possible justification for using a screening test in spite of these problems. The recent, highly publicized Walker spy case is but one example of several recent disasters in our national security system. The consequences of failing to detect leaks of secret information to foreign governments may be severe. A polygraph test that correctly identifies 88% of deceptive individuals tested, misclassifies only 3% and yields 9% inconclusive outcomes could be relied upon to identify most security risks. However, since the base rate of deception in this population is so low, most of the individuals who would fail the test would in fact be truthful. If a deceptive polygraph outcome is more often wrong than it is correct, it is clear that it should not be the sole basis for concluding that a person is a spy, for denying individuals access to secure information or for taking other action against them.

On the other hand, if the screening test is used only to eliminate from further consideration all those who pass the test, then the number of potential security risks would be reduced by a factor of approximately 10 (Raskin and Kircher, 1987). Extensive field investigations would then be required on a much smaller number of individuals with a somewhat higher base rate of deception than in the original sample. With this "successive hurdles" approach (Meehl and Rosen, 1955), polygraph screening tests could be used in the vast majority of cases in lieu of costly field investigations. The required follow-up investigations of those who fail the initial screening test would minimize the risk of false positive errors and probably identify the individuals who are guilty of compromising our national security.

ADDITIONAL REFERENCES

Kircher, J. C., Horowitz, S. W. and Raskin, D. C. (1987). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior*. To appear.

Lykken, D. T. (1979). The detection of deception. *Psychol. Bull.* **86** 47–53.

Meehl, P. and Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.* **52** 194–214.

Podlesny, J. A. and Raskin, D. C. (1977). Physiological measures and the detection of deception. *Psychol. Bull.* **84** 782–799.

Raskin, D. C. (1984, March). Proposed use of polygraphs in the department of defense. Statement before the Committee on Armed Services, U.S. Senate.

Raskin, D. C. (1987). Does science support polygraph testing? In *The Polygraph Test: Truth, Lies and Science* (A. Gale, ed.). Sage, London. To appear.

Raskin, D. C. and Kircher, J. C. (1987). The validity of Lykken's criticisms: Fact or fancy? *Jurimetrics J.* **27** 271–277.

Raskin, D. C. and Podlesny, J. A. (1979). Truth and deception: A reply to Lykken. *Psychol. Bull.* **86** 54–59.

Comment

Janet Wittes

Professor Gastwirth's most interesting paper, coupled with my craving for poppy seed bagels and my passion for our Fourth Amendment right to privacy, has led me to a new appreciation of the importance of specificity θ in medical screening. My work with Dr. Goldberg (Goldberg and Wittes, 1978, 1981) has focused on the sensitivity η; the inverse symmetry of Dr. Gastwirth's equations (3.1) and (3.3) point to diametrically opposed criteria for optimality depending on whether one is interested primarily in the predicted value positive (PVP) or the predicted value negative (PVN). In the former case, Gastwirth shows that θ should be estimated most precisely; in the latter, the emphasis should be placed on η.

The context of the screening determines whether the sensitivity or the specificity is more important. For the cases that Dr. Goldberg and I have considered in the past, screening was performed for the benefit of the screenee. A woman elects to participate in a breast cancer screening, for example, because she is seeking an early diagnosis of a disease for which early diagnosis can translate to her own lengthened survival (Shapiro, Strax, Venet and Venet, 1973). Hence, from her, the consumer's point of view, a screening program consisting of a highly sensitive test, followed by a highly specific test if she is positive, is a sensible course of action. Consideration of the PVP is then secondary to the needs of the consumer. When, however, the consumer is not the screenee, but the society at large, and when that society assumes an implicitly adversarial position with respect to the screenee, Gastwirth's emphasis on the primacy of the specificity

Janet Wittes is Chief, Biostatistics Research Branch, National Heart, Lung, and Blood Institute, Bethesda, Maryland 20892.

is, I believe, apt. My mention of poppy seeds at the beginning of these remarks was therefore not facetious: we federal employees, threatened with the possibility of random urine spot checks for drugs, must beware lest our breakfast include poppy seeds or other foods chemically similar to illicit substances.

