Environmental Health Perspectives
Vol. 52, pp. 247-253, 1983

Analysis of Air Pollution Effects: Uncertainties in Proceeding to Standards

by Stanley V. Dawson*

Uncertainties in the collection and assessment of scientific information make ambient air quality standard setting difficult. Uncertainties occur in the estimation of the medical parameters under test due to the inherent random variability encountered in sampling the parameters. The most common method of dealing with random variability is statistical significance testing. The main caution offered in regard to that analysis is to avoid calling a nonsignificant result negative, unless the circumstances are such that the smallest effect which indicates likely harm to health could have been detected with sufficiently high probability.

Uncertainties also play a crucial role in evaluating the implications that even statistically significant test results have for human health. A signal-detection model, developed to explain expert performance in evaluating the results of such diagnostic tests as X-rays, is presented as an analogy for the situation facing experts who are evaluating the implications of health data that is being considered for use in setting a standard. If criteria are too strict for accepting data as evidence of harm to health, then it is argued that, as a consequence, the decision process will not have sufficient ability to discriminate against false-negative results. False-negative results are those that incorrectly conclude there is no threat when, in fact, a particular level of pollutant is actually a threat to health.

The customary approach to setting such safety levels as ambient air quality standards has been based upon deciding a threshold of effects. The threshold is the lowest pollutant concentration at which harmful effects have been observed. Earlier in this century, the only "effect" that was convenient to detect was death of sensitive laboratory animals after pollutant exposure. In order to provide sufficient protection for humans, standards that were set on the basis of this data generally used a large margin of safety. The margin of safety is the difference between the level of the threshold and of the standard. In recent years, advances in science have allowed ambient air quality standards to be based upon thresholds of such nonlethal effects as increased airway resistance and increased pneumonia incidence. Standards based upon nonlethal effects generally have a small margin of safety.

Recently, Lowrance has discussed a risk analysis approach to standard setting (*1*). In this approach, standard setters first determine the risk of harmful effect at different pollutant exposure levels. A graph of this relationship is made. This graph is then used along with information on societal attitudes to decide on the acceptable level of risk and the accompanying limiting exposure level. Regulators can apply a surrogate for a margin of safety in this approach by using a worst case assumption for the graph.

The key issue in either of these approaches to standard setting is the assessment of the human health implications of the available studies. Specific questions might be as follows: Under what circumstances does a 1% increase in mortality in mice colonies exposed to a pollutant level a high enough increase to indicate that humans should not be exposed to that level? What are the implications to human health of a 10% increase in breathing rate of a group of guinea pigs exposed to another level of another pollutant? The answers to such questions are at the heart of the standard-setting process, and yet they are full of uncertainty. Two kinds of uncertainty involved in answering such crucial questions are discussed in the remainder of this article.

*Biological Effects and Air Standards Branch, Research Division, Air Resources Board, State of California, Sacramento, CA 95812.

The first kind of uncertainty is observational. This is the uncertainty of whether or not a pollutant effect was truly present or absent in the situation being observed. The second kind of uncertainty is about what the studies imply about human health. This is the uncertainty of deciding what effect on human health should be inferred from the results of a study on pollutant effects.

In this paper both kinds of uncertainty are characterized on the common basis of the probability that the uncertainty will permit an incorrect outcome. The discussion of the uncertainty of whether an effect was present or absent is essentially an application of standard methods of statistical significance testing. The interpretation of nonsignificant tests receives particular emphasis. That discussion then serves as a background for the treatment of uncertainties about health implications.

Uncertainties of Whether There Was an Effect

Three types of studies are used in modern standard setting to relate biological effects to air pollution exposure: controlled laboratory studies of animals and tissues, controlled laboratory studies of humans, and epidemiological studies of human populations (*2*). In all three kinds of studies, statistical tests are used to assess the probability that any observed biological effects are associated with pollutant exposure, rather than being simply a chance result attributable to the inherent random variability of the phenomena being studied. One important source of variability is simply the variation among the individuals being tested, whether humans or animals. Even in the same strain of carefully bred laboratory animals, some individuals may be much more responsive to pollutant exposure than others. Another source of variability is the imprecision of the measurement itself. The combined variability is generally characterized by the standard deviation σ of measurements, as defined in textbooks on statistics (*3-5*).

