 0.50 to 1.0 in the same manner as does the proportion of correct responses in a two-alternative forced-choice test.

A slightly revised version of a computer program described by Dorfman and Alf (*9*) provides chi-square measures of the goodness-of-fit of the binormal ROC to ROC data, maximum-likelihood estimates of ROC indices including A_z, and the sampling variances of those estimates.

To measure the accuracy of detection plus localization, of detection plus classification, and of detection plus localization plus classification, we use the "joint" ROC as advanced by Starr, Metz, Lusted, and Goodenough (*10*). The abscissa is the same as that of the conventional detection ROC (namely, the proportion of false-positive responses); the ordinate is the proportion of true-positive responses that are also correct with regard to another dimension or two, that is, localization or classification or both. This joint ROC is plotted without theoretical assumptions about its form, on linear probability coordinates. The joint ROC may also be indexed by the area under the empirical curve, but for convenience here we consider only the index supplied by the ordinate value at a given abscissa value.

Detection

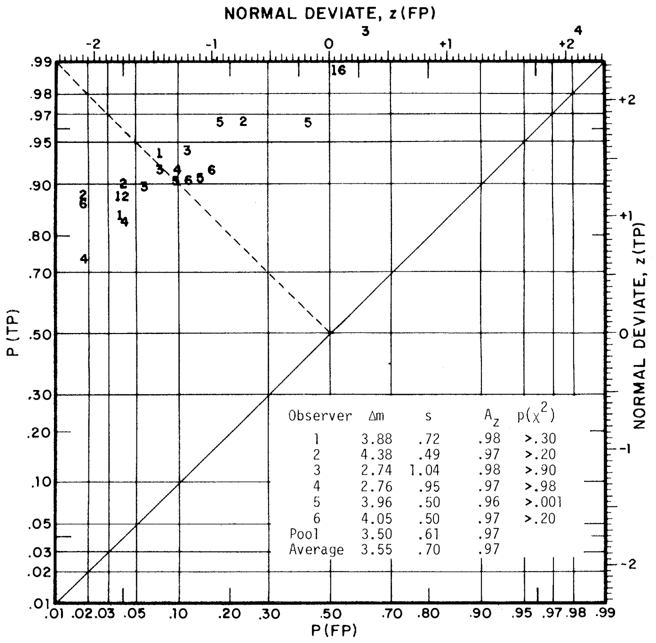
In Fig. 2 individual detection ROC's are shown for the six CT readers, based on just the first reading of the cases. The four points for each reader are indicated by that reader's designated number, 1 to 6. (Readers 3 and 4 each show one point outside the figure's square.) The column labeled $p(\chi^2)$ in the inset of the figure indicates that the empirical ROC's are fitted well by the assumed straight line: the probabilities associated with the chi-square measure of goodness-of-fit are, with one exception, substantially above the level that would reject linearity.

The absolute value of the intercept of the ROC with the axis at *P(TP)* = 0.50, called Δm, and the slope (*s*) of the linear ROC's, are given for each reader; these quantities are expressed in units of the normal deviate as given on the upper and right coordinates. As indicated, Δm and *s* are listed to suggest how ROC curves may be simply but completely characterized. We make no further use of Δm here, but the values of *s* typical of a diagnostic modality are of interest. For example, we would want to establish that there is no substantial difference in *s* between two modalities before comparing them in terms of the single-parameter accuracy index A_z.

