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.*

but inaccurate, estimate will lead us astray in practice. Our inferences can be erroneous and lead us to poor policy decisions.

The large effects on estimation and inference that can be attributed to misclassification suggest that resources should be allocated to estimation of these error rates prior to the implementation of a mass screening program and on an ongoing basis for the duration of the program. The costs of classification errors are high to both individuals and society. The existence of a screening program itself may alter behavior of individuals, and the disease process may change from the intervention after screening and from improvements in both the screening method and therapy. These and other related issues in the evaluation of medical screening procedures are discussed in Goldberg and Wittes (1981).

## ADDITIONAL REFERENCES

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

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.

VECCHIO, T. J. (1966). Predictive value of a single diagnostic test in unselected populations. *New England J. Med.* **274** 1171–1173.

# Comment

**Seymour Geisser**

We are indebted to Professor Gastwirth for an enlightening discussion regarding the reliability of the results of screening tests in two rather important areas: AIDS and lie detectors. His main concern is with the conditional probabilities of correct classification and the sampling error of their frequentist estimators.

I would like to outline an approach that I believe might be more informative and illuminating for inferring the results of such screening tests. For the sake of simplicity, let us assume that there is a properly identified population and a single test (multiple tests and varying populations would only further serve to complicate the situation but not change the conceptual framework for handling such problems).

With the use of Prof. Gastwirth's notation, we have a table exhibiting the following probabilities:

| | $D$ | $\bar{D}$ | |
|---|---|---|---|
| $S$ | $\pi\eta$ | $(1-\pi)(1-\theta)$ | $\pi\eta + (1-\pi)(1-\theta)$ |
| $\bar{S}$ | $\pi(1-\eta)$ | $(1-\pi)\theta$ | $\pi(1-\eta) + (1-\pi)\theta$ |
| | $\pi$ | $1-\pi$ | 1 |

where, e.g., $P(D) = \pi$, $P(S \mid D) = \eta$, $P(\bar{S} \mid \bar{D}) = \theta$; $P(S) = \pi\eta + (1-\pi)(1-\theta) = p$. The critical so-called PVP,

$$P(D \mid S) = \frac{\pi\eta}{\pi\eta + (1-\pi)(1-\theta)} = \tau, \quad \text{say,}$$

and the probability of a false negative,

$$P(D \mid \bar{S}) = \frac{\pi(1-\eta)}{1 - \pi\eta - (1-\pi)(1-\theta)} = \rho, \quad \text{say,}$$

are functions of the three parameters, $\pi$, $\eta$ and $\theta$.

The type of sampling that Professor Gastwirth deals with in the paper presumably would yield a likelihood function for $\theta$, $\eta$ and $\pi$,

$$L(\theta, \eta, \pi) \propto \eta^{r_1}(1-\eta)^{n_1-r_1}\theta^{r_2}(1-\theta)^{n_2-r_2}\left(\frac{\pi\eta}{\tau}\right)^{t}\left(1 - \frac{\pi\eta}{\tau}\right)^{n-t},$$

recalling that $\tau$ is a function of $\theta$, $\eta$ and $\pi$. Suppose a joint prior for $\eta$, $\theta$ and $\pi$, $g(\eta, \theta, \pi)$ is available. Then the posterior density of $\theta$, $\eta$ and $\pi$ is

$$p(\theta, \eta, \pi \mid d) \propto L(\theta, \eta, \pi)g(\eta, \theta, \pi),$$

where $d = (r_1, r_2, n_1, n_2, t, n)$.

Clearly, if we were diligent and clever enough, we could find from $p(\theta, \eta, \pi \mid d)$ the joint posterior density of $\tau$ and $\rho$, say $p(\tau, \rho \mid d)$. Ostensibly then for any set $S$ on the unit square we could find

$$P[(\tau, \rho) \in S] = P,$$

or conversely for any fixed $P$ we could find the "smallest" set $S_P$ such that

$$P[(\tau, \rho) \in S_P] = P.$$

Similar results could be obtained marginally for either $\rho$ or $\tau$. This would be much more informative than the calculation of the approximate standard errors of the estimates $\hat{C}$ and $\hat{F}$. Of course this would require a good deal of heavy calculation involving numerical

*Seymour Geisser is Professor and Director, School of Statistics, University of Minnesota, 270 Vincent Hall, 206 Church Street, S.E., Minneapolis, Minnesota 55455.*