Statistical Tests

The statement of statistical test results is usually one of whether the pollutant effect was or was not statistically significant—in other words, whether the probability of an effect being truly present was sufficiently high. To take a simple and important example of a test, one type of experiment measures the difference between breathing rates of two groups of individuals, one of which is exposed to a specific level of pollutant and the other is not exposed. A test result is said to have been statistically significant if the mean difference d between the groups is large enough relative to its variability that the difference is sufficiently unlikely to have been a chance event. The standard error of the mean σ_m measures the magnitude of random variability of the mean difference in the same way as the standard deviation measures the variability of the underlying random process. The standard error of the mean is given by the formula,

$$\sigma_m = \sigma/n^{1/2}$$

where n is the sample size.

The next step is to determine how large the mean difference must be to be accepted as a significant difference. This step requires assumptions about the general nature of random variation of the effects being observed. It is assumed here, as is customary, that the sampled mean values have a normal distribution, perhaps after being transformed.

Hypothesis Testing

The classical statistical analysis tests a "null" hypothesis: that a measured variable remains the same in the presence or absence of a pollutant. If the null hypothesis is true, then the mean difference would be zero. After an appropriate sampling scheme is established and data are acquired, the statistical analysis of this example depends upon computing the ratio d/σ_m, which, in this context, is called a t-statistic because of its assumed distribution. If the t-statistic is larger than a critical value, then the null hypothesis is rejected, and the result is characterized as being statistically significant. The critical value is set so as to limit the probability of declaring that a pollution effect occurred when, in fact, there was none. Statisticians call this kind of erroneous outcome a Type I error. The somewhat arbitrary choice of an acceptable probability of Type I error depends upon how serious the consequences of the error are considered to be.

The probability of committing a Type I error for any possible choice of the critical value can be read from curve A of Figure 1 in the manner indicated. This curve graphically displays the probability that d/σ_m exceeds the value of the horizontal coordinate if there is, in fact, no pollution effect. Curve A assumes large sample sizes; an appropriate t-distribution must be used for each smaller sample size.

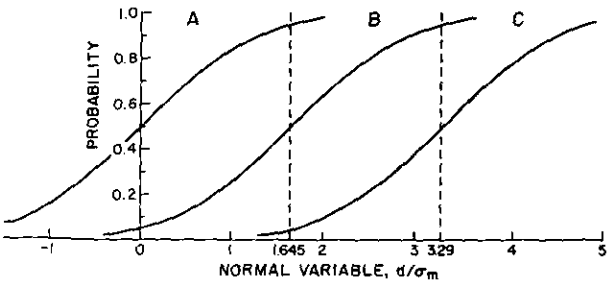

FIGURE 1. Probability that d/σ_m is less than the value indicated on the horizontal axis for three different means (0 for curve *A*, 1.645 for *B*, 3.29 for *C*) of normally distributed d/σ_m. To obtain the probability of Type I error at *X*, subtract the height of curve *A* at the critical value from 1. Given a probability of Type I error and associated critical value, the height of each curve at the critical value is the probability of Type II error for a true effect level at the mean of that curve.

From curve *A* it can be readily seen that the value of d/σ_m will be less than 1.645 for 95% of all samples. Thus, for a critical *t*-statistic (d/σ_m) value of 1.645, the probability of the Type I error is 5%. This is generally the largest Type I error probability used in hypothesis testing. It might be thought appealing to require a greater critical value of *t*-statistic for significance, in order to obtain a smaller probability of Type I error. However, this approach leads to a serious difficulty: a greater proportion of Type II errors. A Type II error occurs if a nonsignificant statistic occurs when, in fact, there was an effect due to pollution.

