

COMPARING NEW AND OLD SCREENING TESTS WHEN A REFERENCE PROCEDURE CANNOT BE PERFORMED ON ALL SCREENEES

EXAMPLE OF AUTOMATED CYTOMETRY FOR EARLY DETECTION OF CERVICAL CANCER

ARTHUR SCHATZKIN, ROBERT J. CONNOR, PHILIP R. TAYLOR, AND BILL BUNNAG

Schatzkin, A. (NCI, Bethesda, MD 20892-4200), R. J. Connor, P. R. Taylor, and B. Bunnag. Comparing new and old screening tests when a reference procedure cannot be performed on all screenees: example of automated cytometry for early detection of cervical cancer. *Am J Epidemiol* 1987;125:672-8.

Direct determination of the sensitivity and specificity of a screening test requires use of a reference procedure (such as biopsy with histopathologic analysis) that provides an estimate of true disease status. The authors present a method for comparing the accuracy of a new screening test to an old one in situations when it is not feasible to apply the reference procedure to all screenees. This method requires that only those persons who test positive on old or new screening tests be further evaluated with the reference procedure. Ratios of sensitivities and specificities are derived for rapid comparison of the two screening tests. It is shown that McNemar's test can be used for significance testing of the differences in sensitivities and specificities between two screening tests. The required sample size for a study that compares the two tests is determined.

cervix neoplasms; cytological technics; flow cytometry; mass screening

How does one determine whether a proposed new screening test is superior to an established test? In this paper, we examine a common practical form of this general problem. We will focus, for purposes of example, on a proposed new procedure for early detection of cervical cancer.

Screening for early detection of cervical cancer has become an accepted part of medical practice. To our knowledge, a controlled clinical trial of Papanicolaou (Pap) smear screening has never been conducted. However, several studies in different countries have suggested that Pap smear programs contributed substantially to reductions in morbidity and mortality associated with cervical cancer (1, 2). Automated cytometry systems have recently been proposed to replace manual reading by cytotechnologists of the standard Pap smear (3). The potential advantages of the automated approach include reduction of false readings that result from fatigue and other human error, compensation for the recent decline in the number of newly-certified cytotechnologists (personal communication, Ann H. Clark, Medical U. of South Carolina, 1986), and an increase in the uniformity and quality of cervical cancer screening in those laboratories with relatively poor quality control. The basic issue is whether or not to replace the old test with the new.

Received for publication February 7, 1986 and in final form July 21, 1986.

Abbreviation: Pap smear, Papanicolaou smear.

From the Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.

Send reprint requests to Dr. Arthur Schatzkin, National Institutes of Health, National Cancer Institute, Blair Building, Room 6A01, 9000 Rockville Pike, Bethesda, MD 20892-4200.

The quest
case, automa
This questio
is being comp
and specific
procedure. E
standpoints.
the manual
increase the
the part of t
cases); this is
delays in tre
must evaluat
parameters is
The sensit
test results a
state. Diseas
atypia. One v
with histopat
procedure is
curettage of
ethical diffic

Our appro
performing b
the true disea
tables show
ignored for
estimated by
because we d
calculated. R

Test
result

+

-

* Brackets in

THE PROBLEM

The question immediately arises whether the accuracy of the new technique (in this case, automated reading) is comparable to that of the old technique (manual reading). This question is potentially relevant to any screening situation in which a new technique is being compared to an older, established one. One would like to know how the sensitivity and specificity of the new, automated procedure compare with those of the manual procedure. Each of these parameters is important from both clinical and economic standpoints. If the automated procedure were to have a markedly lower specificity than the manual procedure, then the greater proportion of resulting false positives would increase the number of diagnostic work-ups that would occur. A reduced sensitivity on the part of the new procedure would increase the proportion of false-negatives (missed cases); this is particularly serious for conditions in which early treatment is effective and delays in treatment are potentially dangerous. In comparing the two procedures, one must evaluate both sensitivity and specificity to assure that an increase in one of these parameters is not outweighed by a reduction in the other.

The sensitivity and specificity of the test are determined by a comparison between the test results and a reference procedure (or "gold standard") that ascertains the true disease state. Disease in this case can refer to the presence of either cervical cancer or cellular atypia. One valid reference procedure for cervical disease is cone biopsy or hysterectomy with histopathologic review of serial sections of the biopsy specimen. Another less drastic procedure is multiple random punch biopsies of the cervix under culposcopic control with curettage of the endocervical canal. Each of these procedures would present clinical and ethical difficulties if it were proposed for use on a series of apparently disease-free women.

