Statistical Science
1987, Vol. 2, No. 3, 213–238

The Statistical Precision of Medical Screening Procedures: Application to Polygraph and AIDS Antibodies Test Data

Joseph L. Gastwirth

Abstract. The increased use of screening tests for drug use or antibodies to the HTLV-III (AIDS) virus, as well as pre-employment polygraph testing, has raised concerns about the reliability of the results of these procedures. This paper reviews the mathematical model underlying the analysis of data from screening tests. In addition to the known formulas for the proportion of positive (negative) classifications that are correct, we provide a large sample approximation to their standard errors. The results reinforce the need for confirmatory tests and indicate that moderately large sample sizes should be used to determine the accuracy rates of screening tests that will be applied to the general population in which the prevalence of the disease or trait is low.

Key words and phrases: Screening tests, polygraph specificity, sensitivity, predictive value of a positive test, large sample theory.

1. INTRODUCTION

As the use of medical screening tests for drug use or exposure to AIDS antibodies and pre-employment polygraph (or "lie detector") testing have become more widespread, concerns have been expressed about their routine use. In part, these concerns arise because the prevalence of persons with the disease or a deceptive character in the general population is far less than that in a prescreened group, e.g., persons in a high risk category for the disease in question, so there may be a high fraction of false positive classifications among the test results. The discussion of screening tests in standard texts (Ingelfinger, Mosteller, Thibodeau and Ware, 1983; Sackett, Haynes and Tugwell, 1985) assumes that the accuracy rates, that is, the sensitivity and specificity of the screening test, and the prevalence π of the disease in the population to be tested are known. The sensitivity (specificity) of a test is the probability that a person having (not having) the disease is correctly classified. In his analysis of polygraph data, Raskin (1986) estimates π from the proportion of persons classified as deceptive using the method developed by Steinhaus who analyzed paternity cases (Finkelstein, 1978; Solomon, 1966); however, the sampling error of the resulting conditional probability that the classification is correct did not play a role in the data analysis.

The purpose of this paper is to show that the estimated conditional probability, C, of interest has an asymptotic normal distribution and to show the effect of this error on the inference one can draw from a screening test. The sampling variability of the estimate of C and the effect of the accuracies of the screening test are considered. Section 2 is devoted to a review of the general framework and the main mathematical result. The results are applied to data on the enzyme-linked immunosorbent assay (ELISA) test used to screen blood for antibodies to the AIDS virus in Section 3. Raskin's data (1986) on the polygraph are reanalyzed in Section 4. Our results indicate that the standard error of the estimate $\hat{C}$ of the conditional probability a person diagnosed as having a disease (or being deceptive) actually has the disease (or is deceptive) *increases* as the prevalence of the disease or trait in the population tested *decreases*. Thus, the problem of a high fraction of false positives is exacerbated by a high degree of uncertainty. An examination of the components of the standard error indicates that the specificity needs to be quite accurately estimated in order to make the standard error of $\hat{C}$ small, e.g., .02, when it is applied to a large population which has a moderate prevalence of the disease. This suggests a change in the usual allocation of samples between diseased and disease-free individuals when the accuracy rates of a screening test are determined or

Joseph L. Gastwirth is Professor of Statistics and Economics, Department of Statistics, C-315, George Washington University, Washington, D. C. 20052.

preferably an increase in the size of the disease-free sample.

2. THE MATHEMATICAL FRAMEWORK USED TO ANALYZE THE RESULTS OF A SCREENING TEST

The purpose of a screening test is to determine whether a person belongs to the class (D) of people who have a specific disease. The test result indicating that a person is a member of the class D will be denoted by S and a result indicating nonmembership by $\bar{S}$. The *accuracy* of a test is specified by two probabilities,

$\eta = P[S \mid D]$
= the probability a person with the disease is correctly diagnosed, called the sensitivity of the test,

and

$\theta = P[\bar{S} \mid \bar{D}]$
= the probability a disease-free individual is correctly diagnosed, called the specificity of the test.

Our focus is on the conditional probability, $P[D \mid S]$, that a person whom the test indicates as having the disease actually has it. Letting $\pi = P(D)$ denote the prevalence of the disease in the population tested, and assuming that a person tested can be regarded as randomly selected from that population, Bayes' theorem yields

$$P[D \mid S] = \frac{P[D \wedge S]}{P[S]}$$

$$= \frac{\pi\eta}{\pi\eta + (1-\pi)(1-\theta)} \tag{2.1}$$

$$= \frac{\pi\eta}{\pi(\eta + \theta - 1) + (1-\theta)}.$$

This probability is called the predictive value of a positive test (PVP).

In mass screening programs the prevalence π may vary among subpopulations so that is must be estimated separately in each one. This is accomplished by using the fact that the probability a person tested will be diagnosed as sick, is

$$p = P[S]$$

$$= \pi P[S \mid D] + (1-\pi)P[S \mid \bar{D}] \tag{2.2}$$

$$= \pi\eta + (1-\pi)(1-\theta).$$

When a sample of n persons are tested, one estimates p by the proportion $\hat{p}$ of those who are classified in S. Solving for π in equation (2.2) yields the estimate of prevalence studied by Rogan and Gladen (1978), which is

$$\hat{\pi} = \frac{\hat{p} + \theta - 1}{\eta + \theta - 1} = \frac{\hat{p} - (1-\theta)}{\eta + \theta - 1}. \tag{2.3}$$

Substituting $\hat{\pi}$ for π in (2.1) yields the estimator $\hat{C}$ for $P[D \mid S]$; namely,

$$\hat{C} = \frac{\hat{\pi}\eta}{\hat{p}} = \frac{\eta}{\eta + \theta - 1}\left[1 - \frac{(1-\theta)}{\hat{p}}\right]. \tag{2.4}$$

It should be noted that formula (2.3) can yield an estimate of prevalence which does not lie between 0 and 1. For example, if η and θ are low, p can be less than $1 - \theta$ and π negative. In practice, when η and θ are moderately large, e.g., $\geq .8$, and π is not too small, this happens very infrequently. To avoid such problems, one can define a truncated version $\hat{\pi}_1$, of $\hat{\pi}$ as $\hat{\pi}_1 = \min[\max(\hat{\pi}, 0), 1]$ and use $\hat{\pi}_1$ in place of $\hat{\pi}$ in solving for $\hat{C}$, except that if $\hat{\pi}_1 = 1$, $\hat{C}$ should be set = 1. In the Appendix we show that $\hat{\pi}_1$ and $\hat{\pi}$ are asymptotically equivalent to order $n^{-1/2}$ so the large sample theory of Rogan and Gladen is valid for $\hat{\pi}_1$. However, when π is small, the sample size n required for the applicability of the large sample results can be quite large.

