Articles

Measuring the Accuracy of Diagnostic Systems

JOHN A. SWETS

Diagnostic systems of several kinds are used to distinguish between two classes of events, essentially "signals" and "noise." For them, analysis in terms of the "relative operating characteristic" of signal detection theory provides a precise and valid measure of diagnostic accuracy. It is the only measure available that is uninfluenced by decision biases and prior probabilities, and it places the performances of diverse systems on a common, easily interpreted scale. Representative values of this measure are reported here for systems in medical imaging, materials testing, weather forecasting, information retrieval, polygraph lie detection, and aptitude testing. Though the measure itself is sound, the values obtained from tests of diagnostic systems often require qualification because the test data on which they are based are of unsure quality. A common set of problems in testing is faced in all fields. How well these problems are handled, or can be handled in a given field, determines the degree of confidence that can be placed in a measured value of accuracy. Some fields fare much better than others.

DIAGNOSTIC SYSTEMS ARE ALL AROUND US. THEY ARE used to reveal diseases in people, malfunctions in nuclear power plants, flaws in manufactured products, threatening activities of foreign enemies, collision courses of aircraft, and entries of burglars. Such undesirable conditions and events usually call for corrective action. Other diagnostic systems are used to make judicious selection from many objects. Included are job or school applicants who are likely to succeed, income tax returns that are fraudulent, oil deposits in the ground, criminal suspects who lie, and relevant documents in a library. Still other diagnostic systems are used to predict future events. Examples are forecasts of the weather and of economic change.

It is immediately evident that diagnostic systems of this sort are not perfectly accurate. It is also clear that good, quantitative assessments of their degree of accuracy would be very useful. Valid and precise assessments of intrinsic accuracy could help users to know how or when to use the systems and how much faith to put in them. Such assessments could also help system managers to determine when to attempt improvements and how to evaluate the results. A full evaluation of a system's performance would go beyond its general, inherent accuracy in order to establish quantitatively its utility or efficacy in any specific setting, but good, general measures of accuracy must precede specific considerations of efficacy (*1*).

I suggest that although an accuracy measure is often calculated in one or another inadequate or misleading way, a good way is available for general use. The preferred way quantifies accuracy independently of the relative frequencies of the events (conditions, objects) to be diagnosed ("disease" and "no disease" or "rain" and "no rain," for instance) and also independently of the diagnostic system's decision bias, that is, its particular tendency to choose one alternative over another (be it "disease" over "no disease," or vice versa). In so doing, the preferred measure is more valid and precise than the alternatives and can place all diagnostic systems on a common scale.

On the other hand, good test data can be very difficult to obtain. Thus, the "truth" against which diagnostic decisions are scored may be less than perfectly reliable, and the sample of test cases selected may not adequately represent the population to which the system is applied in practice. Such problems occur generally across diagnostic fields, but with more or less severity depending on the field. Hence our confidence in an assessment of accuracy can be higher in some fields than in others—higher, for instance, in weather forecasting than in polygraph lie detection.

The Appropriate Measure of Accuracy

Although some diagnoses are more complex, diagnostic systems over a wide range are called upon to discriminate between just two alternatives. They are on the lookout for some single, specified class of events (objects, conditions, and so forth) and seek to distinguish that class from all other events. Thus, a general theory of signal detection is germane to measuring diagnostic accuracy. A diagnostic system looks for a particular "signal," however defined, and attempts to ignore or reject other events, which are called "noise." The discrimination is not made perfectly because noise events may mimic signal events. Specifically, observations or samples of noise-alone events and of signal (or signal-plus-noise) events produce values of a decision variable that may be assumed to vary from one occasion to another, with overlapping distributions of the values associated with the two classes of events, and modern detection theory treats the problem as one of distinguishing between two statistical hypotheses (*2*).

The relevant performance data. With two alternative events and two corresponding diagnostic alternatives, the primary data are those of a two-by-two contingency table (Table 1). The event is considered to be "positive" or "negative" (where the signal event, even if undesirable, is called positive), and the diagnosis made is correspondingly positive or negative. So there are two ways in which the actual event and the diagnosis can agree, that is, two kinds of correct outcomes, called "true-positive" (cell *a* in Table 1) and "true-negative" (cell *d*). And there are two ways in which the actual event and the diagnosis can disagree, that is, two kinds of errors, called "false-positive" (cell *b*) and "false-negative" (cell *c*). Data from a test of a diagnostic system consist of the observed frequencies of those four possible outcomes.

The author is chief scientist at BBN Laboratories Incorporated, Cambridge, MA 02238. He is also lecturer on clinical epidemiology at Harvard Medical School.

Table 1. Two-by-two contingency table.

Diagnosis	Event: Positive	Event: Negative	
Positive	a True-positive	b False-positive	$a + b$
Negative	c False-negative	d True-negative	$c + d$
	$a + c$	$b + d$	$a + b + c + d = N$

However, if we consider proportions rather than raw frequencies of the four outcomes, then just two proportions contain all of the information about the observed outcomes. Take the symbols a, b, c, and d as denoting the actual numbers of each outcome that are observed, and certain ratios, such as $a/(a + c)$, as giving their proportions. Then, whenever the positive event occurs, the diagnosis is either positive or negative, and hence the false-negative proportion, $c/(a + c)$, is simply the complement of the true-positive proportion, $a/(a + c)$, with the two proportions in that column adding to one. Similarly, for the other column, whenever the negative event occurs, the diagnosis is either positive or negative, and so the true-negative and false-positive proportions are complements. Therefore, in a test of a diagnostic system, all of the relevant information with regard to accuracy can be captured by recording only one member of each of the complementary pairs of proportions (one proportion from each column). The usual choices are those of the top row, namely, the true-positive proportion and the false-positive proportion. The language of detection theory is often apt: those two proportions are of "hits" and "false alarms." Any operating system, unless perfect, will lead to false alarms as well as hits. Although other proportions can be drawn from the table, these two proportions are the major ones and the basis for an appropriate accuracy measure.

A measure independent of event frequencies. Converting raw frequencies to proportions in the way just described creates one of two fundamental attributes of a suitable accuracy measure. If it considers only the true-positive and false-positive proportions, an accuracy measure ignores the relative frequencies, or prior probabilities, of positive and negative events—defined, respectively, as $(a + c)/N$ and $(b + d)/N$, where N is the total number of events—and it does not depend on them. This is as it should be. For example, we do not want the accuracy score assigned a particular system for detecting cracks in metal to be specific to the relative frequencies of cracked and sound specimens chosen for the test sample.

A measure independent of the decision criterion. The second fundamental attribute of a suitable accuracy measure is that it be unaffected by the system's decision bias or tendency to favor one or the other diagnostic alternative. It is convenient to think of this bias or tendency as based on the criterion used by the system to establish a positive diagnosis. This decision criterion can be thought of as the critical, or threshold, amount of evidence favoring the occurrence of the positive event that is required to issue a positive diagnosis.

The decision criterion chosen by or for the system should (and usually does) depend on the prior probabilities of the two events. Thus, in situations in which the positive event has a high prior probability, the system should have a lenient criterion for a positive diagnosis. Consider the rain forecaster in Washington (merely a hint of rain leads to a positive prediction) or the mammographer examining a high-risk or symptomatic patient (the minimal suggestion of a lesion leads to further action). Then the quantity from Table 1 that reflects the entire positive row (not column), namely, $(a + b)/N$, will be high relative to its complement in the negative row, namely, $(c + d)/N$. Conversely, a strict criterion should be used when the positive event is unlikely on prior grounds. Then the positive row's probability will be lower relative to the negative row's.