Interpretation of Nonsignificant Results

A statistically significant result is usually reported as offering substantial evidence of the existence of the pollutant effect being investigated. The interpretation of test statistics that are not statistically significant is more difficult. Often, they are reported as offering substantial evidence for the absence of the effect, when in fact the result should be considered indeterminate, as the following discussion will show.

Statistical tests are often characterized only by the probability of Type I error, i.e., of concluding that a difference was significant when no effect was present. However, the opposite type of error, Type II, also needs to be considered in order to gain understanding of nonsignificant test results. Expressing the probability of Type II error is more complicated than expressing the probability of Type I error, because the probability of Type II error depends upon the magnitude of nonzero mean effect, whereas for Type I error the mean effect is fixed at zero. Discussions of hypothesis test performance when there is truly an effect often refer to the probabilities of not making a Type II error at various effect levels (the statistical power of the test for these levels).

The probabilities of Type II error for three different levels of mean effect are illustrated in Figure 1. Curves *B* and *C* are similar to curve *A* in representing probabilities of normal variables having values less than that indicated on the horizontal axis, but they have their mean values shifted away from zero. The assumed mean response of curve *B* to pollutant is 1.645 and the assumed mean response of curve *C* is 3.29. The probabilities of Type II error for a critical value of 1.645 are read from the intersection of that vertical with each of the three curves. For curve *B*, the probability of Type II error is 50%, because half of all observations will fall below the critical value. For curve *C*, the probability of Type II error is only 5%, just matching the probability of Type I error of the assumed critical value. Note that for a very small value of actual mean, as approximated by curve *A* (the no-effects assumption), the probability of Type II error approaches 95%, a very large chance of failing to detect an effect.

Several well-known introductory texts in statistics offer a fuller explanation of both types of errors and call attention to the importance of controlling Type II errors in the crucial task of designing meaningful experiments (*4, 5*). Such considerations are needed to avoid indeterminate results. A particularly useful set of statistical power calculations for the tests that are most often encountered is given in a text by Guenther (*6*). The importance of statistical power in design of agricultural tests has long been known (*7*), but such considerations have not been common in medical tests (*8*). Land has made some important points concerning the analysis of observational studies (*9*). An enlightening approach to actual analysis of statistical results relative to their uncertainty is given in a text by Hays (*10*).

The dependence of the probability β of Type II error upon the choice of the probability α of Type I error has been displayed graphically by Swets (*11*). Such a graph, shown in Figure 2, can be constructed from curves of the type plotted in Figure 1 by reading off the β for each α as specified in the caption to Figure 1.

The curves that are obtained demonstrate the trade-off between Type I and Type II errors. A decrease in the probability of one type of error always results in an increase in the other for the given fixed value of ratio $d' = d/\sigma_m$. It is also apparent that the closer the curve lies towards

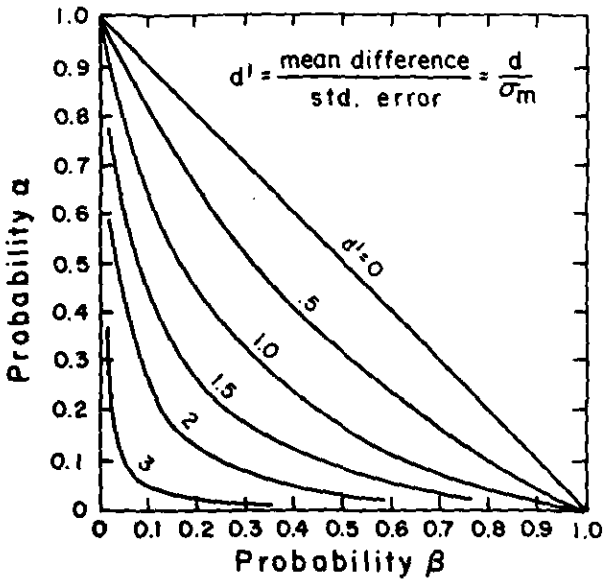

FIGURE 2. In the context of hypothesis testing, α is the probability of Type I error and ß the probability of Type II error. In the context of diagnostic medical tests, α is the probability of false-positives and ß the probability of false-negatives.

the origin, indicating small probabilities of both types of error, the larger the value of the ratio d' must be. Because of the connection between the two types of error, the proper specification of acceptable probabilities of both should depend on public health risks and other societal values of both in an interconnected way. If for example, a high cost is attached to the Type II error (the failure to detect a true effect) and if there is no other feasible way to reduce its probability sufficiently, then it may be prudent to increase the acceptable probability of Type I error to, say, 10%.