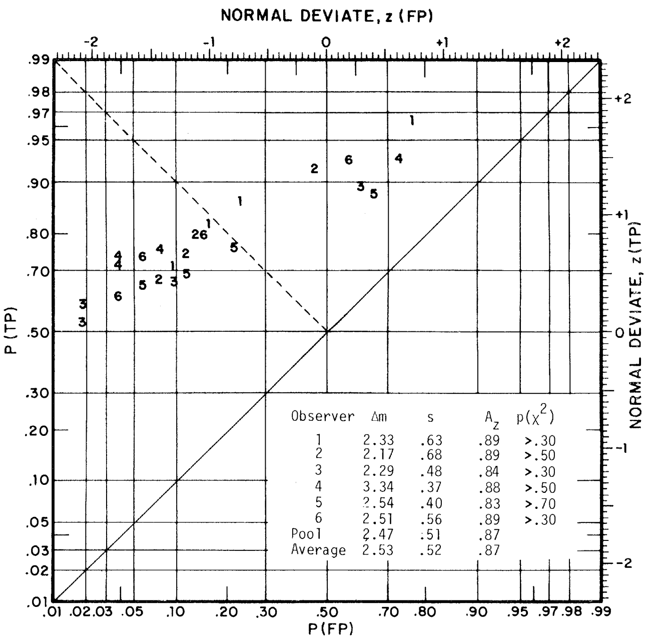
The individual ROC's are seen to lie in a strikingly narrow band. The principal index, A_z, shown in the inset of the figure, varies only from 0.96 to 0.98. The CT readers further agree with one another to a considerable extent on a case-by-case basis: if we collapse the five categories of the rating scale to two (with categories 1, 2, and 3 defined as abnormal and categories 4 and 5 defined as normal), then the average percent agreement of each CT reader with the other five is 91.2. This agreement is only slightly less than the intrareader agreement over two readings of the cases, namely, 93.2 percent. We have estimated the average product-moment correlation between reader pairs to be approximately 0.65. This figure is obtained via Kendall's partial tau (*11*); reader correctness was partialed out to reduce the influence of test difficulty on the correlation estimate.

The six RN readers also show considerable uniformity on the first reading of the cases (Fig. 3), with values of A_z ranging from 0.83 to 0.89. Intrareader and interreader agreement for RN are 90.1 and 86.1 percent, respectively. The average coefficient of product-moment correlation between reader pairs was estimated to be 0.75.

The interreader correlations, within CT and within RN, were taken into account in estimating the standard errors and statistical significance levels reported in the following. That is to say, in combining readers to estimate the statistical significance of overall differences between the modalities, we did not consider the readers to be statistically independent—or, in effect, assume our sample size to be the number of cases times the number of readers—but rather used a formula developed by Jarrett and Henry (*12*) to ascertain how the standard error for one (average) reader is reduced



Observer	Δm	s	A_z	$p(\chi^2)$
1	3.88	.72	.98	>.30
2	4.38	.49	.97	>.20
3	2.74	1.04	.98	>.90
4	2.76	.95	.97	>.98
5	3.96	.50	.96	>.001
6	4.05	.50	.97	>.20
Pool	3.50	.61	.97	
Average	3.55	.70	.97	



Observer	Δm	s	A_z	$p(\chi^2)$
1	2.33	.63	.89	>.30
2	2.17	.68	.89	>.50
3	2.29	.48	.84	>.30
4	3.34	.37	.88	>.50
5	2.54	.40	.83	>.70
6	2.51	.56	.89	>.30
Pool	2.47	.51	.87	
Average	2.53	.52	.87	

Fig. 2 (left). Individual detection ROC's for the six CT readers, with various curve parameters.

Fig. 3 (right). Individual detection ROC's for the six RN readers, with various curve parameters.

by adding other, correlated readers. We find, by the way, that our six correlated readers are the equivalent of about one and one-third independent readers, as concerns the effect of replicated readers on estimates of sample size and standard error. However, the correlation of readers across the modalities, as induced by our use of the same cases in the two modalities, was not taken into account in calculating the significance levels reported in the following, so those levels are conservative. Our significance levels are also conservative in that we have used only the first of the two readings of the cases in our calculations, and thus have not enhanced sample size to the extent allowed by the combination of the two readings.

In the insets of both Figs. 2 and 3, it can be seen that pooling the six readers' raw data, that is, treating the six readers as one reader by merging their rating responses, leads to results much like those obtained when the average (arithmetic mean) of the six individuals' derived indices is taken.

Based on pooled data, the comparison of primary interest is shown in Fig. 4. The ROC for CT is seen to be consistently above that for RN, with respective values of A_z of 0.97 and 0.87. The standard errors of those values of A_z are 0.012 and 0.027, respectively, and so the 95 percent confidence intervals are approximately 0.95 to 0.99 for CT and 0.82 to 0.92 for RN. The difference in A_z has associated a $p = .0007$ under the null hypothesis.

