BIOMETRICS 36, 167–171
March, 1980

Estimating the Error Rates of Diagnostic Tests

S. L. Hui[1] and S. D. Walter
Department of Epidemiology and Public Health, Yale University, 60 College Street, New Haven, Connecticut 06510, U.S.A.

SUMMARY

It is often required to evaluate the accuracy of a new diagnostic test against a standard test with unknown error rates. If the two tests are applied simultaneously to the same individuals from two populations with different disease prevalences, then assuming conditional independence of the errors of the two tests, the error rates of both tests and the true prevalences in both populations can be estimated by a maximum likelihood procedure. Generalizations to several tests applied in several populations are also possible.

1. Introduction

The presence of a disease in an individual often cannot be determined with certainty. Diagnostic procedures with high accuracy are obviously desirable but are frequently too expensive or hazardous to be used on a large scale. When investigating disease in large population groups, therefore, a less sophisticated screening test with greater error rates may be preferred.

When a new diagnostic test is developed, its error rates must be determined and weighed against its cost. The false positive (negative) rate $\alpha(\beta)$ is defined as the proportion of nondiseased (diseased) people who have positive (negative) outcomes in the test. Their complements $1-\alpha$ and $1-\beta$ are, respectively, the specificity and sensitivity of the test. Error rates can be estimated directly if the test can be applied to some individuals whose true disease states are known, but this is usually difficult or not feasible. In such cases, the new test is customarily evaluated against a standard test with its own errors by applying both tests simultaneously to each individual.

Greenberg and Jekel (1969) and Gart and Buck (1966) have investigated the effects of known error rates of the standard test on the estimates of the error rates of the new test. These authors point out that the error rates of the standard test, if not taken into account, can lead to biased estimates of the error rates of the new test. Specifically, if the nonzero false positive (negative) rate of the standard test is assumed to be zero then the false negative (positive) rate of the new test will be overestimated. For example, the errors of computer diagnosis are inflated when it is evaluated against a physician's diagnosis which is taken as correct (e.g. Van Meerten, Durinck and Dewit, 1971).

Goldberg (1975) considered the problem where the test error rates are allowed to take different known values in various population groups. Rogan and Gladen (1978) investigated the properties of the estimate of disease prevalence when one diagnostic test is used in a single population, again with the error rates assumed known. In cases where the error

[1] Present address: Epidemiology and Biometry Program, School of Public Health, University of Illinois at the Medical Center, P.O. Box 6998, Chicago, Illinois 60680.

Key words: Diagnosis; Sensitivity; Specificity; Misclassification; Observer agreement.

rates of the standard test are not known, Greenberg and Jekel (1969) suggest that the false positive (negative) rate of a new test be estimated in a population with low (high) prevalence of the disease. Estimates so obtained are only very slightly biased even if the standard test is assumed to be error-free.

If data are available from any two populations with different prevalences, it is shown in this paper that one can then estimate the error rates of both tests and the prevalences in both populations. The populations might in practice be taken as subgroups of a large population of interest, for example males and females or different age groups.

2. Model and Estimation

Let the standard test (Test 1) and the new test (Test 2) be applied simultaneously to each individual in samples from S populations. Let N_g be a fixed sample size, θ_g be the probability of a diseased individual, and α_{gh} and β_{gh} the false positive and false negative rates of test h $(h=1,2)$ in a sample from population g $(g=1,\ldots,S)$, and let the θ_g's all be distinct. Finally, we assume that, conditional on the true disease state, the two tests on each individual are subject to independent errors; this should be reasonable if the tests have unrelated bases, e.g. X-ray versus blood test.

Under these assumptions, the frequencies of the possible test outcomes in the S populations are distributed as S independent multinomials. The likelihood l is

$$\begin{aligned} l = \prod_{g=1}^{S} &\{\theta_g(1-\beta_{g1})(1-\beta_{g2})+(1-\theta_g)\alpha_{g1}\alpha_{g2}\}^{N_g p_{g11}} \\ &\times\{\theta_g(1-\beta_{g1})\beta_{g2}+(1-\theta_g)\alpha_{g1}(1-\alpha_{g2})\}^{N_g p_{g12}} \\ &\times\{\theta_g\beta_{g1}(1-\beta_{g2})+(1-\theta_g)(1-\alpha_{g1})\alpha_{g2}\}^{N_g p_{g21}} \\ &\times\{\theta_g\beta_{g1}\beta_{g2}+(1-\theta_g)(1-\alpha_{g1})(1-\alpha_{g2})\}^{N_g p_{g22}}, \end{aligned}$$

where $p_{gij'}$ is the observed proportion of sample g with test outcomes j and j' in Tests 1 and 2, respectively: $j=1$ (2) if Test 1 is positive (negative) and $j'=1$ (2) if Test 1 is positive (negative).