A decision on the maximum acceptable magnitude of the two probabilities, α and β, fixes the minimum detectable value of the ratio d'. From Figure 2, the values of β and d' corresponding to a constant α can be determined and plotted to obtain β as a function of d' for each customary value of α (see Fig. 3). Suppose that we have decided upon a maximum acceptable probability β_{max} of Type II error. If we draw a horizontal line on Figure 3 at a height equal to the maximum acceptable probability β_{max}, it will intersect each constant curve at the minimum difference detectable with the specified error probabilities α and β_{max}. If we approximate the true standard error of the mean by its estimate from the data, we may compute $d'\sigma_m$ to obtain an approximation to the minimum effect level d_{min} which the level α hypothesis test can detect with acceptable β.

The ultimate interpretation of a nonsignificant

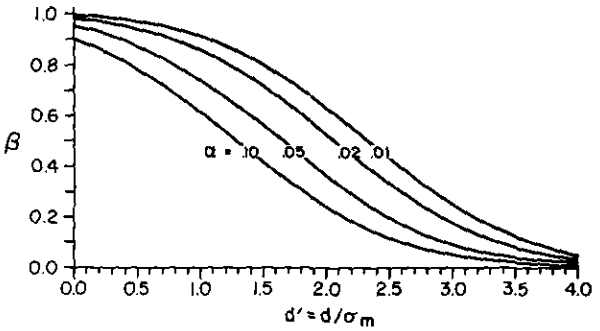

FIGURE 3. Curves of probability ß of Type II error for indicated values of probability α of Type I error, which is the value of d/σ_m that can be detected with the indicated error probabilities, α and ß.

result depends upon the relationship of the minimum detectable level d_{min} to the minimum level of public health concern. If d_{min} is the smaller of the two, then a nonsignificant result should be regarded as essentially negative: within an acceptable probability of error, effects of public health concern are not present. Otherwise the nonsignificant result must be regarded as indeterminate, relative to the minimum level of public health concern.

The decision on the minimum magnitude of an effect that is of public health concern is, along with decisions on the maximum acceptable probabilities, α and β, a public policy judgment. Thus, these limiting values will depend on the context of a particular policy situation. Ordinarily, a scientific investigator assumes a certain nominal value of α in declaring a result nonsignificant. When this occurs, a presentation of estimates of the minimum effects detectable at illustrative values of β, as in Figure 3, is very helpful in interpreting the degree of indeterminateness of a nonsignificant result. If an investigator does venture to characterize a nonsignificant result as strongly supporting the null hypothesis, then a full rationale for the choice of maximum acceptable probabilities, α and β, and the minimum mean difference of public health concern should be clearly stated. In any case it is important for the standard-setting process to be able to review the statistical analysis of studies considered relevant. So a clear statement of variability of the estimate of effect, such as provided by σ_m, is essential for the subsequent analysis to infer a minimum detectible effect d_{min}.

Uncertainty of Implications

Posing the Question

A major problem remaining in the interpretation of the result of a study of pollution effects is

the magnitude of an effect that is to be taken to imply a risk to human health. This decision is at the heart of the standard-setting process and requires a comprehensive analysis of scientific results and social policy. The decision will often have to be taken in the face of great diversity of opinion.