The particular decision criterion that is appropriate depends also on the benefits ascribed to the correct outcomes and the costs ascribed to the incorrect outcomes. Predicting a severe storm that does not occur (a false positive) is typically regarded as having a cost that is small relative to the cost of failing to predict a storm that does occur (a false negative), so the criterion adopted for a positive diagnosis is on the lenient side. Conversely, a strict criterion would be set when the cost of a false-positive outcome is disproprotionately high; for example, the physician wants much to avoid life-threatening surgery on a patient who turns out not to have the suspected disease. Other examples exist in which one or another benefit is paramount (rather than costs as just illustrated) and hence has a major effect on the diagnostic criterion that is adopted.

When a positive diagnosis is made according to a lenient decision criterion, it will be made relatively often and both of the primary proportions in accuracy measurement, the true- and false-positive proportions, will be high. Conversely, positive diagnoses made according to a strict criterion will be made relatively infrequently, and both of these proportions will be low. A system of a fixed capacity to distinguish between positive and negative events cannot increase the true-positive proportion without also increasing the false-positive proportion. Nor can it decrease the false-positive proportion without also decreasing the true-positive proportion. A valid measure of accuracy will acknowledge that the true- and false-positive proportions will vary together, as the decision criterion changes. We desire a measure of accuracy that is valid for all the settings in which a system may operate, with any of the various decision criteria that may be appropriate for the various settings. And, within a single setting, we desire a measure of accuracy that is valid for the different decision criteria, appropriate or not, that may be set by different decision-makers. We must recognize that individuals can differ in their estimates of prior probabilities and of costs and benefits and so adopt different criteria.

Basis for calculating the suitable measure. A measure of accuracy that is independent both of the relative frequencies of the two events and of the decision criterion that is adopted for a positive diagnosis is defined in terms of the graph illustrated in Fig. 1. On this graph, one uses test data to plot the true-positive proportion against the false-positive proportion for various settings of the decision criterion. Thus, a curve on the graph shows the trading relation between true- and false-positive proportions that is characteristic of a particular system. One can see at a glance what proportion (or probability) of true positives the system will give for any particular proportion (or probability) of false positives, and vice versa. The idea then is to extract one number from a curve, which represents the entire curve, to provide a single-valued, general measure of accuracy.

Enough data points to define a curve reliably, say, five or more, are collected by either of two procedures. Under the binary or "yes-no" procedure, the system is induced to adopt a different decision criterion from one group of trials to another (*3*). Under the rating procedure, the system in effect reports which one of several different criteria is met on each trial. It does so by issuing either a rating of likelihood that a positive event occurred—for example, on a five-category scale ranging from "very likely" to "very unlikely"—or effectively a continuous quantity, for example, a probability estimate, that the analyst can convert to a rating. Then, in analysis, one

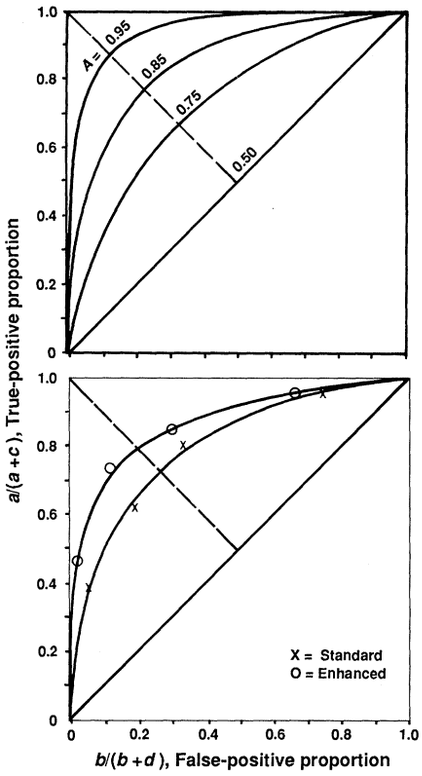

Fig. 1 (**top**). The ROC graph, in which the true-positive proportion is plotted against the false-positive proportion for various possible settings of the decision criterion. The idealized curves shown correspond to the indicated values of the accuracy measure A. **Fig. 2** (**bottom**). Example of empirical ROCs, showing standard and enhanced interpretations of mammograms.

considers different numbers of categories as representing a positive response (*4*). Most of the data reported in this article were obtained by the rating method.

A curve as shown in Fig. 1 is called an "ROC"—sometimes short for Receiver Operating Characteristic, especially in the field of signal detection, and sometimes short for Relative Operating Characteristic, in generalized applications. The history of the ROC and its extensive applications to discrimination tasks in experimental psychology, beginning in the early 1950s, is reviewed elsewhere (*5, 6*). It is well established that measured ROCs generally take a form much like that shown in Fig. 1 (*7*), and a detailed comparison of ROC measures of accuracy with other candidate measures is available (*8*). Methods for applying the ROC analysis to diagnostic systems have been described (*9, 10*).

The preferred measure of accuracy. A suitable, single-valued measure of accuracy is some measure of the locus of an ROC curve on its graph. The now preferred measure, specifically, is the proportion of the area of the entire graph that lies beneath the curve. The measure, denoted A in Fig. 1, is seen to vary from 0.50 to 1.0. Thus, $A = 0.50$ when no discrimination exists, that is, when the curve lies along the major diagonal (solid line), where the true- and false-positive proportions are equal. A system can achieve that performance by chance alone. And $A = 1.0$ for perfect discrimination, that is, when the curve follows the left and upper axes, such that the true-positive proportion is one (1.0) for all values of the false-positive proportion.

Three other, illustrative values of A are indicated by the three intermediate curves of Fig. 1. Observed curves usually differ slightly in shape from the idealized curves shown; they are typically not perfectly symmetrical but are a little higher on one side of the minor diagonal (dashed line) and a little lower on the other. However, the measure A suffices quite well as a single-valued measure of the locus of curves of the sort widely observed. Calculation of the measure can be accomplished graphically but is usually performed by a computer program that accepts as inputs the frequencies of positive and negative diagnoses for each alternative event that are observed for various criteria (*9–11*).

Illustrative Calculation of the Accuracy Measure

Techniques for obtaining an empirical ROC, making a maximum-likelihood fit of a smooth curve to its points, estimating A, estimating the variability in A, estimating components of variability in A due to case and observer sampling and observer inconsistency, and determining the statistical significance of a difference between two values of A are discussed elsewhere (*9, 10*). Here a brief illustration is given of the major aspects.

The data used for this purpose are taken from a study in which six general radiologists attempted to distinguish between malignant and benign lesions as viewed in a set of 118 mammograms (58 malignant, 60 benign), first when the mammograms were viewed in the usual manner and then when they were viewed with two aids. The radiologists came from community hospitals in the Cambridge area to BBN Laboratories where my colleagues and I had assembled an appropriate sample of mammograms from other area hospitals (*12*). The aids were (i) a checklist of perceptual features that are diagnostic of malignancy (obtained from specialists in mammography and confirmed by discriminant analysis), which elicited a scale value from the observer for each feature, and (ii) a computer system that merged the observer's scale values according to their optimal weights and estimated a probability of malignancy.