In general, for R tests applied to S populations, there are $(2^R-1)S$ degrees of freedom for estimating $(2R+1)S$ parameters. Therefore the likelihood with no functional constraint placed on the parameters is over parameterized for R less than three. For R greater than or equal to three, there is no simple closed-form solution for the maximum likelihood estimates but they can be evaluated numerically.

We will now assume that $\alpha_{gh}=\alpha_h$ and $\beta_{gh}=\beta_h$ for all g, in order to study the most important case when $R=2$. In general, the number of parameters then reduces to $2R+S$, which does not exceed the number of degrees of freedom available whenever $S \geqslant R/(2^{R-1}-1)$. Unless equality obtains, the maximum likelihood estimates must be evaluated numerically. In the special case of most practical importance when $R=2$, equality holds with $S=2$. Then the model is saturated with six parameters and the global maximum of the likelihood function occurs at points with the following coordinates:

$$\begin{aligned} \hat{\alpha}_h &= (p_{h1.}p_{\bar{h}.1}-p_{h.1}p_{\bar{h}1.}+p_{211}-p_{111}+D)/2E_h \\ \hat{\beta}_h &= (p_{h.2}p_{\bar{h}2.}-p_{h2.}p_{\bar{h}.2}+p_{122}-p_{222}+D)/2E_h \\ \hat{\theta}_g &= \tfrac{1}{2}+\{p_{g1.}(p_{1.1}-p_{2.1})+p_{g.1}(p_{11.}-p_{21.})+p_{211}-p_{111}\}/2D, \end{aligned}$$

where

$$E_1 = p_{2.1}-p_{1.1}, \qquad E_2 = p_{21.}-p_{11.}$$

and

$$D = \pm\{(p_{11.}p_{2.1} - p_{21.}p_{1.1} + p_{111} - p_{211})^2 - 4(p_{11.} - p_{21.})(p_{111}p_{2.1} - p_{211}p_{1.1})\}^{1/2}$$

are nonzero, $\bar{h} = 2$ (1) if $h = 1$ (2), and . denotes summation over an index. When E_1 and/or E_2 are zero and $D = 0$, some parameters are nonestimable, there being an infinite solution set in this case. When E_1 and/or E_2 are zero but $D \neq 0$, or when D alone is zero, not all of the maximum likelihood equations can be simultaneously satisfied and numerical maximization of the likelihood should be used.

Except when $\hat{\theta}_g = \hat{\alpha}_h = \hat{\beta}_h = \frac{1}{2}$, for g, $h = 1, 2$, two distinct points exist in each solution set given above, such that if $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is a solution so is $(\mathbf{1} - \hat{\boldsymbol{\theta}}, \mathbf{1} - \hat{\boldsymbol{\beta}}, \mathbf{1} - \hat{\boldsymbol{\alpha}})$. Therefore a solution which can be used as an estimate must be defined by appropriate constraints derived from prior knowledge of the error rates of the standard test, e.g. $1 - \alpha_1 - \beta_1 > 0$, which would be a reasonable assumption if the standard test is of practical value (Bross, 1954; Goldberg, 1975). A real solution with coordinates which satisfy these constraints is a maximum likelihood estimate. In other cases when the points of the global maximum are complex or lie outside the unit hypercube, estimates can be obtained by numerical maximization of l with the solution restricted to be within the unit hypercube, again subject to appropriate constraints. Asymptotically, because the observed proportions $p_{gjj'}$ are continuous functions of the parameters, the maximum likelihood solution will converge to $(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

If we define $L = \ln l$, then the information matrix has elements as follows:

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \alpha_h^2}\right) = \sum_g N_g(1-\theta_g)^2 f_{gh}\{\alpha_{\bar{h}}^2, (1-\alpha_{\bar{h}})^2, \alpha_{\bar{h}}^2, (1-\alpha_{\bar{h}})^2\}$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \alpha_h \partial \beta_h}\right) = -\sum_g N_g\theta_g(1-\theta_g) f_{gh}\{\alpha_{\bar{h}}(1-\beta_{\bar{h}}), (1-\alpha_{\bar{h}})\beta_{\bar{h}}, \alpha_{\bar{h}}(1-\beta_{\bar{h}}), (1-\alpha_{\bar{h}})\beta_{\bar{h}}\}$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \alpha_h \partial \alpha_{\bar{h}}}\right) = \sum_g N_g(1-\theta_g)^2 f_{gh}\{\alpha_h\alpha_{\bar{h}}, -\alpha_h(1-\alpha_{\bar{h}}), -(1-\alpha_h)\alpha_{\bar{h}}, (1-\alpha_h)(1-\alpha_{\bar{h}})\}$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \alpha_h \partial \beta_{\bar{h}}}\right) = \sum_g N_g\theta_g(1-\theta_g) f_{gh}\{-\alpha_{\bar{h}}(1-\beta_h), (1-\alpha_{\bar{h}})(1-\beta_h), \alpha_{\bar{h}}\beta_h, -(1-\alpha_{\bar{h}})\beta_h\}$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \theta_g \partial \alpha_h}\right) = N_g f_{gh}\{\alpha_{\bar{h}}(1-\beta_1)(1-\beta_2), (1-\alpha_{\bar{h}})(1-\beta_h)\beta_{\bar{h}}, -\alpha_{\bar{h}}\beta_h(1-\beta_{\bar{h}}), -(1-\alpha_h)\beta_1\beta_2\}$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \beta_h^2}\right) = \sum_g N_g\theta_g^2 f_{gh}\{(1-\beta_{\bar{h}})^2, \beta_{\bar{h}}^2, (1-\beta_{\bar{h}})^2, \beta_{\bar{h}}^2\}$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \beta_h \partial \beta_{\bar{h}}}\right) = \sum_g N_g\theta_g^2 f_{gh}\{(1-\beta_h)(1-\beta_{\bar{h}}), -(1-\beta_h)\beta_{\bar{h}}, -\beta_h(1-\beta_{\bar{h}}), \beta_h\beta_{\bar{h}}\}$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \theta_g \partial \beta_h}\right) = N_g f_{gh}\{\alpha_1\alpha_2(1-\beta_{\bar{h}}), \alpha_h(1-\alpha_{\bar{h}})\beta_{\bar{h}}, -(1-\alpha_h)\alpha_{\bar{h}}(1-\beta_{\bar{h}}), -(1-\alpha_1)(1-\alpha_2)\beta_{\bar{h}}\}$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \theta_g^2}\right) = N_g f_{g1}[\{\alpha_1\alpha_2 - (1-\beta_1)(1-\beta_2)\}^2, \{\alpha_1(1-\alpha_2) - (1-\beta_1)\beta_2\}^2,$$
$$\{(1-\alpha_1)\alpha_2 - \beta_1(1-\beta_2)\}^2, \{(1-\alpha_1)(1-\alpha_2) - \beta_1\beta_2\}^2]$$

$$\mathrm{E}\left(-\frac{\partial^2 L}{\partial \theta_1 \partial \theta_2}\right) = 0,$$

where

$$f_{gh}(a, b, c, d) = a/\bar{p}_{g11} + b/\bar{p}_{gh\bar{h}} + c/\bar{p}_{g\bar{h}h} + d/\bar{p}_{g22}$$

with

$$\bar{p}_{g11} = \theta_g(1-\beta_1)(1-\beta_2) + (1-\theta_g)\alpha_1\alpha_2$$

$$\bar{p}_{gh\bar{h}} = \theta_g(1-\beta_h)\beta_{\bar{h}} + (1-\theta_g)\alpha_h(1-\alpha_{\bar{h}})$$

$$\bar{p}_{g22} = \theta_g\beta_1\beta_2 + (1-\theta_g)(1-\alpha_1)(1-\alpha_2).$$

If the parameters in the above elements are replaced by their estimated values, an estimated information matrix is obtained and this may be inverted numerically to give an estimated asymptotic variance–covariance matrix.

3. Example

In our example, the new Tine test (Test 2) is to be evaluated against the standard Mantoux test (Test 1) for the detection of tuberculosis. Both are skin tests applied to the arms. After 48 hours the presence of an induration larger than a fixed size constitutes a positive result. The criteria were determined from previous experience and they affect the error rates. Data for Population 1 are taken from Greenberg's and Jekel's study in a southern U.S. school district. The original data were reclassified into positive or negative for each test according to the same criteria as those for Population 2 in a second study at the Missouri State Sanatorium (Capobres *et al.*, 1962). The data are presented in Table 1.