When η and θ are not known but are estimated by $\hat{\eta}$ and $\hat{\theta}$ based on samples of size n_1 and n_2, where the screening test is used on persons whose disease status is known, we replace η and θ by these estimated values. The main statistical result will now be stated:

THEOREM. *As n, n_1 and n_2 increase, the sampling distribution of $\hat{C}$ is approximately normal with mean $P[D \mid S]$ and variance*

$$\left\{\frac{\eta(1-\theta)}{p(\eta + \theta - 1)}\right\}^2 \frac{p(1-p)}{p^2 n}$$

$$+ \left\{\frac{\pi(1-\theta)}{p(\eta + \theta - 1)}\right\}^2 \frac{\eta(1-\eta)}{n_1} \tag{2.5}$$

$$+ \left\{\frac{\eta(1-\pi)}{p(\eta + \theta - 1)}\right\}^2 \frac{\theta(1-\theta)}{n_2}.$$

Before illustrating the use of (2.5) we note that when η and θ are *known* the last two terms in (2.5) vanish. These terms reflect the variability in our estimate of the true positive rate due to uncertainty about the true value of the specificity and sensitivity of the screening test. The first term reflects the variation due to the fact that p is estimated by the proportion of persons tested who are classified in D. From formula (2.5), we realize that in the case when η and θ are high, but the prevalence (π) of the disease is low, the third term, the contribution of the variability in the estimate of the sensitivity θ can be the *dominant* term. Moreover, both the first and third terms increase

as π decreases so that the use of screening tests on groups that have a low prevalence rate will often yield a low predictive value positive that has a large standard error.

Of at least as much importance in public health as C is the false negative rate,

$$F = P[D \mid \bar{S}] = \frac{\pi(1-\eta)}{(1-p)}, \tag{2.6}$$

which is 1 minus the probability a negative test is correct or 1 minus the predictive value of a negative test result (PVN). In the context of monitoring, the estimation of F has been studied by Goldberg and Wittes (1978) so that well shall not focus on it. For the sake of completeness, we note that

$$\hat{F} = \frac{\hat{\pi}(1+\hat{\eta})}{(1-\hat{p})} = \frac{\hat{\pi}(1-\hat{\eta})}{\hat{\pi}(1-\hat{\eta}) + (1-\hat{\pi})\hat{\theta}}$$

also has a large-sample normal distribution with mean $P[D \mid \bar{S}]$ and variance given by

$$\begin{aligned} &\frac{(1-\eta)^2\theta^2}{(\eta+\theta-1)^2(1-p)^4}\,\frac{p(1-p)}{p} \\ &+\left[\frac{\pi\theta}{(1-p)(\eta+\theta-1)}\right]^2\frac{\eta(1-\eta)}{n_1} \\ &+\left[\frac{(1-\eta)(1-\pi)}{(1-p)(\eta+\theta-1)}\right]^2\frac{\theta(1-\theta)}{n_2}. \end{aligned} \tag{2.7}$$

In the mass screening applications with which we are concerned, formula (2.7) and the normal approximation should not be used to develop confidence intervals for F unless the adequacy of the normal approximation has been checked since the sample sizes n, n_1 and n_2 required for its accuracy will be large if the accuracies η and θ are high and p and π are expected to be small.

3. APPLICATION TO SCREENING FOR AIDS

The ELISA test for AIDS is used to screen donated blood for the AIDS antibody. An evaluation of this test by Weiss et al. (1985) yielded $\hat{\eta} = {}^{86}/_{88} = .977$ and $\hat{\theta} = {}^{275}/_{297} = .926$ when "borderline" results are classified as having the antibody. Although Barnes (1986) reports that blood banks require three positive ELISA tests before the confirmatory Western blot test is administered, we shall examine the statistical characteristics of just one screening test. Indeed, only one test is often administered in mass screening programs and apparently it is the practice in drug use testing by some employers.

In Table 1 we present the estimated PVP, $\hat{C}$ and its approximate standard deviation as a function of the prevalence rate for two sample sizes (n) of the tested population. The *proportion* of the variance of $\hat{C}$ that is due to uncertainty in our knowledge of θ is also given. The results show that when the prevalence π is reasonably large, e.g., .4 or more, the expected value of $\hat{C}$ is reasonably high, about .9, and its standard error small, e.g., about .025, even accounting for uncertainty in our knowledge of the specificity and sensitivity of the screening test. On the other hand, when π is small, not only is $\hat{C}$ less than .5, indicating a rate of false positives exceeding 50%, its standard error is also quite high. Even when π is .10 and C is expected to be nearly .6, the standard error of $\hat{C}$ remains about .10, even when a sample of 10,000 persons is screened. Indeed, the effect of sampling variability in the estimates of η and θ on the standard error of $\hat{C}$ remains regardless of the size of the population screened.