Each mammogram was rated on a five-category scale of likelihood that the lesion was malignant; the frequencies of the various ratings, as pooled over the six observers, are shown in columns 2 and 3 of Table 2. The procedure described above for converting rating data to the various pairs of true- and false-positive proportions that correspond to various decision criteria was used to generate the data in columns 4 and 5. The ROC points defined by those coordinate proportions are plotted in Fig. 2. (The points span the graph well

Table 2. Data table for illustrative ROC plot.

Rating category	Frequencies		Proportions	
	Malignant cases	Benign cases	$\frac{a}{a+c}$	$\frac{b}{b+d}$
	Standard viewing			
Very likely malignant	132	19	0.38	0.05
Probably malignant	85	50	0.62	0.19
Possibly malignant	63	48	0.80	0.33
Probably benign	53	151	0.96	0.74
Very likely benign	15	92	1.0	1.0
Sum	348	360		
	Enhanced viewing			
Very likely malignant	159	8	0.46	0.02
Probably malignant	102	38	0.75	0.13
Possibly malignant	36	62	0.85	0.30
Probably benign	34	131	0.95	0.66
Very likely benign	17	121	1.0	1.0
Sum	348	360		

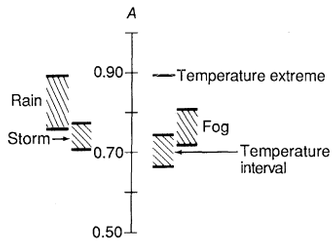

Fig. 3. Measured values of A for forecasts of several different weather conditions. Ranges are shown where multiple tests were made.

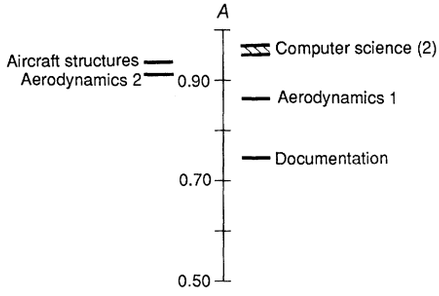

Fig. 4. Measured values of A for a manual information-retrieval system (on the left) and a computer-based information-retrieval system (on the right), for different collections of documents as indicated.

enough to avoid much extrapolation—a five-category rating scale, which yields four points within the graph, is usually adequate; other diagnostic fields often yield more data points, for instance, weather forecasting, where 13 rating categories are the norm, and aptitude testing or information retrieval, where the analyst can often derive a dozen or so ROC points from a nearly continuous decision variable.)

A nonparametric estimate of A can be obtained by connecting the successive ROC points for a given condition by lines and using the trapezoidal rule, or some related formula, to measure the area beneath the connected points (*5*). In general practice, however, empirical ROCs are plotted on other scales, namely, scales that are linear for the normal-deviate values that correspond to probabilities (where probabilities are inferred from observed proportions). A robust result across diagnostic fields is that empirical ROCs are fitted well by a straight line on such a "binormal" graph, as exemplified elsewhere for the fields treated here (*7*). A computer program gives the maximum-likelihood fit of the straight line (with parameters of slope and intercept) along with an estimate of A and its variance (*9, 10*). For the present purposes, the straight lines so fitted were transposed to the smooth curves on ordinary scales in Fig. 2. For the curves of Fig. 2 the χ^2 test of the goodness of fit yielded $P = 0.08$ for the standard viewing condition and $P = 0.29$ for the viewing condition enhanced by the aids, indicating satisfactory fits. The respective values of A for these curves (from pooled ratings) were 0.81 and 0.87 with standard errors of 0.017 and 0.014. (Average A values for curves fitted to individual observers are sometimes slightly different; here they were 0.83 and 0.88.) The difference between observing conditions (assessing the group values of A in either way) is significant by either a critical-ratio test or t test at $P = 0.02$ (*13, 14*).

Measured Accuracies of Some Common Systems

Let us consider now measured accuracies of some common systems in six diagnostic fields. With the exception of medical imaging systems, where upwards of 100 studies have reported ROCs, I include here all of the studies in each field I know that have used (or were later subjected to) ROC analysis and can be represented by a value of A.

Weather forecasting. ROCs for some 20 sets of weather data collected in the United States and Australia were calculated by Mason (*15*). The positive events consisted of rain, temperatures above or below a critical value or within a range, tornadoes, severe storms, and fog. I made estimates of A from his graphs. These estimates are based usually on hundreds, sometimes thousands, of forecasts, so the uncertainty in reported A values is 0.01 or so. Various values for weather events are summarized on the A scale in Fig. 3, showing ranges where available. The average or central values are approximately 0.89 for extreme cold; 0.82 for rain, 0.76 for fog, 0.74 for storms, and 0.71 for intervals of temperature (*16–18*).

Information retrieval. Major tests of information-retrieval systems at two locations were conducted in the mid-1960s (*19*), and their analysis in ROC terms was described shortly thereafter (*20*). The task of such a system is to find the articles and books that are relevant to each of a series of queries that are addressed to it, and to reject the irrelevant documents. In a traditional, manual library system, the queries will be in terms of some indexing language; in a computer-based system, they will contain some combination of key words.

Figure 4 summarizes the results obtained with a computer-based system at Harvard University by Gerard Salton and Michael Lesk (on the right) and results obtained with a manual library system at the Cranfield libraries in England by Cyril Cleverdon and Michael Keen (on the left). The computer-based system measured the degree of relevance of every document in the file to every query addressed to the file, and I established different decision criteria by selecting in turn various values along this nearly continuous variable. Various retrieval methods used synonyms, statistical word associations, hierarchical expansions, and so forth to relate the language of the document to the key words of the query. The collections of documents were in the subject matters of documentation, aerodynamics, and computer sciences. Under each method, a few hundred documents were examined in relation to each of about 20 to 40 queries. With almost no variation among methods, the collection of documents on documentation gave a typical value of A estimated to be 0.75; the aerodynamics collection, 0.87; and two computer-science collections, 0.95 and 0.97. For the manual system, I adopted different criteria by varying the number of query terms a document had to satisfy in order to be retrieved. In this test, 13 retrieval methods were variations of several indexing languages. They were applied with about 200 queries to a collection of some 1500 documents on aerodynamics. Again very consistently over methods, the approximate mean value of A was 0.91. Six retrieval methods applied to a smaller collection on aircraft structures yielded $A = 0.93$. For both of these tests in the library setting, the numbers of observations were large enough to yield very reliable values of A.

Aptitude testing. The validity of aptitude tests is usually measured by a correlation coefficient, because the event predicted, as well as the diagnostic system's output, is usually represented by a continuum of many values, rather than just two. These values are typically school grades or job ratings. However, the prediction of a two-valued event is often required, as when students under individually paced instruction either complete the course or not, or when job performance is measured simply as satisfactory or not. Another example comes from current interest in how much the Scholastic Aptitude Test helps, beyond knowing rank in high school class, in predicting college graduation. For such instances I suggest that the accuracy of prediction in ROC terms is the most appropriate measure of test validity. Figure 5 shows summary results of two studies of school performance on the A scale. Although nearly continuous grades in the course happen to be available here, I simulated a binary outcome of pass-fail by selecting arbitrarily a particular level of course performance as the cutoff for passing (*21*).

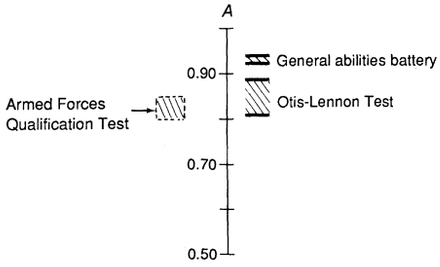

Fig. 5. Measured values of A for two aptitude tests (on the right) that were followed by schooling of all testees; a roughly adjusted range of A values for a test (on the left) that was followed by schooling only of those who achieved a criterion score on the test.