DESCRIPTION OF THE METHOD

Our approach for evaluation of the new automated screening procedure is based on performing both tests on each woman in a series. The results of such an experiment, if the true disease status for all the women were known, are represented in the contingency tables shown in table 1. (Self-matching, whereby each woman receives both tests, is ignored for the moment.) Here the sensitivities of the tests, Π_1 and Π_2 , would be estimated by a'/n_1 and a''/n_1 , the specificities, Θ_1 and Θ_2 , by d'/n_2 and d''/n_2 . But because we do not know the true disease state of all women, these estimates cannot be calculated. Recall that in this setting only women who test positively on either test are

TABLE 1
 Generalized contingency tables for two screening tests*

		a Manual test		b Automated test			
		Disease status		Disease status			
		D	\bar{D}	D	\bar{D}		
Test result	+	a'	b'	a''	b''		
	-	[c']	[d']	[c'']	[d'']		
		[n ₁]	[n ₂]	[n ₁]	[n ₂]	N	

* Brackets indicate unknown value.

WHEN A
 MED

TION OF

HILL BUNNAG

aylor, and
 procedure
 ry for early

ening test
 pathologic
 rs present
 old one in
 creenees.
 old or new
 Ratios of
 wo screen-
 testing of
 ning tests.
 etermined.

ning

s superior to an
 a of this general
 procedure for early

nd part of medical
 ou (Pap) smear
 at countries have
 ons in morbidity
 ry systems have
 of the standard
 ude reduction of
 ensation for the
 nal communica-
 n the uniformity
 vely poor quality
 e new.

Arthur Schatzkin,
 ational Cancer Insti-
 9000 Rockville Pike,

TABLE 2
Generalized contingency tables for comparing two screening tests—by disease status*

		a Diseased		b Nondiseased	
		Test II		Test II	
		+	-	+	-
Test I	+	a ₁	b ₁	a ₂	b ₂
	-	c ₁	[d ₁]	c ₂	[d ₂]
		[N _D]		[N _{\bar{D}}]	

* Brackets indicate unknown value.

Note that these estimates differ only if the observed number of discordant results b_1 and c_1 are unequal. This is reflected in the McNemar test statistic, which, with the usual correction for continuity, is:

$$\chi_c^2 = \frac{(|b_1 - c_1| - 1)^2}{b_1 + c_1}$$

This is a single degree of freedom chi-square (9).

To test the null hypothesis that there is no difference between the specificities of the two screening tests, McNemar's test can again be used. The chi-square statistic (one degree of freedom), with the continuity correction, is:

$$\chi_c^2 = \frac{(|b_2 - c_2| - 1)^2}{b_2 + c_2}$$

Again, only discordant pairs are involved in the statistical test.

Note that if the new test is believed to be superior to the old one, so that both its sensitivity and specificity are greater, the individual tests for sensitivity and specificity can be combined into one global test. The resulting test statistic is a chi-square with one degree of freedom and is given by:

$$\chi_c^2 = \frac{[|(b_1 + c_2) - (b_2 + c_1)| - 1]^2}{b_1 + c_2 + b_2 + c_1}$$

AN EXAMPLE

Because a study comparing automated to manual screening for cervical cancer has not, to our knowledge, been performed, we will apply our approach to data from the Health Insurance Plan Study (10). It should be recognized that these data are presented for illustrative purposes only, and should not be taken as a substantive commentary on breast cancer screening procedures (particularly in light of technical advances that have been made in the last two decades).

Table 3 presents mammography and physical examination findings for 307 women who received biopsies as reported in the Health Insurance Plan Study (10). There were 55 cases detected on biopsy, with 252 women having biopsies negative for breast cancer. The brackets in the lower right cells of tables 3a and b indicate that we do not have biopsy data on persons with negative results on both screening tests.

TABLE 3
Data from the Health Insurance Plan Study (10) used in discussion of McNemar's test for comparison of two screening tests

		a Diseased			b Nondiseased		
		Mammography			Mammography		
		+	-		+	-	
Physical exam	+	10	24	34	13	144	157
	-	21	[?]		95	[?]	
		31		N_D	108		N_B

The sensitivities of the two tests can be compared by applying the McNemar test statistic applied to table 3a:

$$\chi_c^2 = \frac{(|24 - 21| - 1)^2}{45} = 0.089.$$

The sensitivities of the two tests do not differ significantly ($p = 0.76$).

For specificities, the McNemar test applied to table 3b yields:

$$\chi_c^2 = \frac{(|144 - 95| - 1)^2}{239} = 9.6.$$

The difference in specificities is statistically significant ($p = 0.002$). We would conclude from these data that mammography is more specific than physical examination.