TABLE 1

Approximate standard errors of the estimated true positive rate $\hat{C}$ when $\eta = .977$, $\theta = .926$ if they are known or estimated from samples of size $n_1 = 88$ and $n_2 = 297$ as a function of the disease prevalence π and size n of the population to be screened

Prevalence π	$E(\hat{C})$	Standard error if η, θ are known	Standard error if η, θ are estimated	Percentage of the variance of $\hat{C}$ due to estimation of θ
		$n = 500$		
.50	.930	.0065	.0170	84.9
.40	.898	.0094	.0245	85.2
.20	.768	.0241	.0570	82.1
.10	.595	.0491	.1026	77.0
.05	.410	.0817	.1544	72.0
.03	.290	.1056	.1898	69.0
.01	.118	.1433	.2428	65.2
		$n = 10{,}000$		
.50	.930	.0014	.0158	98.5
.40	.898	.0021	.0229	98.9
.20	.768	.0054	.0519	99.0
.10	.595	.0110	.0907	98.5
.05	.410	.0183	.1323	98.1
.03	.290	.0236	.1595	97.8
.01	.118	.0320	.1986	97.4

Note: The low prevalence rates and moderate sample sizes ($n = 500$) are not sufficiently large for $\hat{\pi}_1$ and $\hat{\pi}$ to be equivalent. The standard error using $\hat{\pi}_1$ rather than $\hat{\pi}$ in (2.4) would be smaller; however, the main features of the table and conclusions would not be seriously affected.

In order to assess what might happen if a more accurate test were devised or if multiple positive tests were required (it is hard to model this situation precisely as the tests are unlikely to be statistically

independent), we assumed that $\eta = .99$, $\theta = .98$, $n_1 = 88$, $n_2 = 297$ and $n = 10{,}000$. If $\pi = .05$, $E(\hat{C}) = .923$, however; the approximate standard error of $\hat{C}$ when η and θ were estimated was .116. When $\pi = .02$, $E(\hat{C}) = .503$ and the approximate standard error was .208. Since the normal approximation to the sampling distribution of the proportions $\hat{\eta}$ and $\hat{\theta}$ is not that accurate for samples of these sizes (a Poisson approximation is preferable), one should not use these results to create a formal confidence interval. Recent simulations of Hammick (1987) for the case when η and θ are known ($\eta = .977$, $\theta = .926$) indicate that the normal approximation is valid when π is at least .05. For smaller prevalence rates, the results in Table 1 are underestimates of $E(\hat{C})$ and overestimates of the standard error. When $\pi = .01$, the Monte Carlo standard error was roughly half that in Table 1. Nevertheless, the estimated coefficient of variation of $\hat{C}$ in this case was about 50%, which is quite high.

The results also imply that the sample sizes used to determine the specificity of a screening test may need to be increased, especially if the test will be used on the general population rather than on subgroups known to have an above average prevalence or risk. Not only can this be seen by an examination of Table 1, it can be formally derived from formula (2.5) by regarding the sum $n_1 + n_2 = N$ as fixed and finding the choice of n_1 and n_2 as fractions γ and $1 - \gamma$ of N which minimize the sum of the two rightmost terms in (2.5). Routine calculus yields

$$\frac{n_1}{n_2} \simeq \frac{\pi}{1-\pi}\sqrt{\frac{(1-\theta)(1-\eta)}{\theta\eta}}. \tag{3.1}$$

Formula (3.1) shows that the optimal allocation of the subjects used to determine the accuracy of the screening test is related to the prevalence rate in the population that will be screened and the accuracy rate expected. It also provides the optimal allocation when the objective is to minimize the standard error of $\hat{C}$. If one desired to estimate the prevalence π or the PVN, F, most efficiently, the optimal allocation would be different. For example, from Rogan and Gladen's results, we find that the optimal allocation for estimating π is

$$\frac{n_1}{n_2} \simeq \frac{\pi}{1-\pi}\sqrt{\frac{\eta(1-\eta)}{\theta(1-\theta)}}. \tag{3.2}$$

Similarly, when the focus of attention is the estimation of the false negative rate, the optimal allocation is

$$\frac{n_1}{n_2} = \frac{\pi}{1-\pi}\sqrt{\frac{\eta\theta}{(1-\eta)(1-\theta)}}. \tag{3.3}$$

Formula (3.1) was based on the large sample variance (2.5), so both n_1 and n_2 should be large, e.g., at least 100 and preferably more. The approximate 1:3 ratio for the sample sizes 88:297 in the NIH study (Weiss et al., 1985) is quite appropriate since the test will be applied to a high risk population. Formula (3.1) suggests that if a screening test with values of η and θ near .95 is used to screen a general population with a prevalence less than .05, the number n_2, of disease-free persons used in the determination of θ should be even larger relative to n_1.

So far we have been concerned with minimizing the standard error of $\hat{C}$. Comparing formulas (3.1) and (3.3) reveals that the optimal allocation for minimizing the standard error of $\hat{F}$ differs substantially as the accuracy of the estimated sensitivity $\hat{\eta}$ plays a more important role. Even so, if $\eta = \theta = .95$ and $\pi = .01$, which is reasonable for an AIDS screening test given to a low risk group, formula (3.3) yields an optimal allocation $n_1{:}n_2$ of 1:5. In practical terms these considerations imply that accuracy rates should be determined on larger samples, especially of disease-free individuals.

Finally, we note that our calculations assumed that η and θ were estimated from simple random samples in which the disease status was known. In practice, even the "gold standard" diagnostic or reference test, e.g., the Western blot test for AIDS antibodies, may not be 100% accurate so that its small error rates should be taken into account (Gart and Buck, 1966; Greenberg and Jekel, 1969).

The accuracies of both the screening and reference test can be determined by an experiment that administers both to members of two populations with different prevalences of the disease. The reason one cannot use just one population is that the results are summarized in a single 2×2 table (disease status as determined by the screening and reference tests) so there are only three "independent" counts. However, there are five unknown parameters, the sensitivity and specificity of each test and the prevalence of the disease. With the two-population design with different prevalences, there are six unknown parameters and six "independent" cell counts so the maximum likelihood equations can be solved. The statistical properties of the resulting estimates are given by Hui and Walter (1980) and by Vacek (1985) under different assumptions on the independence of the classification errors of the two tests. Mantel (1951) discussed the estimation of η when several replications of the test were given to each subject.