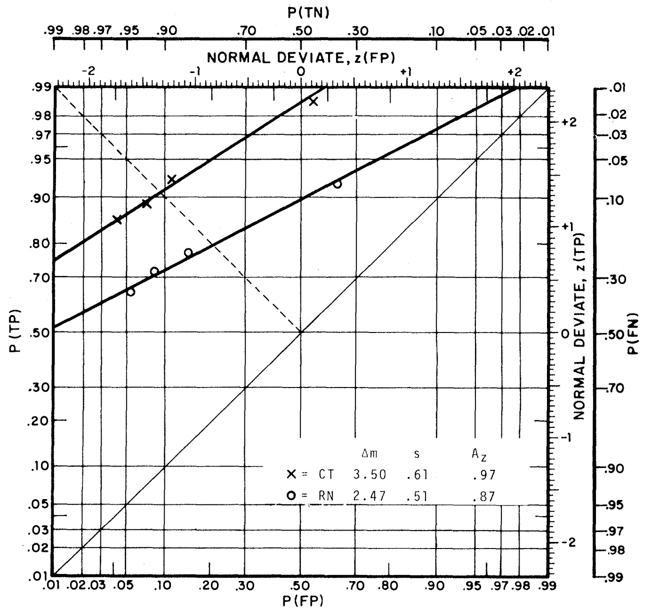
According to the ROC curves of Fig. 4, at a false-positive rate of 0.10, CT would attain a true-positive proportion of 0.91 while RN would attain a true-positive proportion of 0.73. The probability scales given at the top and right of this figure facilitate reading the graph in the other direction: at a false-negative rate of 0.10, the true-negative proportions would be 0.91 for CT and 0.49 for RN.

Let us acknowledge that in the clinic a reader will sometimes avoid both the positive and the negative response and issue an uncertain or "equivocal" response. We can deal with the use of three responses because the proportion of equivocal responses made to positive cases plus the other two proportions based on positive cases, *P(TP)* and *P(FN)*, must add up to 1.0, and similarly the proportion of equivocal responses made to negative cases and *P(TN)* and *P(FP)* must add up to 1.0. Thus, we can calculate the proportion of equivocal responses that would have to be made to satisfy any given limits on the two types of error. For example, according to the data of Fig. 4, CT could maintain both *P(FP)* and *P(FN)* ≤ 0.10 and produce *P(TP)* = *P(TN)* $\cong 0.90$ while sorting all the cases into positive or negative. Meanwhile, to maintain *P(FP)* = *P(FN)* ≤ 0.10, RN would produce *P(TP)* = 0.73 and *P(TN)* = 0.49 and fail to sort definitively 17 percent of the positives and 41 percent of the negatives. Such analyses of response probabilities make clear that an observed difference of 0.10 in A_z has substantial implications for practice.

Localization and Differential Diagnosis

Various joint ROC's for CT and RN (based on pooled data for each modality) are shown in Fig. 5, along with a reproduction of the detection ROC's. They are (i) detection plus localization, (ii) detection plus "classification," or differential diagnosis, when any of the first four classification responses is scored as correct, (iii) detection plus classification when only the first choice of a differential diagnosis is scored as correct, and (iv) ROC's determined when the response must be correct with respect to detection and localization and first-choice classification in order to be scored as a true-positive response. Each curve consists of three points because localization and classification responses were given only for cases placed in categories 1 to 3 of the rating scale in item 1 of the response form.

