LETTERS

Environmental Analysis

In reading articles on environmental issues, I have repeatedly felt that three areas receive insufficient attention.

1) *Extrapolation.* "Everyone knows" that extrapolation is statistically unsound and that curves apply only to regions validated by data. In environmental issues extrapolation is routinely used, since no data exist in some important areas. The fallacy becomes obvious in looking at the case of trace elements in the human body. Copper and zinc, for example, are absolutely necessary to life. In larger doses they are dangerous poisons. No extrapolation could predict this reversal at low dosage. We need data, not bad guesses.

2) *Analysis.* Environmental risk assessment is a cost-benefit analysis, but not a very good one. Cost-benefit analysis was popularized by industrialists, for whom it was a good tool. It consists of stating a mathematical inequality, or sometimes a simultaneous series of inequalities, which may be compared to determine the course most favoring some objective. Industrialists state their data in dollar amounts, quite easy to contrast. But this is not the case in environmental matters. Consider cancer; in what unit do we state its risk? The probability of developing it? The probability of dying from it? The dollar loss due to disability? The amount of suffering? And can this last be quantified at all? How do we state as a number the value of saving the whale from extinction? Or the snail darter? Are these equal? There is no "unit of risk" as distinguished from the nature of the risk itself.

3) *Risk.* The "no risk" concept has no real existence. We live at risk from conception to death. Adding a specific risk may or may not be justified by circumstance. We save no lives by eliminating carcinogens from our environment. The mortality of those having and not having cancer is exactly the same—100 percent. The real issue, length and quality of life, is hard to evaluate. Consider insecticides; on one side is the risk of poisoning; on the other is the risk of malnutrition or starvation. Which is cost? Which benefit? In choosing a course that will minimize outcomes perceived as "bad" and maximize those perceived as "good" there will be risks on both sides. There is often no conclusive way to identify and select an optimum.

LEWIS G. JACOBS

1701 Vallejo Street,
San Francisco, California 94123

Assessing Diagnostic Technologies

The method offered by Swets *et al.* (24 Aug. 1979, p. 753) to describe the diagnostic efficiency of a test is an elegant way to graph the sensitivity and the specificity (Table 1) of a diagnostic test where observers are not or cannot be standardized to diagnostic criteria.

Unfortunately the "fundamental index, termed A_z," advocated by Swets *et al.* is less useful for comparing diagnostic methods than is visual inspection of their figure 4 or of the usual graphs such as their figure 5, which plots sensitivity against specificity on arithmetic paper.

The best diagnostic criterion is never determined by consideration of sensitivity or specificity alone (*1*). If there are no cost considerations relative to diagnosis and therapy and there is no disadvantage to falsely diagnosing someone as ill, one tries to attain 100 percent sensitivity even though the specificity of that diagnostic criterion may be low. If one is only concerned about screening for healthy individuals one tries to attain 100 percent specificity even at the expense of a low sensitivity. In practice one is usually somewhere between these extremes.

Sometimes the best criterion depends also upon the prevalence of the disease, as when a specific positive predictive value (Table 1) is desired. In that case the optimum combination of sensitivity and specificity for the method changes with prevalence and therefore the appropriate best criterion changes. A graph of sensitivity against specificity such as the graphs in Swets *et al.* is useful to determine the method that delivers the best combination of sensitivity and specificity for a particular diagnostic method given the objectives of the diagnosis (*2*).

However, the A_z index is never useful for any of these choices. The reason can best be understood if two methods (I and II) being compared have lines that cross when sensitivity is plotted on one axis against specificity on the other. At low specificities one method (I) has a higher sensitivity for a given specificity than has the other method (II); the opposite is true at high specificities. If one favors a criterion with a high sensitivity one will choose method I; method II is better when one favors a high specificity. The A_z index gives no information on this important matter.