The method used in the Appendix to derive the asymptotic normality and variance of $\hat{C}$ also applies to the estimates of η and θ yielded by the two-population experimental design. One must now

include the covariance term

$$\frac{-2\pi(1-\theta)\eta(1-\pi)\mathrm{Cov}(\hat{\eta},\hat{\theta})}{p^2(\eta+\theta-1)^2}.$$

The numerical results in Hui and Walter (1980) suggest that $\mathrm{Cov}(\eta, \theta)$ is relatively small but negative so that the values in the tables are likely to be slight underestimates of the standard error of the asymptotic distribution of $\hat{C}$.

4. APPLICATION TO THE USE OF LIE DETECTORS

In the murder trial of James Frye, *Frye* v. *United States*, 293 F.1013(D.C. Cir. 1923), defense counsel desired to introduce the result of a systolic blood pressure test for deception as it was favorable to his claim that Mr. Frye was innocent. The trial judge did not admit this evidence on the basis that it was not sufficiently reliable. The appeals court affirmed this ruling and set out the following criteria for the admissibility of scientific evidence:

> Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.

Whether the criteria of general acceptance of a scientific principle or technique set forth in the *Frye* case is the most suitable one has been subject to an extensive debate in legal circles (Imwinkelried, 1981; Lacey, 1984; Raskin, 1986). Most commentators agree that the scientific evidence presented to a court should be reliable; the issue is how to determine the degree of reliability required for admissibility. These considerations are not of a purely statistical nature as courts are also concerned with the possibility that the probative or evidentiary value of a scientific nature could be outweighed by the danger of its prejudicing or confusing a jury. This can easily occur with scientific evidence if jurors are unaware of the possible errors inherent in this technique. Focusing on C or the PVP, as well as on the sensitivity and specificity, may help in the determination of whether a technique is sufficiently reliable in a particular case.

Ever since polygraph evidence was not accepted as sufficiently reliable in the *Frye* case, the judicial acceptance of scientific evidence in general and polygraph tests in particular has been somewhat mixed (Lacey, 1984; Imwinkelried, 1981). Recent proposals to use polygraph testing more often to screen current employees and new applicants has renewed interest in and concern about the reliability of the method (Brooks, 1985; Hurd, 1985; Holden, 1986; Simon, 1983) because of a possible high rate of false positive classifications. The statistical principles underlying the analysis of polygraph data follow the framework presented in Section 2. The class D now denotes the set of deceptive persons and S denotes the set of persons the test classifies as being deceptive. The usual analysis of the data and the proper way to describe the results to a jury are described in Raskin (1982, 1986). The analysis assumes that the accuracy rates η and θ are known, while the prevalence π of deceptive persons in the group tested is estimated from the data using (2.3) and the conditional probability $P[D \mid S]$ is obtained from (2.4). Usually the sampling variability inherent in the estimates of π, η and θ are ignored, and we now examine their effect on the data in Raskin (1986) to assess their effect.

The accuracy rates η and θ are estimated from special studies carried out on students or on convicted criminals. Several studies in which a total of 120 guilty and 120 innocent persons were tested by the control question polygraph yielded $\hat{\eta} = .88$ and $\hat{\theta} = .86$, if inconclusive results are classified as errors and $\hat{\eta} = .97$ and $\hat{\theta} = .92$ if inconclusive results are ignored. Since firms may not hire persons in the inconclusive category we will use $\hat{\eta} = .88$ and $\hat{\theta} = .86$ in our analysis. Moreover, these results are somewhat superior to the accuracies of $\eta = .89$ and $\theta = .80$ reported as typical in a review by the Office of Technology Assessment (1983).

Over a 12-year period, Raskin (1986) conducted 292 polygraph tests on a confidential basis for attorneys defending suspected criminals and classified 193 or 66.1% of them as deceptive. Setting $p = .66$, $\eta = .88$ and $\theta = .86$ in equation (2.4) yields $\hat{C} = .93$ with a standard error of .01, if η and θ are assumed known and .02 when the sampling variability of $\hat{\eta}$ and $\hat{\theta}$ is incorporated. As before we see that the effect of estimation of the parameters is π, η and θ is small when a group with a high prevalence is screened.

In Table 2 we report the expected value of $\hat{C}$ and its standard error for various values of the prevalence rate, emphasizing low rates. The trend in the results is similar to that in Table 1. However, due to the lower accuracy rates of the polygraph, at every prevalence rate $\hat{C}$ is smaller and its standard error is higher. In addition, the smaller sample of persons used to determine the accuracies of the polygraph further increased the standard error of $\hat{C}$ when the estimation of η and θ was considered. When the prevalence π of deceptives

TABLE 2

Approximate standard errors of the estimated true positive rate $\hat{C}$ when $\hat{\eta} = .88$, $\hat{\theta} = .86$ if they are known or are estimated from samples of size $n_1 = n_2 = 120$ as a function of the prevalence π and the size n of the population screened

Prevalence π	Expected fraction p of persons classified D	$E(\hat{C})$	$n = 292$			$n = 10{,}000$		
			Standard error if η, θ are known	Standard error if η, θ are estimated	Percentage of the variance of $\hat{C}$ due to estimation	Standard error if η, θ are known	Standard error if η, θ are estimated	Percentage of the variance of $\hat{C}$ due to estimation
.50	.510	.863	.0187	.0418	78.2	.0032	.0375	96.7
.40	.436	.807	.0254	.0580	78.9	.0043	.0523	98.2
.20	.288	.611	.0532	.1174	79.4	.0091	.1051	99.1
.10	.214	.411	.0873	.1809	76.8	.0149	.1591	99.1
.05	.177	.249	.1187	.2344	74.4	.0203	.2032	99.0
.03	.162	.163	.1365	.2634	73.1	.0233	.2265	98.9
.01	.147	.060	.1590	.2988	71.7	.0272	.2544	98.9

TABLE 3

The expected value and standard error of the true positive rate for a polygraph with accuracy rates $\eta = .88$, $\theta = .86$ when they are known and when they are estimated from a sample of 960 persons and 10,000 people are tested[a]