Gastwirth's expression for the variance of $\hat{C}$, the estimated PVP, has an immediate social consequence: when screening is performed at society's behest, and not voluntarily at the request of the screenee, and when, further, a positive test carries a strong negative social presumption of improper behavior, as would false positive polygraphs or false positive drug tests, the specificity should be very high and the PVP precisely estimated. We must take seriously Gastwirth's admonition that the usual allocation of subjects for the purpose of estimating sensitivity and specificity is inefficient when the population prevalence is low. As we have learned from the famous yellow Lincoln case of the California Supreme Court (1968), those who deal with other people's lives and reputations must keep their conditional probabilities straight.

AIDS represents a special situation. Testing is currently usually performed not because the individual has requested to be screened, but rather because blood banks routinely screen all blood donations. Clearly, society would be ill-served by a screening program that emphasized specificity. Therefore, the cut-off value for declaring the enzyme-linked immunosorbent assay (ELISA) to be positive is set so the test is highly sensitive at the expense of having rather low specificity. Screening programs for AIDS achieve high specificity by following a positive ELISA with a confirmatory Western blot. Currently, positivity on ELISA is not reported to the screenee on the grounds that most ELISA positives are false positives. As screening for AIDS becomes more common, so long as screeners continue to understand that most people positive on ELISA are not carriers of HIV, and so long as positivity on ELISA alone is not reported to the screenee, his employer or his insurer, then low specificity at the initial ELISA is an appropriate approach to achieving high sensitivity.

Gastwirth's paper implicitly raises a problem caused by the nondichotomous nature of many disease processes. Sensitivity and specificity have clear meanings when a disease can be characterized either as present or absent. For diseases with stages, sensitivity and specificity are more difficult to interpret. More generally, if we view disease as a process unfolding in time, we may imagine sensitivity and specificity to change with the course of disease. Letting t denote the time from onset of the disease, then $\theta(t)$ represents the sensitivity as a function of time. In some diseases, $\theta(t)$ is not monotone. Moreover, a test may be specific to a particular stage of disease, so that η may also change with time. Further, two different screening tests may have different patterns of change in $\theta(t)$ with time. Both the statistical and epidemiologic literatures have discussed the effect of population prevalence on estimates of the PVP; the literature has not grappled sufficiently seriously with the consequence on the estimation of sensitivity and specificity when the measures are made from artificially homogeneous populations of affected and unaffected individuals. This phenomenon may partially explain the highly discrepant estimates of sensitivity and specificity noted by some authors (see Goldberg, 1975). Thus, if sensitivity and specificity are calculated from patients who are all in a similar stage of disease, the screening test may exhibit unexpected operating characteristics when it is applied in a more heterogeneous setting. Sensitivity is often estimated by using a hospitalized group of patients, but the screen is then applied to a group of people with no overt signs of disease. In such situations, the estimated PVP may be even more variable than equation (2.5) would suggest. One useful extension of Professor Gastwirth's results would be to estimate C for diseases with nonconstant $\theta(t)$ and $\eta(t)$.

Some diseases are dichotomous themselves but their associated screening test measures a continuous variable related to the presence of disease. For example, in testing for AIDS, the HIV antibody is either present or absent, but ELISA measures the level of certain antigens in the blood. An arbitrary value of the level is then chosen as an indication that the antibody is present. As mentioned above, this level has been selected so that the screening program, ELISA followed by blot, is both extremely sensitive and extremely specific.

Sometimes apparent false positives may not be false positive at all. For example, informal clinical folklore holds that women who are falsely positive at a screening examination are more likely to develop breast cancer than those who are screened negative. (In this case, a false positive means that the mammogram is positive but the confirmatory biopsy is negative.) A possible explanation for elevated risk among false positive screenees is that apparently nonspecific screening tests may indeed be identifying a cohort characterized by a stage of disease the so-called gold standard fails to recognize. Mammography, which identifies some noncancerous lesions, may be identifying a type of lesion that is likely to become malignant.

Professor Gastwirth's valuable paper should provide a starting point for the estimation of the predictive value of a positive screening test. His insight that the variance of the estimate is large when the population prevalence is low is, I believe, especially relevant when screening is not voluntary.

