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.