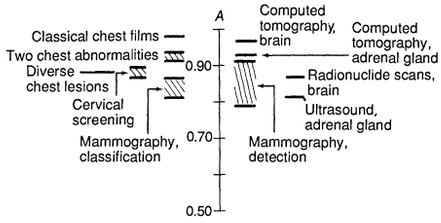

Fig. 6. Measured values of A for several imaging tests in clinical medicine.

The testees in the study shown on the right in Fig. 5 were approximately 450 students entering the seventh grade in four classes in each of three schools in Barquisimeto, Venezuela, all of whom would take a special course on thinking skills (*22*). The initial tests included the Otis-Lennon School Abilities Test (OLSAT) and an extensive battery of general aptitude tests (GAT). The year-end measure of performance came from an extensive test consisting of items specially written to tap the thinking abilities the course was intended to teach, called the Target Abilities Test (TAT). Actually, three values of A were computed for both OLSAT and GAT, one for each of three different percentile cuts on TAT performance that I took (arbitrarily) to define passing and failing. For GAT, the three corresponding A values ranged from 0.93 to 0.94 and for OLSAT they ranged from 0.81 to 0.89. The correlation coefficients are 0.87 for GAT and 0.71 for OLSAT. Other data so far analyzed in ROC terms are on the ability of the Armed Forces Qualification Test to predict pass-fail performance in four Navy schools (*7*) and are shown on the left in Fig. 5 (*23, 24*).

Medical imaging. Thirty published studies of medical imaging techniques or of image interpreters, in which the ROC was used for accuracy evaluation, were reviewed several years ago (*25*). There may now be as many as five times that number in the medical literature (*5, 10*), and so the summary here must be selective. I have tended to choose studies with enough clinical cases and image readers, and with enough points per ROC, to yield relatively reliable estimates of A. They focus on the abilities of computed tomography, mammography, and chest x-rays to discriminate lesions indicative of disease from normal variations of organ appearance. Values of A are summarized in Fig. 6.

In an early study of computed tomography (CT) relative to radionuclide scanning (RN) for detecting brain lesions, images of both modalities were obtained from the same patients. With the sponsorship of the National Cancer Institute, images of 136 cases were selected from about 3000 cases collected at five medical centers. These images were interpreted retrospectively by six specialists in each modality who gave ratings of likelihood that a lesion was present (*26*). About 60% of the cases were abnormal as proven by histology; the normal cases showed no symptoms after 8 months of follow-up. Both the pooled and average A values across observers were 0.97 for CT and 0.87 for RN. A study comparing CT to ultrasound in detecting adrenal disease (for the most part, with both examinations made of the same patients) was based on cases at two medical centers and on interpretations made in the course of diagnosis. Values of A were 0.93 for CT and 0.81 for ultrasound (*27–35*).

Materials testing. "Materials testing" here means testing metal structures, such as aircraft wings, for cracks. There is one major study in the field, in which a set of 148 metal specimens, each regarded to be with or without cracks, was tested at 16 bases of the U.S. Air Force. The diagnostic systems consisted of ultrasound and eddy current devices used by upwards of 100 technicians in two separate tests (*36*).

Because the technicians made only binary decisions, without manipulation of their diagnostic criteria, just one point on each individual's ROC is available. To calculate A, I assumed that that point lay on a symmetrical ROC, as shown in Fig. 1 (not a crucial assumption here). The average A values across sites are 0.93 for the eddy-current technique and 0.68 for the ultrasound technique, but accuracy varied widely from one base to another, across the ranges shown in Fig. 7. Indeed, the extent of the range may be the salient result: a case could be made for analyzing the expertise at the more proficient sites in order to export it to the less proficient.

Polygraph lie detection. Studies of polygraph accuracy in lie detection are of two types. In so-called "field" studies, for various real crimes, the polygraph examiners' decisions about deception or truth are compared either to actual judicial outcomes, panel decisions about guilt, or confessions. So-called "analog" studies are of mock or role-playing crimes in a laboratory setting, for example, stealing a $20 bill from an office down the hall. The obvious differences between the two types concern the surety of the "ground truth" about the positive event of guilty and the negative event of not guilty, and the severity of the consequences of failing the test.

Figure 8 shows summary values for both types of study. About ten published studies exist in each category. Most of the field studies were reviewed in the context of signal detection theory and ROC analysis (*37*), and both categories were reviewed for the U.S. Congress Office of Technology Assessment (*38*). I have calculated values of A, again under the assumption that the single ROC points available lie on a symmetric curve as shown in Fig. 1. (The possible impact of that assumption is lessened by the fact that the studies generally gave points near the negative, dashed-line, diagonal of the ROC graph.) Of the field studies, four gave A values near 0.95, with one as high as 0.98; these were conducted by a commercial agency. Five field studies conducted in university settings gave A values between 0.70 and 0.92. Nine of eleven analog studies produced A values ranging from 0.81 to 0.95; A values for the two outlying analog studies were 0.64 and 0.98. One of the analog studies used the rating-of-likelihood procedure to yield full ROCs, based on six ROC points. In this study, six examiners yielded A values ranging from 0.55 to 0.75; four of them were between 0.64 and 0.68 (*39*).

Qualifications of Measured Accuracies

Most, if not all, of the values of A listed for the various fields should be qualified in one or more ways. Certain definitions, operations, and assumptions made in conducting the tests have served to bias the calculated values—often in unknown directions to unknown extents, sometimes in known directions, and, infrequently, to an extent that may be estimated. So, calculated values of A are neither perfectly reliable, in the sense of being repeatable across different tests of the same system, nor perfectly valid, in the sense of measuring what they are supposed to measure. There is constant, as well as variable, error. The difficulty lies not, I have argued, with the measure: as far as we can tell, there are no intrinsic limits on its reliability, beyond ordinary statistical considerations, or on its

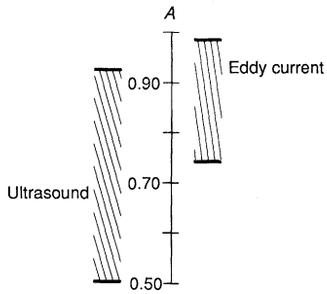

Fig. 7. Measured values of A for detecting cracks in airplane wings by two techniques, from several Air Force bases.

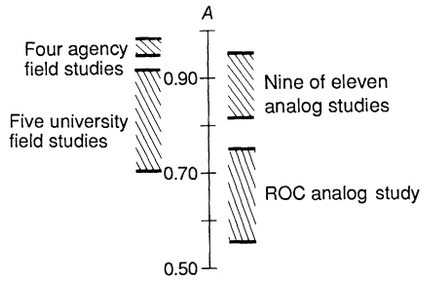

Fig. 8. Measured values of A for polygraph lie detection in several field studies (on the left) and several analog studies (on the right).

validity. The difficulty arises, rather, because the quality of test data is not everything it should be. I consider here four ideals for test data and comment on how well each of the diagnostic fields treated here matches up to these ideals.

Adequacy of truth. The tester should know certainly for every item in the test sample whether it is positive or negative. Incorrectly classifying test items will probably depress measures of accuracy.