As the true-positive response is de-



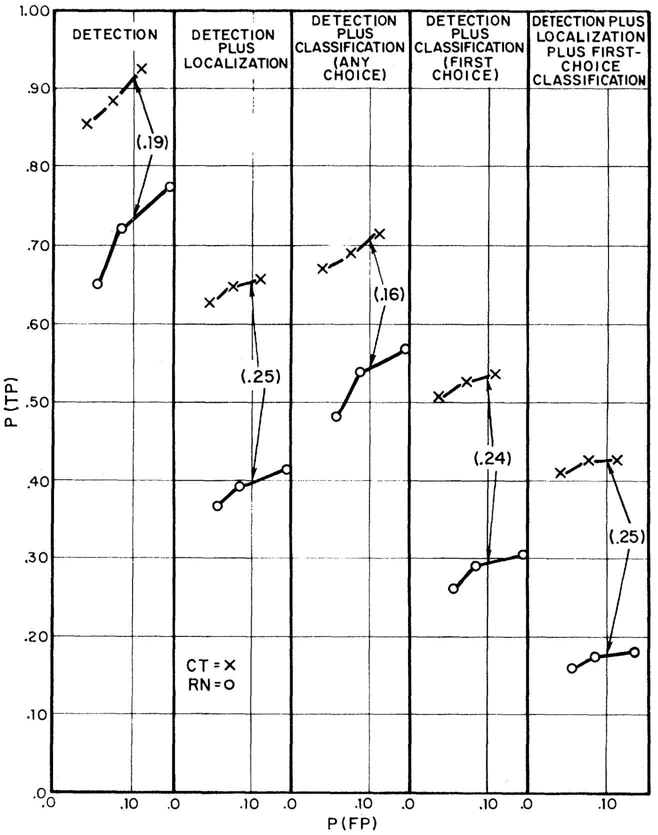


Fig. 4 (left). Detection ROC's for CT and RN, based on pooled data for the six readers of each modality, with various curve parameters. Fig. 5 (right). Joint ROC's for CT and RN, along with detection ROC's, based on pooled data for the six readers of each modality.

fined in progressively more demanding fashion, the drop in $P(TP)$ is precipitous, amounting to about 0.50 for each modality. At $P(FP) = 0.10$, the advantage of CT over RN in $P(TP)$ is on the order of 0.20 to 0.25 throughout the various levels of performance measurement. On the basis of the binomial estimate of sampling variance, the probabilities under the null hypothesis across the five panels of Fig. 5, for the differences in $P(TP)$, are 0.00008, 0.0001, 0.01, 0.0002, and 0.00006. While the difference between modalities is fairly constant, the ratio of $P(TP)$ for CT to $P(TP)$ for RN grows as the detection ROC is extended to include localization, classification, and both—from about 1.25 to 2.5.

Other Analyses

In principle, detection ROC's may be calculated for various subsets of a broad sample of cases—in which the abnormal cases correspond to single types and sites of pathology—provided that the subsets are large enough to yield reliable measures. We calculated detection ROC's for primary tumor (56 cases), glioma (31 cases), meningioma (13 cases), and metastatic tumor (18 cases); and for lesions above (71 cases) and below (9 cases) the tentorium. The differences between CT and RN in A_z were significant at the 0.05 confidence level for the subsets with more than 30 cases and not significant for the smaller samples. None of the differences between the several types and sites within a modality reach significance in our sample. However, the pattern of results suggests that in both modalities the amount by which the detection of primary tumors exceeds that of metastatic tumors, and the amount by which the detection of supratentorial lesions exceeds that of infratentorial lesions, would be statistically significant in a slightly larger sample.

Our experimental design also permitted a comparison of CT cases read with and without contrast enhancement. In brief, detection performance between contrast and noncontrast scans differed insignificantly; however, contrast scans were significantly superior to noncontrast scans for detection plus classification.

To test whether using CT and RN together would improve detection performance over the better system alone, two decision rules were analyzed: (i) make a positive response ("abnormal") if either system's response is positive and (ii) make a positive response only if both systems are positive. For the entire sample of cases, the ROC for the combined performance resembled very closely the ROC for CT alone.

We also asked whether combining the detection reports of three readers within a modality would show an advantage over a single reader. The decision rules were to make a positive team response if (i) any one reader was positive, (ii) if any two readers were positive, and (iii) only if all three were positive. The readers' reports were independent in the sense that the readers did not communicate with each other. However, as would be expected from the high correlation among readers mentioned earlier, three readers were no better than one. The various decision rules affected the position of the points on the ROC curve, of course, but not the location of the curve.

Discussion

Three factors must be considered as possible qualifications in the interpretation of our main results. Two of them stem from an extensive analysis, described elsewhere (*1*) in detail, of differences between our test sample and the much larger sample of the collaborative study. That analysis indicates, first, that the lesions in our test sample tended to be more apparent clinically, hence probably larger and more visible for both CT and RN, than lesions in the original collaborative sample. Second, there was a significant underrepresentation in our test sample, relative to the whole collaborative sample, of (i) cases falsely considered negative by the RN readers at the collaborative sites and (ii) cases for which the RN reader in the collaborative study failed to give the site of a lesion—that is, cases relatively difficult for RN. A third possible qualification stems from the large variability of image format in our cases as presented by RN and the opinion of our RN readers that many of those presentations were of relatively poor quality.