SAMPLE SIZE ESTIMATION

Having shown how the McNemar test is used to test the sensitivity (or specificity) of two screening tests when only "positives" are evaluated, we now consider the issue of overall sample size for such a study. That is, how many asymptomatic people should be entered into the study and screened by the two procedures. Our discussion is limited to testing the sensitivities, because in all practical cases the overall sample size required for comparing specificities will be less than that required for comparing sensitivities (11). Moreover, the argument for determining overall sample size for testing specificities is identical to that used for testing sensitivities.

We assume the following conditions: 1) the two screening tests will be performed on all those in the sample but only those with positive tests will be further evaluated; 2) the difference in sensitivities will be tested using the McNemar test with significance level α ; 3) the sensitivity of the first test, Π_1 , is known; 4) the difference, δ , between the sensitivity of the second test, Π_2 , and Π_1 , that we wish to detect is specified in advance; 5) the test is to have a power of $1 - \beta$ of detecting the difference, δ ; and 6) the prevalence, P , of disease in the population being studied is known.

Under these conditions, the overall sample size is calculated in two steps. In the first step, the number of cases N_D required to meet conditions 1-5 above is determined. Miettinen (12) derived an approximate sample size formula for the McNemar test for a specified difference between the proportions when the probability of discordant results

is known. When it is determine sample size sample result (9). The

where $Z_{\alpha/2}$ and Z_{β} are probabilities, $\alpha/2$ and

The second step in certain preselected pr our first step. This is overall sample size be sampling variation as possible that the obser If this were the case, t required to give the d subjects under this cor

Recall that the num normal distribution. T

for N^* , where P is corresponding to the sufficiently large to give the previously specified

To illustrate the two $P = 0.1$, $\Pi_1 = 0.90$, $\delta =$ For a two-tailed test

We now solve for N^*

and

The key to our app testing positive on a biopsies were performed information available would c' , c'' , and d' , d'' . There would be no w proportion of those w insufficient informatio the two tests.

McNemar's test for comparison of two

		b		
		Nondiseased		
Mammography				
+	-			
13	144			157
95	[?]			
108				N_D

Applying the McNemar test

(= 0.76).
As:

(0.002). We would conclude
physical examination.

sensitivity (or specificity) of
now consider the issue of
asymptomatic people should be
our discussion is limited to
all sample size required for
comparing sensitivities (11).
for testing specificities is

tests will be performed on
be further evaluated; 2) the
test with significance level
difference, δ , between the
effect is specified in advance;
e, δ ; and 6) the prevalence,

d in two steps. In the first
1-5 above is determined.
for the McNemar test for a
ability of discordant results

is known. When it is unknown, as is the case here, he gave a bound that can be used to determine sample size. This result is conservatively approximated by the independent sample result (9). Thus, we have:

$$N_D = \frac{(Z_{\alpha/2} + Z_\beta)^2 [\Pi_1(1 - \Pi_1) + \Pi_2(1 - \Pi_2)]}{(\Pi_1 - \Pi_2)^2}, \tag{3}$$

where $Z_{\alpha/2}$ and Z_β are the standard normal deviates corresponding to the one-tailed probabilities, $\alpha/2$ and β , respectively.

The second step involves the calculation of the overall sample size that gives us a certain preselected probability, ϕ , of obtaining the number of cases, N_D , determined in our first step. This is conservative. A less conservative approach would be to let the overall sample size be N_D/P , where P is the true prevalence. However, this ignores the sampling variation associated with the selection of a study population. That is, it is possible that the observed prevalence, P_s , would be less than P , so that $(N \times P_s) < N_D$. If this were the case, then the number of cases observed would be fewer than the number required to give the desired power. The effort and expense of including the additional subjects under this conservative approach should be small.

Recall that the number of successes in a binomial can be approximated by the standard normal distribution. Therefore, to take the sampling variation into account, we solve

$$\frac{N_D - N^*P}{\sqrt{P(1 - P)N^*}} = Z_\phi \tag{4}$$

for N^* , where P is the population prevalence and Z_ϕ is a standard normal deviate corresponding to the upper-tail probability ϕ . This yields a sample size, N^* , that is sufficiently large to give us a probability ϕ of generating enough cases to be able to detect the previously specified difference, δ , with a power of at least $1 - \beta$.

To illustrate the two-step sample size calculation, suppose that $\alpha = 0.05$, $1 - \beta = 0.90$, $P = 0.1$, $\Pi_1 = 0.90$, $\delta = \Pi_1 - \Pi_2 = 0.09$, $\Pi_2 = \Pi_1 - \delta = 0.81$, and $\phi = 0.975$.