How are truly guilty and truly innocent parties to be determined for tests of the polygraph? Judicial outcomes and panel decisions may categorize erroneously, and even confessions can be false. Hence, one may resort to the analog study, which sacrifices realism to gain sure truth.

Sure truth about cracks in metals can only be obtained destructively, by sacrificing the specimens. Destructive testing tends not to be done, because then the next diagnostic technique, or the next group of inspectors, must be tested on another, different set. A set of specimens for which fairly good truth is felt to be available is acquired only painstakingly, and a new set will not be a common ground for comparing a potential diagnostic technique with existing ones, or new inspectors with old. Just how truth is determined in this field, short of sacrifice, is not clear to me. I believe that it is based on the same diagnostic techniques one hopes to test, perhaps in the hands of experts and in combination.

The so-called "gold standard" for truth in medical imaging is usually regarded to be surgery or autopsy and analysis of tissue. It is recognized, however, that imagery and pathology are not perfectly correlated in space or time. The image interpreter and pathologist may look at different locations, and the pathological abnormality observed may not have been present when the image was taken. Moreover, the pathologist's code or language for describing lesions differs from the radiologist's. Of course, this histology standard is applied primarily to positive cases. Negative truth is usually based on months or years of follow-up without related symptoms.

For assessments of aptitude testing in terms of the measure A I think the problems of truth data are slight. If handing in the assigned work means passing, then we know who passed and who failed; if staying on the job constitutes success, likewise. We may assume that errors in determining who did what are infrequent.

The definition of a document's relevance to a query—or, in this context, what should constitute truth—has had a controversial history in the field of information retrieval. In the studies reviewed here, the relevance of every document in the file was judged by subject-matter experts for each query. In some instances, the degree of relevance was estimated on a four-category scale. Other studies have drawn queries directly from documents in the file, a procedure that better defines those documents as relevant than it does all others as irrelevant. In any event, the dependence of truth on judgment suggests that it will be more adequate for some subject matters, probably those with a highly technical language, than for others.

Problems in assessing truth in weather-forecasting arise primarily from logistic limitations on establishing in a fine-grained manner whether a weather event occurred throughout the area of the forecast. One knows rather surely how many millimeters of rain there are in a can at the airport, but the forecast is often made for a larger area. Similarly, tornadoes may touch down, or storms may be severe, in unobserved places. In short, it is difficult to correlate the forecast and the truth determination in space. The correlation of forecast and truth determination in time is not simple either but seems easier.

Independence of truth determination and system operation. The truth about sample items should be determined without regard to the system's operation, that is, without regard to the system's decisions about test cases. If this condition is not met, the truth will be inappropriate for scoring the system and will probably inflate its measured accuracy.

When confessions are used to determine guilt and innocence, the likelihood of a confession depends on whether the polygraph test is judged to be positive or negative. Examiners work hard to elicit a confession from suspects who appear to test positively and believe that the existence of a positive test is often the main factor in securing a confession. (Hence, they can argue that the system's efficacy is high even if its accuracy is low.) The result for accuracy measurement is that the system is scored against a determination of truth that it helped to make. That test procedure treats the polygraph system very generously—it ought to do well. Values of A will be inflated, to an unknown, but conceivably large, extent.

If panel decisions based on all available evidence are used to establish truth in materials testing, then truth is determined in part by the operation of the system or systems under test.

In good practice in medical imaging, the truth is determined independently of system operation. Occasionally, truth is determined by all of the case evidence, including the results of the systems under test. That practice can favor CT, say, over the alternative, if the CT result is dominant in calling a case positive or negative. CT is then scored against itself.

Independence of test sample and truth determination. Procedures used to establish the truth should not affect the selection of cases. Thus, the quest for adequate truth may bias the sample of test cases, perhaps resulting in an easier sample than is realistic.

Many criminal investigations do not result in a confession. When confession is the sole basis for determining truth and hence dictates the sample, the sample will probably not represent the population of cases to which the polygraph is typically applied. As one specific, it is possible that the more positive a test appears to be, the greater the likelihood of a confession. So the sample will tend to consist of the easier cases to diagnose. Again, the possibility exists of substantial inflation of measured accuracy.

In materials testing, the use of panel decisions based on all available evidence would serve to condition the constitution of the sample by the procedure for determining truth.

In medical imaging, potential biases in the sample may result from the procedures for establishing truth. If tissue analysis is the standard, the sample will be made up of cases that achieve that advanced stage, quite possibly cases that show relatively clear

lesions. For negative cases, a sample may reflect the population more or less well, depending on how long one waits for follow-up. A formula for eliminating these biases was recently proposed (*40*).

A problem for aptitude testing arises from the fact that testing is carried out to make selections, and ground truth is available only for those selected. How the persons scoring below the selection criterion on the aptitude test would have performed in school or on the job is usually not known. Procedures used to establish truth—observing and grading the individual in school or on the job—determine the sample completely. The sample for assessing the diagnostic system is biased relative to the population to which the system is applied.

Representativeness of the sample. The sample of test items should fairly reflect the population of cases to which the diagnostic system is usually applied. The various types of events should occur in the sample in approximately the same proportions as they do in practice.

A representative of criminal cases could, in principle, be obtained prospectively. If all criminal cases in the country were collected for a sufficiently long time starting now, then the various types of crimes against objects and people would appear in appropriate numbers. But that would be a very long time from the standpoint of someone desirous of an accuracy measure soon. Selection of criminal cases retrospectively in the appropriate proportions would depend on a common and adequate coding of case types across the country, or a central file, and an ability to acquire full records for cases in which the polygraph was used. A reasonable deduction from sample sizes in the polygraph assessment literature, ranging typically from about 20 to 50 in field studies, is that sample items are very difficult to acquire. This is an instance of a potential bias in the accuracy measure of an unknown direction, let alone extent (*41*).

For reasons given earlier, it would seem to be difficult in materials testing to specify a representative sample of types and sizes of cracks and to ensure that one exists.

The problem in medical imaging of achieving representative samples with respect to type and extent of lesion mirrors the problem for criminal cases. Prospective sampling is expensive and time-consuming. Indeed, a new and advanced model of the imaging device may be available before enough cases are collected with the tested one. Retrospective sampling requires first that the data be accessible, and so far they usually have not been. Such sampling also requires great care. For example, rare cases must be present in at least minimal numbers to represent the rarity fairly, and having that number of them may distort the relative proportions.

In information retrieval, it is difficult to say whether a representative sample of documents is acquired for a general assessment of a system. Working with special subject matters seems appropriate for a given test, but most systems, as illustrated earlier, are tested with just a few of them. Across the few mentioned above, accuracy varies considerably and seems to covary with the "hardness," or technical nature, of the language used for the particular subject matter.

The ability of weather forecasters to assemble large and representative samples for certain weather events is outstanding. Prediction of precipitation at Chicago was tested against 17,000 instances, and even individual forecasters were measured on 3,000 instances. Of course, some weather events are so rare that few positive events are on record, and for such events the precision as well as the generality of the measurements will be low (*42*).