Concerning the first point, that our sample of cases may have been somewhat easier to diagnose than is realistic for either CT or RN, we observe again that cases in our test were read with little specific case background. Both elements considered, CT performed very well. Quite possibly RN was affected more than CT by the lack of case background.

Regarding the second and third factors mentioned as qualifications, we believe that the factor favoring RN relative to CT had at least as much effect as the factor hampering RN relative to CT and, therefore, that our comparative results may be taken at face value. These results show CT performing substantially better than RN in the detection, localization, and differential diagnosis of intracranial lesions in patients in whom tumor is suspected.

This interpretation is consistent with our analysis of a sample of some 1200 cases, as read originally by participants in the collaborative study at their respective sites, with case background available, and scored according to what was regarded as the most likely diagnosis (based on all the accumulated evidence, not just histology) at the time of our analysis. In that collaborative sample the index A_z was 0.94 for CT (compared to 0.97 in our sample) and 0.82 for RN (compared to 0.87 in our sample).

We have illustrated ROC analysis and associated study procedures in a psychophysical approach to evaluation of the accuracy of imaging techniques. Our approach obtains an ROC accuracy index that is relative to independent and credible truth, yielded by real readers, based on real cases, and measured under controlled conditions. The ROC analysis can be used to measure accuracy in detecting characteristics of a phantom test object or simulated lesions, but then it indexes the potential of the modality to affect diagnosis rather than giving a direct estimate of how the modality affects diagnosis. The ROC analysis can be used to measure the accuracy of human diagnostic judgments based on evidence in addition to, or entirely other than, image information. And it can be used to evaluate mechanized diagnostic tests that yield a single number as a result (for example, 24-hour thyroid uptake, various protein levels), with a number greater than some criterion number taken as an indication of abnormality. The present application is among the more complex of the applications mentioned, and the ROC-centered approach is seen to be practical in this application. Thus, a rigorous and objective evaluation method can be used consistently across the range of diagnostic systems.

Although our methods would apply, we did not try to develop a protocol to aid the physician in deciding whether to order CT or RN in individual cases. That development would require measures of accuracy on separate categories of cases divided according to presenting history, signs, and symptoms. It would require a larger number of cases than the number available to us that met the requirements of our test sample.

While confining our present assessment to an index of accuracy, we have mentioned that the ROC is the proper basis for an evaluation of the usefulness of a diagnostic system, which would include elements of medical efficacy, risk, and cost. In such an evaluation one proceeds from probabilities of responses based on the diagnostic system under study, through a further decision-therapeutic tree, to health outcomes. The ROC is a means of determining the response probabilities appropriate to the best available estimates of the values, costs, and event probabilities that inhere in the relevant diagnostic and therapeutic context—and not merely the response probabilities associated with whatever decision criterion might have been employed in a test of the system (*13*). Costs and benefits may, therefore, be determined for a system operating at its best.