Prevalence π	$E(\hat{C})$	Standard error of $\hat{C}$ if η, θ are known	Standard error of $\hat{C}$ if η, θ are estimated for three allocations of the subjects		
			$n_1 = 840$ $n_2 = 840$	$n_1 = 240$ $n_2 = 720$	$n_1 = 120$ $n_2 = 840$
.50	.863	.0032	.019	.0159	.0153
.40	.807	.0043	.026	.0219	.0207
.20	.611	.0091	.053	.0438	.0407
.10	.411	.0149	.0806	.0664	.0618
.05	.249	.0203	.1031	.0850	.0791
.03	.163	.0233	.115	.0949	.0883
.01	.060	.0272	.1294	.1068	.0994

[a] The standard errors were calculated from formula (2.5) with $n = 10{,}000$ and the values of n_1 and n_2 specified in the appropriate column.

was less than .10, the standard error was at least half of the expected PVP. Thus, our analysis supports the skepticism of Brooks (1985) and other legislators and psychologists (Holden, 1986) about the reliability of routine polygraph examinations. With prescreening, e.g., when π is at least .5, the proportion $(1 - \hat{C})$ of false positive classifications that are erroneous is reduced to about 10% and its standard error is small.

Because of the lower specificity and sensitivity of the polygraph relative to the ELISA test for AIDS antibodies, we explored the potential for reducing the standard error of $\hat{C}$ by changing the relative proportion of the sample sizes n_1 and n_2 from 1:1 to 1:3 or 1:7. Table 3 presents the expected value of $\hat{C}$ and its standard error for each of the three allocations of a total of 960 subjects used to obtain the accuracy rates for a polygraph test subsequently applied to a group of $n = 10{,}000$ people. We assume that $\eta = .88$ and $\theta = .86$, as in Table 2, and note that although the expected value of $\hat{C}$ is the same in all cases there is an appreciable (25%) reduction in the standard error in the low prevalence situation. However, the results in Table 3 remind us that more precise determination of the accuracy rates will not reduce the high fraction of false positive classifications that are expected to occur when a moderately accurate screening device is used in a population with a low prevalence of the characteristic for which is it screened.

5. OTHER ISSUES AND IMPLICATIONS

As the same statistical paradigm can be used to analyze urinalysis tests for drug use or preemployment tests to determine whether a job applicant possesses sufficient knowledge or skills to perform the job, our results have wide applicability. They reinforce the known statistical fact that the prevalence of the trait in the population tested is a major factor in the determination of the PVP and PVN (Bross, 1954; Fleiss, 1981) with an increased standard error of the estimated PVP in populations with a low prevalence. Before examining further applications we need

to address the similarities and differences in the analysis of polygraph and medical screening data.

In order to estimate the accuracy of the polygraph, Raskin (1986) combined the results of several smaller studies which appeared mutually consistent in the sense that the accuracies differed from one another by amounts which could arise by chance. Since experimental or field conditions can vary substantially, the χ^2 test of homogeneity (Mosteller and Rourke, 1973) should be used to check that the sample proportions are not too diverse. Indeed, the studies listed in Saxe, Dougherty and Cross (1985) show a wide variation in the estimates of η and θ obtained by different investigators.

The same problem also arises in the determination of the specificity and sensitivity of the ELISA test. The total sample of 297 healthy donors consisted of a group of 228 persons from Vermont and 69 from Denmark. The results are given in Table 4. Using the normal approximation to the Mann-Whitney form of the Wilcoxon test yields the value $Z = 5$ showing that the distribution of mean absorbance ratios (the basis of the ELISA test) in the Vermont donors is significantly greater than that of the Danish donors. Thus, it is not proper statistical procedure to pool the two samples. If we only consider the Vermont data, the estimated specificity becomes $^{207}/_{288} = .904$. This difference reduces the PVP and reinforces the concerns about the widespread use of screening tests without confirmatory ones. For example, when $\pi = .05$, the expected value $E(\hat{C})$ is .348 in contrast with .410 in Table 1, and when $\pi = .01$, $E(\hat{C})$ is only .093.

Similarly, the sensitivity of the ELISA test may depend on whether or not the AIDS patient has Kaposi's sarcoma (KS). Of the sample of 88 persons used to determine η, 51 had KS. The ELISA ratios of the KS and non-KS patients are also reported in Table 4 and the same version of the Wilcoxon test yields the value $Z = 3.6$ indicating that the ELISA ratio units of AIDS patients with KS are higher than non-KS patients. Weiss et al. (1985) are aware of this, as they remarked that KS patients had more extremely high ratios but that the proportions of both groups with ELISA ratios of 3.0 or more (the cut-off point for a borderline classification) were not statistically significantly different. (*Note*: In response to an earlier version of this article Dr. S. Weiss kindly called my attention to the recent work of Blaser, Cohn, Cody, Penley, Judson, Saxinger and Weiss (1986). In order to assess the usefulness of repeating the ELISA test, these investigators restandardized the cut-off values using the data on the 228 healthy blood donors from Vermont and set the criteria for a positive classification at 8.0 ratio units or more, borderline as 4.5 to 7.99 ratio units and negative as less than 4.5 units.)

Another potential issue arises when the group used to determine the accuracy rates differs from the population on which the test will be used. Raskin (1986) obtained his accuracy rates from mock crime studies, often on subjects from prison populations, who should have the demographic and other relevant background characteristics of persons accused of crimes. This may no longer be the case if the same polygraph test is used on employees of a bank. Thus, the accuracy rates need to be verified on a population similar to the one on which it will be used.