Concluding Remarks

Can we say how accurate our diagnostic systems are? According to the evidence collected here, the answer is a quite confident "yes" in the fields of medical imaging, information retrieval, and weather forecasting, and, at least for now, a "not very well" in most if not all other fields, as exemplified here by polygraph lie detection, materials testing, and (except for the few analyses mentioned above) aptitude testing for predicting a binary event. ROC measures of accuracy are widely used in medical imaging (*5, 10, 24*), have been advocated and refined within the field of information retrieval (*20, 43*), and have been effectively introduced in weather forecasting (*15, 17, 18, 44*). Although problems of bias in test data do not loom as large in information retrieval and weather forecasting as elsewhere, those fields have shown a high degree of sophisticated concern for such problems, as has medical imaging, where the problems are greater (*45*). So, in medical imaging we can be quite confident for example, about A values of 0.90 to 0.98 for prominent applications of CT and chest x-ray films and A values of 0.80 to 0.90 for mammography. Similarly, in weather forecasting, confident about A values of 0.75 to 0.90 for rain, depending largely on lead time, and of 0.65 to 0.80, depending on definitions, for temperature intervals and fog; and in information retrieval, A values ranging from 0.95 to 0.75 depending on subject matter. A positive aspect of the field of polygraph lie detection is that it recognizes the need for accuracy testing and attempts to identify and cope with inherently difficult data-bias problems, and the field of materials testing is making some beginnings in these respects. Of course, for other than the special case considered here, the field of aptitude testing devotes a good deal of sophisticated effort to validity questions.

What will the future bring? A basic assumption of this article is that testing the accuracy of diagnostic systems is often desirable and feasible and is sometimes crucial. Although individual diagnosticians are treated here only in passing, a similar case could be made for the importance of testing them. I suggest that a wider and deeper understanding of the needs and the possibilities would be beneficial in science, technology, and society, and that it is appropriate for scientists to take the lead in enhancing that understanding. Scientists might help society overcome the resistance to careful evaluation that is often shown by diagnosticians and by designers and managers of diagnostic systems, and help to elevate the national priority given to funding for evaluation efforts. Specifically, I submit that scientists can increase general awareness that the fundamental factors in accuracy testing are the same across diagnostic fields and that a successful science of accuracy testing exists. Instead of making isolated attempts to develop methods of testing for their own fields, evaluators could adapt the proven methods to specific purposes and contribute mutually to their general refinement.

REFERENCES AND NOTES

1. The measurement of efficacy in the context of the present approach to accuracy is treated in some detail elsewhere (*9, 10*). The usefulness of empirical measures of diagnostic, and especially predictive, accuracy was further set in a societal context in a recent editorial: D. E. Koshland, Jr., *Science* **238**, 727 (1987).
2. W. W. Peterson, T. G. Birdsall, W. C. Fox, *IRE Trans. Prof. Group Inf. Theory* **PGIT-4**, 171 (1954).
3. With a human decision-maker, one can simply give instructions to use a more or less strict criterion for each group of trials. Alternatively, one can induce a change in the criterion by changing the prior probabilities of the two events or the pattern of costs and benefits associated with the four decision outcomes. If, on the other hand, the decision depends on the continuous output of some device, say, the intraocular pressure measured in a screening examination for glaucoma, then, in different groups of trials, one simply takes successively different values along the numerical (pressure) continuum as partitioning it into two regions of values that lead to positive and negative decisions, respectively. This example of a continuous output of a system suggests the alternative to the binary procedure, namely, the so-called "rating" procedure.
4. Thus, to represent the strictest criterion, one takes only the trials given the highest category rating and calculates the relevant proportions from them. For the next strictest criterion, the trials taken are those given either the highest or the next highest category rating—and so on to what amounts to a very lenient criterion for a

positive response. The general idea is illustrated by probabilistic predictions of rain: first, estimates of 80% or higher may be taken as positive decisions, then estimates of 60% or higher, and so on, until the pairs of true- and false-positive proportions are obtained for each of several decision criteria.

5. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966; reprinted with updated topical bibliographies by Krieger, New York, 1974, and Peninsula Publishing, Los Altos, CA, in press).
6. J. A. Swets, *Science* **134**, 168 (1961); *ibid.* **182**, 990 (1973).
7. ———, *Psychol. Bull.* **99**, 181 (1986).
8. ———, *ibid.*, p. 100.
9. ——— and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory* (Academic Press, New York, 1982).
10. C. E. Metz, *Invest. Radiol.* **21**, 720 (1986).

11. Just how much accuracy a particular value of A represents takes on some intuitive meaning if one examines the various combinations of true- and false-positive proportions it represents. Values of A between 0.50 and 0.70 or so represent a rather low accuracy–the true-positive proportion is not much greater than the false-positive proportion anywhere along the curve. Values of A between about 0.70 and 0.90 represent accuracies that are useful for some purposes, and higher values represent a rather high accuracy. Further intuitive meaning of any values of A arises from the fact that it can be viewed as the percentage of correct decisions in a "paired comparisons" task. In this task, a positive event and a negative event are always presented on each trial (side by side or one after the other), and the diagnostic system must say which is which. Although "percent correct" is usually a misleading measure for diagnostic tasks, it is a useful measure for the paired-comparisons task because here relative frequencies are not at issue and the decision criterion can be expected to be symmetrical rather than biased one way or the other (toward the left or right side or toward the first or second time interval). The value of A obtained in the way described above can be shown, in general, to equal the percentage correct in this special task (*5*). Thus, $A = 0.82$ for rain predictions means that if the forecaster were presented a pair of randomly sampled weather conditions on successive trials—always with one condition that led to rain and one that did not—the forecaster could say which is which 82% of the time.

    Because the slope of the ROC decreases smoothly, the numerical value of the slope at any point on the ROC is an index of the decision criterion that yields that point. Various definitions of optimal decision rules specify the optimal criterion for any specific situation (as characterized, for example, by particular prior probabilities and benefits and costs) in terms of its corresponding ROC slope (*5, 9*).

12. D. J. Getty, R. M. Pickett, C. J. D'Orsi, J. A. Swets, *Invest. Radiol.* **23**, 240 (1988).

13. A straight line on a binormal graph is predicted from theory when the distributions of samples or observations of the two diagnostic alternatives (for example, signal and noise) are Gaussian. However, almost any form of distribution, including logistic, triangular, and exponential, yields very nearly a straight line on a binormal graph, and, in particular, a form that is indistinguishable from a straight line with ordinary amounts of data (*8*). Observed slopes of the linear ROC range for the most part from about 0.70 to 1.0 (*7*). A slope of 1.0 corresponds to an ROC on ordinary scales that is symmetrical about the minor diagonal (as in Fig. 1), and slopes of less than 1.0 correspond to ROCs on ordinary scales that rise more steeply from the origin than does a symmetrical curve (*5*), as do the ROCs in Fig. 2 (whose slopes on a binormal graph are 0.92 and 0.71).

    The measure here denoted A (for convenience) is denoted A_z in the primary literature, where the subscript z serves as a reminder that the measure was obtained from a binormal graph. A_z can be estimated from the two parameters of the linear ROC by the formula $z(A) = a/(1 + b^2)^{1/2}$, where b is the slope and a is the intercept of the linear ROC, and $z(A)$ is the normal-deviate value of the cumulative standardized normal distribution that corresponds to a tabled area beneath the normal distribution equal to A_z (*9*). For some crude measurement purposes, one can estimate A from a single ROC point by assuming that the linear ROC has a slope of 1.0, or, equivalently, that the ROC on ordinary scales is symmetrical about the negative diagonal. A simple procedure is to use a pair of tables published elsewhere (*14*): one can take the true- and false-positive proportions to one table to obtain a quantity called d' (a measure of accuracy based on a linear binormal ROC of unit slope) and take that quantity to the second table to obtain A (there presented as the percentage of correct choices in a paired-comparisons task, as mentioned earlier).