The maximum likelihood estimates with their standard errors were found to be

$$\hat{\alpha}_1 = 0.0067 \pm 0.0038 \qquad \hat{\beta}_1 = 0.0339 \pm 0.0069 \qquad \hat{\theta}_1 = 0.0268 \pm 0.0071$$

$$\hat{\alpha}_2 = 0.0159 \pm 0.0056 \qquad \hat{\beta}_2 = 0.0312 \pm 0.0062 \qquad \hat{\theta}_1 = 0.7168 \pm 0.0128.$$

The upper triangle of the variance–covariance matrix and the lower off-diagonal triangle of the correlation matrix are jointly shown as follows:

$$\begin{array}{c} \\ \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\alpha}_2 \\ \hat{\beta}_2 \\ \hat{\theta}_1 \\ \hat{\theta}_2 \end{array} \begin{array}{c} \begin{array}{cccccc} \hat{\alpha}_1 & \hat{\beta}_1 & \hat{\alpha}_2 & \hat{\beta}_2 & \hat{\theta}_1 & \hat{\theta}_2 \end{array} \\ \begin{pmatrix} 1.45\times10^{-5} & -1.41\times10^{-7} & 2.72\times10^{-7} & -6.68\times10^{-6} & -2.07\times10^{-6} & -4.43\times10^{-6} \\ -0.0054 & 4.83\times10^{-5} & -1.37\times10^{-5} & 1.26\times10^{-6} & 2.17\times10^{-6} & 1.00\times10^{-5} \\ 0.0127 & -0.3516 & 3.15\times10^{-5} & -1.40\times10^{-7} & -2.36\times10^{-6} & -0.96\times10^{-5} \\ -0.2820 & 0.0290 & -0.0040 & 3.88\times10^{-5} & 1.86\times10^{-6} & 4.99\times10^{-6} \\ -0.0763 & 0.0438 & -0.0590 & 0.0420 & 5.08\times10^{-5} & 1.56\times10^{-6} \\ -0.0911 & 0.1126 & -0.1338 & 0.0626 & 0.0171 & 1.64\times10^{-4} \end{pmatrix} \end{array}.$$

One check of the validity of the model in the example is to compare our estimates of β_1 and β_2 to those of Capobres *et al.* (1962), who applied the Mantoux and Tine tests to 362 patients with pulmonary tuberculosis proven by positive culture of acid-fast bacilli.

Table 1

Results of Mantoux and Tine tests for tuberculosis in two populations

	Population 1			Population 2		
	Tine test			Tine test		
Mantoux test	Positive	Negative	Total	Positive	Negative	Total
Positive	14	4	18	887	31	918
Negative	9	528	537	37	367	404
Total	23	532	555	924	398	1322

Assuming that the culture has zero false positive rate (i.e. $\alpha_1 = 0$), they estimated false negative rates with standard errors for the Mantoux and Tine tests to be 0.0304 ± 0.0090 and 0.0331 ± 0.0094, respectively. Our estimates 0.0339 ± 0.0069 and 0.0312 ± 0.0062 show satisfactory agreement.

Following Greenberg's and Jekel's (1969) suggestion, we assumed $\alpha_1 = \beta_1 = 0$ and estimated α_2 (β_2) from Population 1 (2) with low (high) disease prevalence. The estimates for the Tine test are $\hat{\alpha}_2 = 0.0168 \pm 0.0055$ and $\hat{\beta}_2 = 0.0338 \pm 0.0050$, compared to $\hat{\alpha}_2 = 0.0159 \pm 0.0056$ and $\hat{\beta}_2 = 0.0312 \pm 0.0062$ from our model. This shows that their method is adequate when the prevalences are as extreme as in this example. When no such populations can be found, our method can still be used as it only requires unequal θ's.

Acknowledgements

We are grateful to Dr J. Jekel of this department for providing the original data from the study in Population 1. The helpful comments of the referees and Editor on an earlier draft are also acknowledged.

Résumé

On a souvent besoin d'évaluer la précision d'un nouveau test diagnostique relativment à un test standard à taux d'erreur inconnu. Si l'on applique simultanément des deux tests aux mêmes individus de deux populations à fréquences de la maladie différentes, on peut estimer, par une procédure du maximum de vraisemblance, les taux d'erreur des deux tests et les vraies fréquences dans les deux populations, en supposant l'indépendance conditionnelle des erreurs des deux tests. On peut aussi généraliser la procédure à plusieurs tests appliqués à plusieurs populations.

References

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

Capobres, D. B., Tosh, F. E., Yates, J. L. and Langeluttig, H. V. (1962). Experience with the tuberculin Tine test in a sanatorium. *Journal of American Medical Association* **180,** 1130–1136.

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. *American Journal of Epidemiology* **83,** 593–602.

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. *Journal of the American Statistical Association* **70,** 561–567.

Greenberg, R. A. and Jekel, J. F. (1969). Some problems in the determination of the false positive and false negative rates of tuberculin tests. *American Review of Respiratory Disease* **100,** 645–650.

Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology* **107,** 71–76.

Van Meerten, R. J., Durinck, J. R. and Dewit, C. (1971). Computer guided diagnosis of asthma, asthmatic bronchitis, chronic bronchitis and emphysema: computing methods and results. *Respiration* **28,** 399–408.

Received March 1978; revised March 1979 and October 1979