For example, consider a study that shows a statistically significant 10% shift in mean resistance to pulmonary flow immediately after a pollutant exposure. One expert may state: "This result indicates possible serious harm because the pollutant has clearly interfered with the lung's homeostasis." Another expert might offer a differing interpretation of the same study: "This small shift, though statistically significant, is not likely to be of physiological significance because a greater change could be achieved simply by breathing at a lower lung volume." This type of statement is often seen in the "discussion" section of an article appearing in a scientific journal that encourages relevance. Such statements are also heard in regulatory proceedings. The statements made might well be more equivocal, but the question does arise as to how the experts could come to such divergent conclusions.

Analogy of a Diagnostic Medical Test

Some insight into how differing expert opinions come to be offered in connection with developing a health-based standard can be obtained in an analogy to interpretation studies of medical diagnostic tests (*12-17*). One of these studies (*12*) measured the ability of both experienced and inexperienced radiologists to diagnose breast cancer from X-ray films. The accuracy of the X-ray diagnosis was checked against the pathological findings, which were assumed to be correct. When the X-ray and the pathological diagnoses agreed on whether or not breast cancer was present, a true-positive or a true-negative response occurred. Conversely, when the X-ray and the pathological diagnoses disagreed, a false-positive or a false-negative response occurred (see Table 1 for a summary of outcomes). The probability of false-positive diagnoses and probability of false-negative diagnoses for each radiologist was then calculated and plotted (Fig. 4), according to the following equations:

$$P_{FP} = \text{False positives / Cases of tumor absent} = FP/(FP + TN)$$

$$P_{FN} = \text{False negatives / Cases of tumor present} = FN/(FN + TP)$$

Table 1. Classifications of outcomes of tumor diagnosis by X-ray.

	Diagnosis positive	Diagnosis negative
Tumor absent	FP = false positive	TN = true negative
Tumor present	TP = true positive	FN = false negative

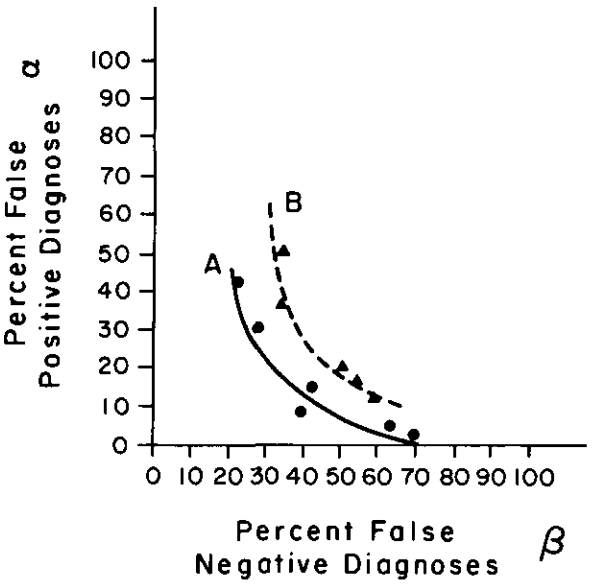

FIGURE 4. Characteristic curves of uncertainty. Frequency of false-positive outcomes versus frequency of false negative outcomes. Diagnosis of breast cancer from mammogram (—) by radiologists and (– –) by trainees. Adapted from Lusted (*12*).

Each point on the plot in Figure 4 represents an individual radiologist. As expected, experienced radiologists (curve *A*) generally predicted fewer false positives and fewer false negatives than inexperienced radiologists (curve *B*). For either level of expertise, however, there was a trade-off between the number of false negatives and number of false positives that occur. Within each level of expertise the individuals that tended to avoid making a false-negative diagnosis made more false-positive diagnoses, and vice versa. An individual therefore reduces one type of false diagnosis at the expense of increasing the other.

A simple signal-detection model of this situation permits the calculation of a mathematical relationship between the probability of false-positive responses and the probability of false-negative responses. If a normal distribution is assumed for a measure of the clarity of the "signal" of X-rays that indicate disease, then the theoretical result is a diagram that is quantitatively identical to that shown in Figure 2 for signifi-

cance testing (*11*). In this diagram α is analogous to P_{FP} while β is analogous to P_{FN}. The signal-to-noise ratio of the detection process is the d' of the figure. As in significance testing, curves representing increasing certainty of detection are closer to the origin. From comparing the shape of the curves in Figures 2 and 4, it appears that the predictions of the model are only qualitatively in agreement with the radiological test results.