    The advantage of ROC analysis and A over other common measures of accuracy can be partially illustrated by the data of Fig. 2. Clearly, the true-positive proportion alone is inadequate as a measure: for example, for the standard condition it varies over the four observed data points from 0.38 to 0.96 although those points represent changes only in the decision criterion while accuracy remains constant. Contrary to fact, the standard condition could appear more accurate than the enhanced condition, if, for example, the third-left point were obtained in the standard condition yielding a value of 0.80 and the second-left point were obtained in the enhanced condition yielding a value of 0.75. The overall percentage of correct detections, $(a + d)/N$ in Table 1, also varies across the four points: if the relative frequencies of noise and signal trials are equal, then this measure gives 0.67, 0.72, 0.74, and 0.61 for the four points of the standard condition, and 0.72, 0.81, 0.78, and 0.65 for the enhanced condition. Were the second-left point on each curve the only one obtained, the difference between conditions would be seen as 0.09, whereas, if the third-left point were the only one obtained, the difference would be 0.04, so that the difference between conditions might be interpreted to be statistically significant in one case and insignificant in the other. Meanwhile, had the observed points been based on relative frequencies of 75 signal trials and 25 noise trials, the standard condition would yield percentages of 0.52, 0.67, 0.77, and 0.79 for the four points and the enhanced condition would yield 0.59, 0.78, 0.81, and 0.80, values quite different from those predicated on equal frequencies and either higher or lower depending on the decision criterion.

    Taking both coordinate proportions from a single observed point (or from a single two-by-two contingency table) is also a common and usually inadequate practice: in medicine, the true-positive proportion is called "sensitivity" and the true-negative proportion (the complement of the false-positive proportion) is called "specificity." Such an approach is adequate only when the diagnostic tests in question have an agreed-upon and invariant decision criterion, such as a test that yields a number along a scale. A difficulty for tests that do not (for example, imaging systems with human interpreters) is that differences are often indeterminate; for example, in Fig. 2, the first-left point of the standard condition could turn out to be on the same ROC as the second-left point of the enhanced condition if the full curves were obtained, so with a single point per condition one might not know if the conditions were equally accurate or not. Taking the inverses of the two proportions just mentioned—that is, the probability of the signal event (or the noise event) given the positive (or negative) decision, which is called the positive (or negative) "predictive value" of the system—is similarly problematic: these values vary over a wide range depending both on the decision criterion and on the relative frequencies of the two kinds of events. Indeed, the positive predictive value varies from the prior probability (relative frequency) of the signal event (which may be low) when the decision criterion is lenient on up to 1.0 when the criterion is strict (*9*). Similar problems plague the measure called "relative risk" or "odds ratio," which is a ratio of two inverse probabilities (*8, 9*). In general, single-valued measures taken from a single point (or two-by-two table), such as the odds ratio and percentage correct and various measures of association in statistics, can be shown to predict ROCs of too restricted a form (for example, a binormal slope of 1.0) or of a form very different from those observed in actual practice (*8*).

14. J. A. Swets, Ed., *Signal Detection and Recognition by Human Observers* (Wiley, New York, 1964; reprinted by Peninsula Publishing, Los Altos, CA, in press).
15. I. Mason, *Aust. Meteorol. Mag.* **30**, 291 (1982).

16. Several tests of rain are included, in which the likelihood of rain during the daytime hours was forecast the evening before. Their values of A are generally: 0.85 to 0.90 at the Canberra (Australia) airport, when any measurable trace of rain was defined as rain; 0.89 there when rain was defined as ≥2.5 mm; 0.83 at Seattle; 0.85 at Chicago; and 0.89 at Great Falls, Montana. When accuracy was measured for a fairly large area (for example, at any of 25 official rain gauges in the city of Canberra), $A = 0.74$. Recently published data on rain forecasts from the Environmental Research Laboratory at Boulder, Colorado, were based on events at different periods after issue of the forecast. The lead times and corresponding A values were: 0 to 2 hours, 0.86; 2 to 4 hours, 0.77; 4 to 6 hours, 0.75; and 6 to 8 hours, 0.75 (*17*). Similar findings were reported separately by two laboratories of Canada's Atmospheric Enviroment Service, showing monotonically decreasing values of A for two major prediction systems over six periods extending to 72 hours (*18*). Such details are also available for predictions of the other weather events.

17. G. M. Williams, in *Ninth Conference on Probability and Statistics in Atmospheric Sciences* (American Meteorological Society, Boston, 1985), p. 214.
18. N. Brunet, R. Verret, N. Yacowar, in *Tenth Conference on Probability and Statistics in Atmospheric Sciences* (American Meteorological Society, Boston, 1987); p. 12; R. Sarrazin and L. J. Wilson, *ibid.*, p. 95.
19. C. Cleverdon, *Assoc. Spec. Libr. Inf. Bur. Proc.* **19**, 6 (1967); ——— and M. Keen, *Factors Determining the Performance of Indexing Systems: Test Results* (Association of Special Libraries and Information Bureaux, Cranfield, England, 1966), vol. 2; G. Salton and M. E. Lesk, *J. Assoc. Comput. Mach.* **15**, 8 (1968).
20. J. A. Swets, *Am. Doc.* **20**, 72 (1969).

21. A problem faced in validating aptitude tests is that measures of school or job performance usually exist for only a subset of those tested, namely, those selected to matriculate or work. To alleviate this problem, the correlation coefficient is usually corrected (for example, adjusted upward), according to an accepted formula, for the restriction in the range of aptitudes considered. Similarly, one might devise a way to adjust values of A upward, because the discrimination task is made more difficult by considering only the more homogeneous group of testees who are selected for the school or job. I avoid the problem here, for the first study to be described, by analyzing an unrestricted set of data: all testees took the course and were graded in it.

22. R. J. Herrnstein, R. S. Nickerson, M. de Sánchez, J. A. Swets, *Am. Psychol.* **41**, 1279 (1986).

23. The approximate average value for the Navy data is $A = 0.70$, but that figure is depressed by the restriction of range (*21*). The correlations for the four schools are 0.27, 0.09, 0.31, and 0.35. I have not calculated their corrected values, but in another Navy study median uncorrected and corrected coefficients were 0.43 and 0.73 (*24*). As a very rough guess, applying some appropriate adjustment to the average $A = 0.70$ of the present data might bring it into the range 0.80 to 0.85, more like the unrestricted A values reported above. The Navy data are placed tentatively at that level (broken lines) on the summary scale of Fig. 5.

24. L. Swanson, *Armed Services Vocational Aptitude Battery, Forms 6 and 7, Validation Against School Performance in Navy Enlisted Schools* (Technical Report 80-1, Navy Personnel Research and Development Center, San Diego, 1979).
25. J. A. Swets, *Invest. Radiol.* **14**, 109 (1979).
26. J. A. Swets *et al.*, *Science* **205**, 753 (1979).
27. H. L. Abrams *et al.*, *Radiology* **142**, 121 (1982).