References and Notes

1. The materials and methods of this study are described in more detail, along with additional results and discussion, in *Technical Report No. 3818* from Bolt Beranek and Newman Inc., to the National Cancer Institute (1979).
2. W. W. Peterson *et al.*, *Trans. IRE Prof. Group Inf. Theory* **PGIT-4**, 171 (1954).
3. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966; reprinted by Krieger, New York, 1974).
4. J. A. Swets, *Science* **182**, 990 (1973).
5. ______ and D. M. Green, in *Psychology: From Research to Practice*, H. L. Pick, Jr., H. W. Leibowitz, J. E. Singer, A. Steinschneider, H. W. Stevenson, Eds. (Plenum, New York, 1978), pp. 311-331.
6. L. B. Lusted, *Introduction to Medical Decision Making* (Thomas, Springfield, Ill., 1968); B. J. McNeil, B. Keeler, Jr., S. J. Adelstein, *N. Engl. J. Med.* **293**, 211 (1975); J. A. Swets, *Invest. Radiol.* **14** (No. 2), 109 (1979).
7. C. E. Metz, *Semin. Nucl. Med.* **8**, 283 (1978).
8. Principal investigators and institutions participating in the collaborative study were H. L. Baker, Jr., Mayo Clinic; D. O. Davis, George Washington University Medical Center; S. K. Hilal, Columbia University; P. F. J. New, Massachusetts General Hospital; and D. G. Potts, Cornell University Medical Center. Data coordination for the collaborative study was supplied by C. R. Buncher, University of Cincinnati Medical Center.
9. D. D. Dorfman and E. Alf, Jr., *J. Math. Psychol.* **6**, 487 (1969).
10. S. J. Starr, C. E. Metz, L. B. Lusted, D. J. Goodenough, *Radiology* **116**, 533 (1975).
11. M. G. Kendall, *Rank Correlation Methods* (Hafner, New York, 1962).
12. R. F. Jarrett and F. M. Henry, *J. Psychol.* **31**, 175 (1951).
13. J. A. Swets and J. B. Swets, in *Proceedings of the Joint American College of Radiology—IEEE Conference on Computers in Radiology* (IEEE Computer Society, Silver Spring, Md., in press).
14. Actively representing the sponsor to the project were R. Q. Blackwell, W. Pomerance, J. M. S. Prewitt, and B. Radovich. We are indebted to Dr. Prewitt for recognizing the need for a general protocol and for suggesting the present application. Members of an advisory panel, who contributed extensively to the protocol development and the design of this study, were S. J. Adelstein, H. L. Kundel, L. B. Lusted, B. J. McNeil, C. E. Metz, and J. E. K. Smith. J. A. Schnur participated throughout the project; consultants for the study reported here were W. B. Kaplan and G. M. Kleinman. We are indebted also to the principal investigators in the collaborative study (*8*). We thank the following radiologists, who participated as readers in the tests conducted at Bolt Beranek and Newman Inc.: computed tomography—L. R. Altemus, R. A. Baker, A. Duncan, D. Kido, J. Lin, and T. P. Naidrich; radionuclide scanning—H. L. Chandler, J. P. Clements, T. C. Hill, L. Malmud, F. D. Thomas, and D. E. Tow.

Molecular Thermodynamics for Chemical Process Design

The properties of fluid mixtures must be understood for economic manufacture of chemical products.

J. M. Prausnitz

Chemical engineering design is concerned with finding economic and efficient methods for producing on a large scale what the chemist or materials scientist produces in small quantities. In meeting this concern, the chemical engineer must conceive, design, and build large units of equipment for processing large quantities of gases, liquids, and solids. Rational design of this equipment requires quantitative knowledge of the properties of the materials to be processed. Usually, these materials are mixtures.

Summary. Chemical process design requires quantitative information on the equilibrium properties of a variety of fluid mixtures. Since the experimental effort needed to provide this information is often prohibitive in cost and time, chemical engineers must utilize rational estimation techniques based on limited experimental data. The basis for such techniques is molecular thermodynamics, a synthesis of classical and statistical thermodynamics, molecular physics, and physical chemistry.

While test tubes and glass retorts are useful for manufacturing a few grams of a chemical substance, entirely different equipment is needed when manufacturing tons of that substance. Similarly, the chemical processes used for large-scale production are often highly dissimilar from those used for small-scale production. For example, a standard laboratory method for producing a few grams of hydrogen is to react hydrochloric acid with zinc, giving zinc chloride in addition to hydrogen. However, when hydrogen is needed in large amounts this procedure is inefficient, not only because the reactants are too expensive, but also because the available supply of zinc is too small to meet the world's need for hydrogen and because there would be a severe problem of what to do with vast amounts of zinc chloride. In principle, the zinc chloride could be reduced to recover zinc, but the expense of that operation would be prohibitive. For large-scale production of hydrogen, the traditional industrial method is to react steam with coal or other hydrocarbons or to oxidize partially natural gas with air.

In addition to dominant economic factors, there is a significant technical difference between laboratory-scale and industrial-scale production of typical chemical products. In the laboratory, the chemist generally uses pure reactants (for example, pure hydrochloric acid from one bottle and pure zinc from another) and the reaction is often so carried out that the desired product (gaseous hydrogen) and the undesired side product (solid zinc chloride) appear as separate phases, thereby avoiding any cumbersome separation operation. In industrial

The author is a professor in the Department of Chemical Engineering, University of California, Berkeley 94720.