Similar problems arise in determining the sensitivity and specificity of the ELISA test. The sensitivity is determined from AIDS patients. In view of the differential response between KS and non-KS patients, asymptomatic carriers of AIDS antibodies might have a somewhat different distribution of ELISA ratios than either category of AIDS patients. At first one might believe that the specificity of the test could be determined from a larger sample of well individuals; however, the difference between the two groups of blood donors indicates that there may be some real biological differences among healthy populations. Differences in the accuracies of various versions of ELISA test kits (Kuritsky, Rastogi, Faich, Schoor, Menitove, Reilly and Bove, 1986) also affect their use by blood banks. Indeed, screening tests may suffer from a decreased performance level in the field relative to the laboratory. Moreover, the tests may

TABLE 4

Number of individuals having a given mean absorbance ratio in the ELISA for HTLV antibodies[a]

Group	Sample size	Number of individuals in ratio range							Median ratio
		<2	2–2.99	3–3.99	4–4.99	5–5.99	6–11.99	12+	
Healthy Donors									
Vermont	228	134	72	15	3	2	2	0	1.86
Denmark	69	68	1	0	0	0	0	0	.61
AIDS patients									
With KS	51	0	1	3	4	2	22	19	9.12
Without KS	37	0	1	4	3	13	14	2	5.76

[a] Adapted from Table 1 of Weiss et al. (1985).

respond to factors other than the one screened for. For instance, the ELISA test detects HLA antibodies as well as those for the HTLV-III (AIDS) (Sayers, Beatty and Hanson, 1986) and the polygraph may reflect a general level of stress. The sensitivity of the polygraph may be diminished as there are ways deceptive persons can use to defeat it (Raskin, 1986).

Most medical discussions of the effect of false positive and false negative classifications (see Sackett, Haynes and Tugwell, 1985) evaluate the decision to give further tests or treatment in terms of maximizing expected utility. Typically the consequences of a false positive are less serious than those of a false negative as the error is quite likely to be corrected when the confirmatory test is given. Thus, the false negative rate $(1 - \hat{F})$ has received more attention in the biostatistical literature concerned with periodic screening programs (Goldberg and Wittes, 1978; Zelen and Feinleib, 1969). Moreover, the dependence of PVP and PVN on the prevalence of the characteristic in the population to be screened may render them less useful than specificity and sensitivity in determining the cutoff value for the classification (DeLong, Vernon and Bollinger, 1985).

Since medical screening tests are often given to persons who have some risk factors or symptoms of a disease and employment tests are given to persons who apply for a job and presumably believe they possess the requisite skills, screening tests typically are given to populations with a prevalence higher than that of the general population. The increased use of screening tests and related procedures on populations with a low prevalence rate motivated our analysis emphasizing $\hat{C}$, the estimated PVP, rather than the estimated PVN.

To illustrate the potential problem, James and Morgenstern (1985) reported that 5 of 100 United States Army blood donors tested positive on the ELISA test for AIDS antibodies, but only one of them was confirmed by the Western blot test. Since the prevalence of AIDS antibodies in the Army is undoubtedly less than .01, the 4 out of 5 misclassifications from a single screen may well be an underestimate. Similar results were reported by Marwick (1985) who noted that only the proportion .002 of slightly over 1.5 million units of donated blood were *repeatedly* positive on the ELISA screening test. Of these units, 2552 were given the reference Western blot test and only 587, or just 23%, were confirmed as having AIDS antibodies. The problem of assuring that an ample supply of disease-free blood is available has many more dimensions; we note that statistical considerations have an important role and support the recommendation of Weiss et al. (1985) that a confirmatory assay be developed and used.

In addition to the possibility of contracting AIDS accidentally by a blood transfusion, public concern with the accuracy of screening tests was increased when President Reagan issued Executive Order 12564 (Labor Law Reporter, 1986a) authorizing mandatory drug screening tests for Federal employees in sensitive positions. The guidelines issued by the Justice Department (Labor Law Reporter, 1986b) stated that confirmatory tests would be required. However, the accuracy of the mass drug tests has been questioned by Altman (1986) who noted that a study by Hanson, Caudill and Boone (1985), which sent blind samples with known amounts of various drugs to laboratories, found that the sensitivity and specificity of the tests were noticeably lower than the values of η and θ on the samples used to certify the testing procedures. For example, only 1 of 11 testing firms met the standard that both η and θ exceed .8. Usually, the specificity remained high but the sensitivity was low (ranging from .31 to .88 depending on the particular substance).

The accuracy of the screening test is quite important even when a confirmatory test is used since confirmatory tests are not always correct. To approximate a confirmatory test we assumed that $\eta = .99$ and $\theta = .98$ and obtained the results corresponding to those in Tables 1 and 2. Assuming that the error rates of the confirmatory and screening tests are independent when the prevalence of true positives in the population (prescreened) to which the confirmatory test is applied was .1, the expected value of $\hat{C}$ was .85 and the expected value of $\hat{F}$ was .0013. Assuming that η and θ were estimated from samples of size 240 and 720, respectively, the approximated standard errors were .04 for $\hat{C}$ and .001 for $\hat{F}$. These results suggest that using a confirmatory test will mitigate but not eliminate the false positive problem. However, the screening test has to be sufficiently accurate so that the proportion of true positives among the positive classifications is .10 (and certainly greater than .05). When screening a population with low prevalence ($\pi \leq .01$), this condition needs to be checked carefully. This is especially true for testing employees in sensitive jobs as they often have been subject to a personnel investigation so the prevalence of drug use among them may be well below the rate of all employees. Ironically, the fraction of false positive classifications and the standard error of the estimated PVP will be *larger* for this population than the general workforce.

Fortunately, the field performance of the ELISA test for AIDS antibodies has been as good as the levels indicated by the NIH study, so the growth in transfusion-related cases has leveled off (Marwick, 1985). This is due to the high accuracy of the ELISA test, the very high accuracy of the Western blot test and the fact that high risk groups have been screened

out of the donor pool by self-selection and by confidential questionnaires (Nusbacher, Chiavetta, Naiman, Buchner, Scalia and Horst, 1986).

The accuracy of drug screening tests has played a role in legal cases. In *Capua* v. *City of Plainfield* CA-86-2992 (D.C.N.J. 1986), Judge Sarokin reviewed drug testing cases and decided that random testing is not permissible and that there should be probable cause before an employee is tested. In statistical terms, probable cause increases the prevalence or prior probability π and will increase the expected value of $\hat{C}$ and decrease its standard error.