Once the optimum method is chosen from a sensitivity-specificity plot, the statistical significance of the difference between the methods can be tested at the particular sensitivity or specificity chosen as appropriate for the intended use of the diagnoses. Neither the A_z index nor its variance gives any information about this statistical significance even when the sensitivity-specificity lines do not cross. In fact, comparison of A_z indices is meaningless unless there is a good likelihood that the sensitivity-specificity lines are parallel. This appears unlikely in the comparison of computer tomography and radionuclide scanning for the detection of brain lesions (figure 4 of Swets *et al.*), and unascertainable from the data presented for their other comparisons.

Even when the use of the A_z index is statistically permissible, it is, as noted above, not the logical measure of comparison when choosing a method for its cost-benefit in screening, for its precision in estimating prevalence, for its sensitivity in monitoring change, or for any other characteristic. Is there, then, any use for the A_z index in comparing medical diagnostic methods?

JEAN-PIERRE HABICHT

Savage Hall, Cornell University,
Ithaca, New York 14853

Table 1. Conventional terms used in describing the accuracy of diagnostic methods.

*A. "Truth table"**

Diagnosis by method used	"True" condition	
	Ill	Healthy
Ill	*TP*	*FP*
Healthy	*FN*	*TN*

B. Conventional terms

Sensitivity = $Se = TP/(TP + FN) = P(TP)\dagger = 1 - P(FN)\dagger$

Specificity = $Sp = TN/(TN + FP) = P(TN)\dagger = 1 - P(FP)\dagger$

Negative predictive value = $TN/(TN + FN)$

Positive predictive value = $TP/(TP + FP)$

C. Prevalence estimate of true disease

$Pr = (TP + FN)/(TP + FN + TN + FP)$

D. Interrelation between conventional terms and prevalence

Positive predictive value = $PrSe/[PrSe + (1 - Pr)(1 - Sp)]$

Negative predictive value = $PrSp/[PrSp + (1 - Pr)(1 - Se)]$

*T = true, F = false, P = positive, N = negative.
†Notation as per Swets *et al.*

References

1. R. S. Galen and S. R. Gambino, *The Predictive Value and Efficiency of Medical Diagnosis* (Wiley, New York, 1975), pp. 49-51.
2. J.-P. Habicht, *Am. J. Clin. Nutr.*, in press.

Although the article by Swets *et al.* is excellent methodologically in its demonstration of how to develop relative (or receiver) operating characteristic (ROC) curves for use in assessing diagnostic techniques, the authors distort the role of radionuclide imaging in the diagnosis of brain lesions. In the study "under-

taken both to refine and to illustrate" a protocol, their comparisons of computed tomography (CT) with radionuclide (RN) scanning of the brain were so arranged that the outcome was inevitable.

For the CT readings they had available approximately equal slices and uniform data element resolution, whereas the RN material they used represented a hodgepodge of techniques, some current and some obsolete. Some of the RN readings included immediate and delayed views (delay unspecified) plus flow studies, for some there were no flow studies, and for some there were no immediate studies. In contrast, all the CT readings included images with and without contrast media. We are not told how many patients were in each of the RN categories; nor is there any effort to determine whether different RN techniques gave different results. The radioisotope used was sodium pertechnetate, a radiopharmaceutical that many nuclear-medicine physicians would consider suboptimal. Nothing is told about the scintillation cameras or collimators used; the RN images could have been made with obsolete equipment, cameras having only 19 photomultiplier tubes with poor resolution or collimators having inappropriate resolution characteristics. Modern RN brain imaging requires a high-resolution camera with at least 37 photomultiplier tubes and a high-resolution collimator. A flow study should accompany every RN brain study. Similarly, delayed views must be taken at least 3 and preferably 4 hours after the injection of the radiopharmaceutical in order to obtain a high-quality study. That the authors' RN readers report that "many of [the] presentations were of relatively poor quality" supports these comments.

Because the authors' conclusion that computed tomography is "substantially more accurate than radionuclide scanning" has not been proved by this study, we urge that another, similar study be done comparing current computed tomography with current nuclear-medicine imaging in order to determine the true relative accuracies of these techniques in diagnosing brain lesions.