For a two-tailed test, we substitute in formula 3 and obtain

$$N_D = \frac{(1.96 + 1.282)^2 (0.252)}{(0.09)^2} = 327.$$

We now solve for N^* in formula 4:

$$\frac{327 - 0.1N^*}{0.3\sqrt{N^*}} = -1.96$$

and

$$N^* = 3,624.$$

DISCUSSION

The key to our approach is performing a reference procedure (biopsy) on persons testing positive on a new screening test even if they test negative on the old test. If biopsies were performed only on persons testing positive on the old test, the amount of information available for comparing the two tests would be severely limited. Not only would c' , c'' , and d' , d'' be unknown (table 1), but a'' and b'' would be unknown as well. There would be no way of determining if the new test actually detected a greater proportion of those with true disease (reduced missed cases), and there would be insufficient information available to quantitatively estimate the ratio of sensitivities of the two tests.

(A minor point is that, to test the difference of sensitivities (or specificities), we need the true disease status only for those with discordant test results. Of course, ethical considerations would mandate further evaluation for those with positive results on both tests.)

The approach presented here is applicable to the valuation of a variety of new screening techniques, for example, HemoQuant versus Hemocult for the early detection of colon cancer (13). We emphasize that our method for comparing screening parameters for two tests does not provide any direct information on the impact of screening using either test on mortality from the disease in question. However, if the old test has been shown through independent studies of screening to be effective in reducing mortality, and the new test has been shown to have greater sensitivity and specificity by our method, then it is likely that the new test will have at least as favorable an impact on mortality as the old test.

In summary, it is possible to make a meaningful comparison between a new screening test and an established one, even in the absence of data from a reference procedure ("gold standard") on all screening participants, as long as persons testing positive on either of the two tests are given the more definitive procedure to determine actual disease status.

REFERENCES

1. Miller AB, Lindsay J, Hill GB. Mortality from cancer of the uterus in Canada and its relationship to screening for cancer of the cervix. *Int J Cancer* 1976;17:602-12.
2. Johansson G, Geirsson G, Day N. The effect of mass screening in Iceland, 1965-74, on the incidence and mortality of cervical carcinoma. *Int J Cancer* 1978;21:418-25.
3. Wheelless LL, Patten SF, Berkan TK, et al. Multidimensional slit-scan prescreening system: preliminary results of a single blind clinical study. *Cytometry* 1984;5:1-8.
4. Buck AA, Gart JJ. Comparison of a screening test and a reference test in epidemiologic studies. I. Indices of agreement and their relation to prevalence. *Am J Epidemiol* 1966;83:586-92.
5. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *Am J Epidemiol* 1966;83:593-602.
6. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1980;36:167-71.
7. Staquet M, Rozenzweig M, Lee YJ, et al. Methodology for the assessment of new dichotomous diagnostic tests. *J Chronic Dis* 1981;34:599-610.
8. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;39:207-15.
9. Snedecor GW, Cochran WG. *Statistical methods*. Ames, IA: Iowa State University Press, 1980.
10. Strax P, Venet L, Shapiro S, et al. Mammography and clinical examination in mass screening for cancer of the breast. *Cancer* 1967;20:2184-8.
11. Cole P, Morrison AS. Basic issues in population screening for cancer. *JNCI* 1980;64:1263-72.
12. Miettinen OS. On the matched pairs design in the case of all-or-none responses. *Biometrics* 1968;24:339-52.
13. Ahlquist DA, McGill DB, Schwartz S. HemoQuant, a new quantitative assay for fecal hemoglobin: comparison with Hemocult. *Ann Intern Med* 1984;101:297-302.

PARTITIONING T

MICHAEL B

Boehnke, M. (P. Moll, B. A. Ke plasma glucose *Epidemiol* 1987; Fasting plasma individuals in the were analyzed. environmental fa To that end, fa inverse normal s and females on height²), season found that 27.7% pedigrees is exp shared environm of the variability ing for at least 2 diabetes, for wh these familial fa normalized faste concomitants.

blood glucose

Fasting plasma glucose the diagnosis of diabetes important predictor of fasting plasma glucose level than tests of glucose tolerance work suggests that

Received for publication final form July 25, 1986.

¹ Department of Biostatistics, The University of Minnesota, Minneapolis, MN 55455. (Send reprint requests to this address.)

² Department of Epidemiology, The University of Minnesota, Minneapolis, MN 55455.

³ Department of Human Genetics, The University of Michigan, Ann Arbor, MI 48109.

⁴ Mayo Clinic and Mayo Foundation, Rochester, MN.