The potential harm from errors in polygraph tests has been discussed by Hurd (1985) who noted that the problem is compounded by the fact that these results are not confidential (i.e., restricted to the employer giving the test) and the fact that legal recourse for the individual when the employer gives out the results to other potential employers often is unavailable. Again statistical analysis focuses our attention on the high expected fraction of false positives $(1 - C)$ in Table 2 when the prevalence is low as well as on the high standard error of an individual diagnosis. Hopefully, polygraphers and employers will consider these results in their interpretation of the results of screening and polygraph tests.

APPENDIX: OUTLINES OF THE PROOFS

We now provide the mathematical basis of our assertion that $\hat{\pi}$ and $\hat{\pi}_1$ are asymptotically equivalent and the derivation of formula (2.5).

We first prove

LEMMA A.1. *If* $\eta > \frac{1}{2}$, $\theta > \frac{1}{2}$ *and* $\pi > 0$, *then* $\sqrt{n}(\hat{\pi} - \pi)$ *and* $\sqrt{n}(\hat{\pi}_1 - \pi)$ *have the same limiting distribution.*

PROOF. From (2.3) it follows that $\hat{\pi} = \hat{\pi}_1$ whenever $0 < (\hat{p} + \theta - 1) \leq (\eta + \theta - 1)$, i.e., when $1 - \theta < \hat{p} < \eta$. Recall that $\hat{p}$ is the maximum likelihood estimate of $p = \pi\eta + (1 - \pi)(1 - \theta)$ and that $1 - \theta < p < \eta$ since $\eta > 1 - \theta$ and $\pi > 0$. From Bernstein's inequality

$$P[|\hat{p} - p| > \varepsilon] \leq 2e^{-2n\varepsilon^2}.$$

Therefore, for any $\varepsilon < \min[\eta - p, p - (1 - \theta)]$, $\hat{\pi} - \hat{\pi}_1 = 0$ except for a set of probability no greater than $2e^{-2n\varepsilon^2}$. Hence, $P[|\sqrt{n}(\hat{\pi}_1 - \hat{\pi})| > 0] \leq 2e^{-2n\varepsilon^2}$ so that $\sqrt{n}(\hat{\pi}_1 - \pi) - \sqrt{n}(\hat{\pi} - \pi)$ converges to zero in probability and the result follows.

REMARK. The condition $\eta > \frac{1}{2}$, $\theta > \frac{1}{2}$ holds for about any useful screening test (Bross, 1954; Goldberg, 1975). When π is small, η and θ are near 1 and $\pi > (1 - \theta)$, a large n may be needed before the asymptotic results are valid.

We next state

THEOREM A.1. *Let the accuracy rates* α *and* β *be estimated by the proportions* $\hat{\eta}$, $\hat{\theta}$ *of correct classifications in samples of size* n_1 *and* n_2 *from persons in class* D *and* $\bar{D}$, *respectively. Let* π, p *and* $\hat{p}$ *be as described earlier. Then, as* n, n_1 *and* n_2 *increase, the sampling distribution of* $\hat{C}$ *tends to a normal distribution with mean* $P[D \mid S]$ *and variance* (2.5).

PROOF (OUTLINE). The formal development uses the delta method (Rao, 1973), which is based on expanding the estimate of $\hat{C}$ around the parameter it is estimating and using the fact that $\hat{\eta}$ $\hat{\theta}$ and $\hat{p}$ are independent sample proportions with respective means η, θ and p and variances $\eta(1 - \eta)/n_1$, $\theta(1 - \theta)/n_2$ and $p(1 - p)/n$. As sample averages they obey the central limit theorem. The Taylor expansion shows that the difference between

$$\text{(A.1)} \qquad \hat{C} = \frac{\hat{\eta}}{\hat{\eta} + \hat{\theta} - 1}\left[\frac{\hat{p} - (1 - \hat{\theta})}{\hat{p}}\right] = g(\hat{\eta}, \hat{\theta}, \hat{p})$$

and $P[D \mid S]$ is representable as

$$\text{(A.2)} \qquad \left.\frac{\partial g}{\partial \hat{\eta}}\right|_{(\eta,\theta,p)}(\hat{\eta} - \eta) + \left.\frac{\partial g}{\partial \hat{\theta}}\right|_{(\eta,\theta,p)}(\hat{\theta} - \theta) + \left.\frac{\partial g}{\partial \hat{p}}\right|_{(\eta,\theta,p)}(\hat{p} - p) + o_p(n^{-1/2}).$$

Formal calculation and evaluation of the three derivatives in (A.2) yields

$$\left.\frac{\partial g}{\partial \eta}\right|_{(\eta,\theta,p)} = \frac{p + \theta - 1}{p}\,\frac{(\theta - 1)}{(\eta + \theta - 1)^2},$$

$$\left.\frac{\partial g}{\partial \theta}\right|_{(\eta,\theta,p)} = \frac{\eta}{p}\,\frac{(\eta - p)}{(\eta + \theta - 1)^2}$$

and

$$\left.\frac{\partial g}{\partial p}\right|_{(\eta,\theta,p)} = \frac{(1 - \theta)}{(\eta + \theta - 1)p^2}.$$

The result follows from the fact that expression (A.2) is a linear combination of three independent sample averages with variance equal to the sum of the individual variances.

ACKNOWLEDGMENTS

This research was initiated when I was a Visiting Scholar in the Division of Statistics, University of

California at Davis, on a Guggenheim Foundation Fellowship and completed with the support of a grant from the National Science Foundation. It is a pleasure to thank the Division for their hospitality and support and to acknowledge helpful conversations with Professors P. K. Bhattacharya and W. Johnson of the University of California at Davis, S. W. Greenhouse and P. Hammick of George Washington University and Dr. M. H. Gail of the National Cancer Institute.

REFERENCES

ALTMAN, L. K. (1986). Expensive drug tests often inaccurate. *The New York Times*, Sept. 16.