PAUL M. WEBER
Kaiser-Permanente Medical Center, Oakland, California 94611

We agree with much of what Habicht says, as indicated in the concluding paragraph of our article and the reference made there (*1*). Surely, when one is concerned with a particular diagnostic situation, one should attempt to define the operating point on the ROC curve that is optimal for that situation, and then focus analysis on that particular point rather than on the curve as a whole. We do not agree, however, that all important questions of accuracy pertain to a particular situation, nor that optima can usually be clearly established, and so we find useful an index (A_z) that reflects accuracy in general, through a range of possible operating points.

A general index of accuracy is required when evaluating a diagnostic method from a general vantage point, say, a government health agency. The agency must recognize that the optimum operating point for a given disease will vary considerably from one diagnostic setting to another (for example, community hospital to teaching hospital) and, within one setting, from one patient population to another (for example, from a population being screened to a population at high risk). The agency no doubt will also recognize that optimum operating points for particular situations are frequently not precisely determined, and that the operating point used in a given type of diagnostic situation will vary across locations and perhaps across diagnosticians at one location. It hardly needs to be said that optima are difficult to establish in medicine, depending as they do on prevalence figures that are often not readily available and, more important, on values and costs related to morbidity and mortality. Moreover, translating an established optimum into consistent practice is not simple. Thus an index of accuracy that represents a range of operating points is desirable, if not necessary. At issue, often, is whether a given diagnostic system is generally more accurate than another (for a given disease) and to what extent. The index A_z has convenient and well-studied statistical properties, which make it preferable to the several other general indices that have been considered.

Habicht's concern for the slopes of ROC curves being compared reflects our own. We stated that a general index should be used only if the slopes are not materially different. The slopes of the ROC's presented in figure 4 of our article do not differ enough to begin to reach statistical significance, and the curves do not cross anywhere near an operating point of possible interest. In our opinion, those slopes are similar enough, relative to the distance between the curves, for A_z to be a very useful index of detection accuracy in the context we described. In our experience, ROC slopes are essentially similar in most comparisons.

Weber questions the adequacy of the case studies we used to represent the RN modality. Part of his general concern is that those studies did not employ current RN technology. We stated that our investigation was based on case studies previously assembled over a 3-year period. Perhaps we should have emphasized that these case studies therefore reflect RN and CT technology as of 5 or 6 years ago. Basic to our position is that a confident determination of the performance of newer CT and RN equipment and practices would require another study like ours. And such a study should be undertaken if new developments in either of the modalities are widely thought to effect a substantial increase in diagnostic accuracy.

The second main aspect of Weber's general concern is that, as we reported, the RN case studies did not reflect the same degree of standardization as did the CT studies. This discrepancy is a shortcoming in the design of the original collaborative investigation in which the cases were collected and, indeed, points up the need for active participation from the beginning by members of all relevant disciplines to ensure access to the best available techniques in each discipline. This discrepancy in standardization clearly precluded our comparing CT to a single form of RN regarded as the best available then, but we think that it permits a reasonable comparison of CT and RN as both modalities operated in the field at that time. As far as the general adequacy of the RN studies is concerned, we observe that they were obtained in leading medical centers for diagnostic purposes and that, to our knowledge, none of them was rejected at those institutions for being incomplete or otherwise inadequate in design for the case at hand or because the images were of inferior quality.

Weber has us concluding that CT "is" more accurate than RN. Actually, we wrote that CT "was found to be" more accurate than RN. Our phrase, as it usually does, signifies an appreciation of the fact that conclusions are conditional. Our article generally reflects a concern to describe experimental operations as fully as is practically possible and to highlight those that may have an effect on the meaning and generality of the results. Given the complexity of medical studies, we cannot join in Weber's call for a determination of the "true" relative accuracies of these techniques.

JOHN A. SWETS
RONALD M. PICKETT
Bolt Beranek and Newman Inc., Cambridge, Massachusetts 02238

References

1. J. A. Swets and J. B. Swets, in *Proceedings of the 6th IEEE Conference on Computer Applications and Radiology, 18–21 June 1979, Newport Beach, Calif.* (Institute of Electrical and Electronics Engineers, New York, 1979), p. 203.