Application of the Analogy to Air Quality Standards

The same sort of problems encountered in the interpretation of medical tests are also seen when expert opinion is offered concerning health implications of the results of data on air pollution effects. In interpreting health effects of air pollution, an expert may emphasize avoiding false-positive judgment by rejecting statistically significant studies that may suggest harm to health but are far from establishing a serious effect on human health. Such an expert is, whether consciously or not, increasing the probability that he or she will make more false-negative judgments (not detecting effects when they are there). For example, such an expert might not accept the relevance of animal tests in quantifying standards. The expert who pursues the opposite course, emphasizing avoidance of false-negative judgments about effects while risking more false-positive judgments, will generally be more protective of the public health with respect to that pollutant. With reference to Figures 2 and 4, the protectiveness with regard to public health increases as an individual or group performance point moves towards the upper left-hand corner and decreases as the point moves toward the lower right-hand corner.

Alternative approaches to the use of diagrams such as Figure 2 have been developed for medical decision criteria (*18, 19*). In a more comprehensive approach to a related problem of environmental risk, Page has presented a useful discussion of the probabilities outlined in the present approach (*20, 21*). More traditional approaches to risk assessment are also mentioned in Page's work and in the monograph by Lowrance (*1*). A diverse set of newer approaches to risk in connection with standard setting was gathered together by the U.S. Environmental Protection Agency Risk Analysis Program and made available in report form in 1980 (*22*). One approach, for example offered by Feagans and Biller, is based on still other probability considerations than those of the present work. Each expert interviewed is asked to produce his or her own curve of the probability (vertical axis) that a key effect really occurs at or below the level of pollutant indicated on the horizontal axis of this graph. [See also a related workshop proceeding (*23*).]

Industrial interests have emphasized that air quality standards should be based on "solid scientific evidence." Such a suggestion might seem to imply that the scientific basis of present standards is not now solid enough, despite the fact that an expert scientific review of the basis of federal standards generally performed, as is required by law. Using the present thesis, this suggestion might be interpreted to be a call for experts to increase avoidance of false positives or for decision makers to rely more on experts who do so judge. This approach would tend to relax standards because decision makers would focus on the need for strict criteria of acceptability rather than on requiring a convincing demonstration of the safety of an exposure.

An example of a specific form that such an approach can take among scientists is the suggestion by Ferris and Speizer to narrow the definition of adverse health effects to "medically significant physiologic or pathologic changes generally evidenced by permanent damage or incapacitating illness to the individual" (*24*). Such a recommendation would certainly have the consequence of reducing the probability of false-positive judgments. Thus, decisions using this criterion would have less tendency to be based on an effect of air pollution when none should have been attributed. According to the present analysis, however, such a reduction of the probability of false-positive judgments would inevitably have the consequence of increasing the probability of false-negative judgments. This consequence is of special concern, because such an approach would lead to a direct increase of the risk to human health.

Conclusion

This analysis makes the case that overly strict criteria for judging an effect thought to be of health concern tend to increase risk of harm to public health. The examples given in the text were controlled laboratory experiments, but the same principle of the risk of overly strict criteria for acceptance also applies to epidemiological studies. A crucial example is an observational study which did not control for a potentially confounding variable. If such a study is otherwise of sufficient overall quality, the study may still need to be given weight if an air quality standard is to prevent risk to public health.

In this discussion many simplifications have been made. For example, the emphasis has been on a single effect of a pollutant. In actual standard setting, the full set of effects must be considered together. This is partly in order to test one observation against another and partly to assess the range of effects.

Even with the adoption of a policy of avoiding overly strict criteria for studies accepted as having an effect on health, a substantial margin of safety may still be needed to obtain a standard that assures sufficient protection of public health. In addition, the choice of margin of safety needs to take into account a number of other practical factors, such as the spatial and temporal variation of pollutants and the effects of combinations of pollutants.