ACKNOWLEDGMENT

I thank Dr. Charles Hollingsworth of the Blood Resources Branch, Division of Blood Diseases and Resources, National Heart, Lung, and Blood Institute, for his information about ELISA and Western blot.

ADDITIONAL REFERENCES

GOLDBERG, J. D. and WITTES, J. T. (1981). The evaluation of medical screening procedures. *Amer. Statist.* **35** 4–11.

PEOPLE VS. COLLINS. (1968). *California Reporter* **66** 242–253. West Publishing Co., St. Paul, Minn.

SHAPIRO, S., STRAX, P., VENET, L. and VENET, W. (1973). Changes in five-year breast cancer mortality in a breast cancer screening program. *Seventh Nat. Cancer Conf. Proc.* 663–678.

Comment

Judith D. Goldberg

Professor Gastwirth addresses an important and interesting problem, the evaluation of medical screening procedures and programs.

The examples of AIDS screening with enzyme-linked immunosorbent assay (ELISA) and the use of the polygraph to detect deceptive individuals raise the broader questions of how to estimate the sensitivity and specificity of a screening test and how to implement and monitor the widespread use of a test in a population. The difficulties in the estimation of C arise from the practical issues of obtaining useful estimators of the sensitivity and specificity from incomplete data often in the absence of confirmatory testing of negatives on screening. The precision of $\hat{C}$ is a function of the precision of the component estimators that may themselves have large variances depending upon the method of estimation.

I would like to clarify the terminology used by Gastwirth. The traditional false negative rate or Neyman-Pearson type I error is defined as the proportion of "diseased" individuals who are negative on screen or (1 − sensitivity); the false positive rate, analogous to the type II error, is defined as the proportion of "nondiseased" individuals who are positive on screen or (1 − specificity) (Goldberg, 1975). These rates do not depend on the prevalence. The predictive value of a positive test, Gastwirth's C, the quantity of interest in this paper, was originally defined by Vecchio (1966) and does depend on the disease prevalence as does the predictive value of a negative test or $(1 - F)$ in Gastwirth's discussion. Gastwirth's F is not the traditional false negative rate nor is C the traditional true positive rate (Tables 1 to 3).

Goldberg and Wittes (1978) estimate the traditional false negative rate (1 − sensitivity) of a screening program, and not F as indicated by Gastwirth. The estimator depends on a capture-recapture estimator of the number of diseased individuals in the population. The observed data used to obtain the estimate are the numbers of positives on each of two distinct screening mechanisms; the prevalence of the disease is not estimated. The proposed estimators are useful when no confirmatory test is administered to individuals who are negative on the dual screening.

For example, in the Health Insurance Plan breast cancer screening program, a randomized trial designed to evaluate periodic screening with mammography and clinical examination, negatives on screen were returned to the population pool (Shapiro, Strax, Venet and Venet, 1973). The false negative rates estimated from this study vary and have wide confidence intervals even when the population is stratified into reasonably homogeneous groups.

In his paper, Gastwirth examines the estimated standard errors of $\hat{C}$ when prevalence and sample size vary with sensitivity and specificity held fixed (and assumed known or estimated from another source). Because $\hat{C}$ depends on the error rates as well, the sensitivity analysis should address the implications on estimation of the range of possible error rates for useful screening tests.

For diseases of low prevalence ($\pi \leq .05$), the bias in the estimator of the proportion positive on screening when there is misclassification depends primarily on the false positive rate (Goldberg, 1975). $\hat{C}$ depends on the misclassification rates both directly and indirectly through the estimator of the proportion positive on screening.

Gastwirth points out that prevalence can vary from group to group. It is just as likely that the false negative and false positive rates will vary from group to group for the same test (Goldberg, 1975). Thus, the analysis of the sensitivity of the precision of $\hat{C}$ should address first the sensitivity of $\hat{C}$ itself to underlying prevalence and error assumptions because a precise,

Judith D. Goldberg is Director, Statistical Design and Analysis, Medical Research Division, American Cyanamid Company, Pearl River, New York 10965.