28. Three mammography studies treated the search for and detection of lesions then judged to be malignant. Egan's pioneering book (*29*) reported data I translate into $A = 0.91$. The remaining two studies were based on cases obtained in a national screening program during the past 10 years: one showed $A = 0.85$ for mammography alone, $A = 0.73$ for physical examination alone, and $A = 0.91$ for the two examinations combined (*30*); the other examined only "incidence" lesions (the generally smaller lesions that appeared only after the initial screening) and gave $A = 0.79$ (*31*). The study mentioned earlier of the benign-malignant classification of lesions in specified locations (*12*) is also represented in Fig. 6. A study comparing automatic computer classification of photomicrographs of cervical cells

with the performance of ten cytotechnologists found that the distinction between abnormal and normal cells was made with $A = 0.90$ by the computer and $A = 0.87$ by the cytotechnologists (*32*). The classical studies of "observer error" in radiology were made for chest films in the diagnosis of tuberculosis, and were conducted by Garland and by Yerushalmy and associates in the late 1940s [L. H. Garland, *Radiology* **52**, 309 (1949); J. Yerushalmy, J. T. Harkness, J. H. Cope, B. R. Kennedy, *Am. Rev. Tuberculosis* **61**, 443 (1950)]. They were reanalyzed in ROC terms by Lusted in his book (*33*), which introduced students of medical decision-making to the ROC. Although the data points from the two studies were in different portions of the ROC graph, for both studies $A = 0.98$. (Meanwhile, values of "percentage correct" for the two studies are 83 and 91%, indicating, spuriously, a difference in accuracy.) One recent study of chest films (*34*) gave A values of 0.94 and 0.91 for the detection of two specific abnormalities (pneumothoraces and interstitial infiltrates). Another, with diverse lesions, showed $A = 0.79$ when the observer was not prompted by the case history and 0.88 when he or she was (*35*).

29. R. L. Egan, *Mammography* (Thomas, Springfield, IL, 1964).
30. J. E. Gohagan *et al.*, *Invest. Radiol.* **19**, 587 (1984).
31. J. E. Goin, J. D. Haberman, M. K. Linder, P. A. Lambird, *Radiology* **148**, 393 (1983).
32. J. W. Bacus, *Application of Digital Image Processing Techniques to Cytology Automation* (Technical Report, Rush Presbyterian–St. Luke's Medical Center, Medical Automation Research Unit, Chicago, 1982). Some results of this study were reported by J. W. Bacus *et al.*, *Anal. Quant. Cytol. Histol.* **6**, 121 (1984).
33. L. B. Lusted, *Introduction to Medical Decision Making* (Thomas, Springfield, IL, 1968).
34. H. MacMahon *et al.*, *Radiology* **158**, 21 (1986).
35. K. S. Berbaum *et al.*, *Invest. Radiol.* **21**, 532 (1986).
36. J. A. Swets, *Mater. Eval.* **41**, 1294 (1983).
37. G. Ben-Shakhar, I. Lieblich, Y. Bar-Hillel, *J. Appl. Psychol.* **67**, 701 (1986).
38. L. Saxe, D. Dougherty, B. Scott, *Am. Psychol.* **40**, 794 (1985).
39. J. J. Szucko and B. Kleinmuntz, *ibid.* **36**, 488 (1981).
40. R. Gray, C. B. Begg, R. A. Greenes, *Med. Decis. Making* **4**, 151 (1984).
41. A review of the polygraph assessment literature, which treats the problems mentioned here and some more specific, concludes that measured accuracy values are likely to be inflated by a wide margin (*37*). That inflation is even more marked when a measurement of "percentage of correct decisions" is made rather than a measurement of A. That percentage measure sets aside cases having polygraph records that the examiner deems inconclusive—in effect, cases on which he or she chooses not to be scored.
42. I should observe in passing that accuracy measures may be biased by certain general factors other than the four treated here. The one that comes first to mind is that system tests are often conducted under laboratory, rather than field, conditions. Thus, radiologists may participate in laboratory tests under unusually good conditions: quiet, no interruptions, and no patient treatment depending on the diagnosis. Such a test may be conducted purposely, in order to test a system in a standard way and at its full potential, but the measured value may be higher than can be expected in practice.
43. Since the concern for evaluation in the field of information retrieval heightened in the 1950s, this field has used several different measures of accuracy intended to be independent of the decision criterion; see, for example, a review by J. A. Swets, *Science* **141**, 245 (1963). ROC measures have been considered extensively, for example, by B. C. Brookes, *J. Doc.* **24**, 41 (1968); M. H. Heine, *ibid.* **29**, 81 (1973); A. Bookstein, *Recall Precision* **30**, 374 (1974); M. H. Heine, *J. Doc.* **31**, 283 (1975); M. Kochen, Ed., *The Growth of Knowledge* (Wiley, New York, 1967); T. Saracevic, Ed., *Introduction to Information Science* (Bowker, New York, 1970); B. Griffith, Ed., *Key Papers in Information Science* (Knowledge Industry Publications, White Plains, NY, 1980).
44. Various scores for binary forecasts that were developed in this field were related to ROC measures by I. Mason, in *Ninth Conference on Weather Forecasting and Analysis* (American Meteorological Society, Boston, 1982), p. 169; M. C. McCoy, in *Ninth Conference on Probability and Statistics in Atmospheric Sciences* (American Meteorological Society, Boston, 1985), p. 423; *Bull. Am. Meteorol. Soc.* **67**, 155 (1986).
45. D. F. Ransohoff and A. Feinstein, *New Engl. J. Med.* **299**, 926 (1978); G. Revesz, H. L. Kundel, M. Bonitabus, *Invest. Radiol.* **18**, 194 (1983). Recommendations for handling various bias problems are advanced by C. B. Begg and B. J. McNeil, *Radiology* **167**, 565 (1988).
46. This article was prepared during a sabbatical leave granted under the Science Development Program of BBN Laboratories Incorporated.

The El Niño Cycle: A Natural Oscillator of the Pacific Ocean–Atmosphere System

NICHOLAS E. GRAHAM AND WARREN B. WHITE

Research conducted during the past decade has led to an understanding of many of the mechanisms responsible for the oceanic and atmospheric variability associated with the El Niño–Southern Oscillation (ENSO). However, the reason for one of the fundamental characteristics of this phenomena, its quasi-periodicity, has remained unclear. Recently available evidence from a number of sources now suggests that the ENSO "cycle" operates as a natural oscillator based on relatively simple couplings between the tropical atmospheric circulation, the dynamics of the warm upper layer of the tropical ocean, and sea surface temperatures in the eastern equatorial Pacific. This concept and recent field evidence supporting the natural coupled oscillator hypothesis are outlined.

ONE OF THE FUNDAMENTAL ASPECTS OF THE EL NIÑO–Southern Oscillation (ENSO) phenomena is its quasi-periodicity that is marked by the repeated appearance of warm or cool water in the equatorial eastern and central Pacific Ocean at intervals of 3 to 5 years. There is now evidence that this quasi-cyclic behavior is due to the operation of a natural coupled oscillator of the ocean-atmosphere system in the tropical Pacific. In the context of this concept, warm and cool water episodes are considered as phases of a self-sustaining cycle; this contrasts with the more traditional view of El Niños as discreet warm events superimposed on a mean background state. This cycle is maintained by interactions, both immediate and delayed as well as local and remote, between three fields in the tropical Pacific; sea surface temperature (SST), surface wind, and the thickness of the warm upper layer of the ocean. Representative examples of time series from each of these fields are presented in Fig. 1, which shows departures from the long-term monthly means of eastern Pacific SST (*1*), zonal (east-west) wind in the central equatorial Pacific (*2*), and sea level at Truk Island in the western tropical Pacific (*3*) for the period from 1950 through the mid-1980s (*4*). How these curves

The authors are at the Scripps Institution of Oceanography, La Jolla, CA.