BARNES, D. M. (1986). Keeping the AIDS virus out of the blood supply. *Science* **233** 514–515.

BLASER, M. J., COHN, D. L., CODY, H. J., PENLEY, K. A., JUDSON, F. N., SAXINGER, W. C. and WEISS, S. H. (1986). Counter-immunoelectrophoresis for detection of human serum antibody to HTLV-III. *J. Immunol. Methods* **91** 181–186.

BROOKS, J. (1985). Polygraph testing: thoughts of a skeptical legislator. *Amer. Psychol.* **40** 348–354.

BROSS, I. (1954). Misclassifications in 2 × 2 tables. *Biometrics* **10** 478–486.

DELONG, E., VERNON, W. B. and BOLLINGER, R. R. (1985). Sensitivity and specificity of a monitoring test. *Biometrics* **41** 947–958.

FINKELSTEIN, M. O. (1978). *Quantitative Methods in Law*. The Free Press, New York.

FLEISS, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd ed. Wiley, New York.

GART, J. J. and BUCK, A. A. (1966). Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *Amer. J. Epidemiol.* **83** 593–603.

GOLDBERG, J. D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *J. Amer. Statist. Assoc.* **70** 561–567.

GOLDBERG, J. D. and WITTES, J. T. (1978). The estimation of false negatives in medical screening. *Biometrics* **34** 77–86.

GREENBERG, R. A. and JEKEL, J. F. (1969). Some problems in the determination of the false positive and false negative rates of tuberculin tests. *Amer. Rev. Respir. Dis.* **100** 645–650.

HAMMICK, P. (1987). Ph.D. dissertation. Dept. Statistics/C&IS, George Washington Univ. To be submitted.

HANSON, H. J., CAUDILL, S. P. and BOONE, J. (1985). Crisis in drug testing: results of a CDC blind study. *J. Amer. Med. Assoc.* **253** 2382–2387.

HOLDEN, C. (1986). Days may be numbered for the polygraphs in the private sector. *Science* **232** 705.

HUI, S. L. and WALTER, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics* **36** 167–171.

HURD, S. E. (1985). Use of the polygraph in screening job applicants. *Amer. Bus. Law J.* **22** 529–549.

IMWINKELRIED, E. J. (1981). A new era in the evolution of scientific evidence—a primer on evaluating the weight of scientific evidence. *William and Mary Law Rev.* **23** 261–290.

INGELFINGER, J. A., MOSTELLER, F., THIBODEAU, L. A. and WARE, J. H. (1983). *Biostatistics in Clinical Medicine*. MacMillan, New York.

JAMES, J. J. and MORGENSTERN, M. A. (1985). HTLVIII antibodies in U. S. blood donors in West Germany. *J. Amer. Med. Assoc.* **254** 1449.

KURITSKY, J. N., RASTOGI, S. C., FAICH, G. A., SCHOOR, J. B., MENITOVE, J. E., REILLY, R. W. and BOVE, J. R. (1986). Results of nationwide screening of blood and plasma for antibodies to HTLV-III. *Transfusion* **26** 205–207.

LACEY, F. B. (1984). Scientific evidence. *Jurimetrics J.* **24** 254–272.

MANTEL, N. (1951). Evaluation of a class of diagnostic tests. *Biometrics* **7** 240–246.

MARWICK, C. (1985). Blood banks give HTLV-III testing positive appraisal after 5 months. *J. Amer. Med. Assoc.* **254** 1681–1683.

MOSTELLER, F. M. and ROURKE, R. E. K. (1973). *Sturdy Statistics*. Addison-Wesley, Reading, Mass.

NUSBACHER, J., CHIAVETTA, J., NAIMAN, R., BUCHNER, B., SCALIA, V. and HORST, R. (1986). Evaluation of a confidential method of excluding blood donors exposed to human immunodeficiency virus. *Transfusion* **26** 539–541.

OFFICE OF TECHNOLOGY ASSESSMENT. (1983). *Scientific Validity of Polygraph Testing: A Research Review and Evaluation*. U.S. Government Printing Office, Washington, D. C.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

RASKIN, D. C. (1982). The scientific basis of polygraph techniques and their use in the judicial process in reconstructing the past. In *The Role of Psychologists in Criminal Trials* (A. Trankell, ed.). Norstedt and Soners, Stockholm.

RASKIN, D. C. (1986). The polygraph in 1986: Scientific, professional and legal issues surrounding application and acceptance of polygraph evidence. *Utah Law Rev.* **1986** 29–74.

ROGAN, W. J. and GLADEN, B. (1978). Estimating prevalence from the results of a screening test. *Amer. J. Epidemiol.* **107** 71–76.

SACKETT, D. L., HAYNES, R. B. and TUGWELL, P. (1985). *Clinical Epidemiology*. Little, Brown, Boston.

SAXE, L., DOUGHERTY, D. and CROSS, T. (1985). The validity of polygraph testing: Scientific analysis and public controversy. *Amer. Psychol.* **40** 355–366.

SAYERS, M. H., BEATTY, P. G. and HANSON, J. A. (1986). HLA antibodies are a cause of false positive reactions in screening enzyme immunoassays for antibodies to HTLV-III. *Transfusion* **26** 113–115.

SIMON, M. A. (1983). Shall we ask the lie detector? *Sci. Technol. Human Values* **8** 3–13.

SOLOMON, H. (1966). Jurimetrics. In *Research Papers in Statistics: Festschrift for J. Neyman* (F. N. David, ed.). Wiley, New York.

VACEK, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* **41** 959–968.

WEISS, S. H., GOEDERT, J. J., SARNGADHARAN, M. G., BODNER, A. J., THE AIDS SEREOEPIDEMIOLOGY WORKING GROUP, GALLO, R. C. and BLATTNER, A. (1985). Screening test for HTLV-III (AIDS agent) antibodies. *J. Amer. Med. Assoc.* **253** 221–225.

ZELEN, M. and FEINLEIB, M. (1969). On the theory of screening for chronic diseases. *Biometrika* **56** 601–614.