The author is grateful for suggestions by A. Alexis, J. K. Moore, H. Griffin, H. Ozkaynak and S. C. Morris and by the anonymous reviewers.

REFERENCES

1. Lowrance, W. W. Of Acceptable Risk. William Kaufman, Inc., Los Altos, CA, 1976.
2. Shy, C. M., Goldsmith, J. R., Hackney, J. D., Liebowitz, M. S., and Menzel, D. B. Health Effects of Air Pollution. American Thoracic Society, American Lung Association, New York, 1978.
3. Brownlee, K. A. Statistical Theory and Methodology in Science and Engineering. John Wiley and Sons, New York, 1960.
4. Armitage, P. Statistical Methods in Medical Research. Blackwell Scientific Publications, Oxford, 1971.
5. Hoel, P. G. Introduction to Mathematical Statistics. John Wiley and Sons, New York, 1947.
6. Guenther, W. C. Concepts of Statistical Inference. McGraw-Hill, New York, 1965.
7. Chew, V. Comparison among Treatment Means. ARS/H/6 U.S. Department of Agriculture, Washington, DC, 1977.
8. Freiman, J. A., Chalmers, T. C., Smith, H., and Kuebler, R. R. The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. New Engl. J. Med. 299: 690–694(1978).
9. Land, C. E. Estimating cancer risks from low doses of ionizing radiation. Science 209: 1109–1203 (1980).
10. Hays, W. L. Statistics for Psychologists. Holt, Rinehart and Winston, New York, 1963.
11. Swets, J. A. The relative operating characteristic in psychology. Science 182: 990-1000 (1973).
12. Lusted, L. B. Decision-making studies in patient management. New Engl. J. Med. 284:416–424 (1971).
13. McNeil, B. J., Keeler, E., and Adelstein, S. J. Primer on certain elements of medical decision making. New Engl. J. Med. 293: 211-215 (1975).
14. Goodenough, D. J., Rossman, K., and Lusted, L. B. Radiographic applications of signal detection theory. Radiology 105: 199-200 (1972).
15. Starr, S. J., Metz, C. E., Lusted, L. B., and Goodenough, D. J. Visual detection and localization of radiographic images. Radiology 116: 533-538 (1975).
16. Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B., and Freeman, B. A. Assessment of diagnostic technologies. Science 205: 753–759 (1979).
17. Swets, J. A. ROC analysis applied to the evaluation of medical imaging techniques. Invest. Radiol. 14: 109–121 (1979).
18. Metz, C. E., Goodenough, D. J., and Rossman, K. Evaluation of receiver operating characteristic curve data in terms of information theory with applications in radiography. Radiology 109: 297–303 (1973).
19. McNeil, B. J., Varady, P. D., Burrows, B. A., and Adelstein, S. J. Measures of clinical efficacy. Cost effectiveness calculations in the diagnosis and treatment of hypertensive renovascular disease. New Engl. J. Med. 293: 216–221 (1975).
20. Page, T. A generic view of toxic chemicals and similar risks. Ecology Law Quart. 7: 207–244. (1978).
21. Page, T. A framework for unreasonable risk in the Toxic Substances Control Act (TSCA). Ann. N. Y. Acad. Sci. 363: 45–166 (1981).
22. U.S. Environmental Protection Agency. Science Advisory Board. Approaches to Health Risk Assessment for Alternative National Ambient Air Quality Standards, A Report of the Subcommittee on Health Risk Assessment. December 1980. Ambient Standards Branch, MD-12. U.S. Environmental Protection Agency, Research Triangle Park NC 27711.
23. Morgan, M. G., Henrion, M., and Morris, S. C. Expert Judgments for Policy Analysis. Brookhaven National Laboratory, BNL 51358, 1979.
24. Ferris, B. G., and Speizer, F. E. (Suggested) criteria for establishing standards for air pollutants, In: The Business Roundtable Air Quality Project, Vol. 1, National Ambient Air Quality Standards, November 1980.