Estimation of sensitivity and specificity of site-specific diagnostic tests

P. P. Hujoel[1], L. H. Moulton[1] and W. J. Loesche[2]

[1]Department of Biostatistics, School of Public Health, [2]Department of Biologic and Material Sciences, The University of Michigan, School of Dentistry, Ann Arbor, MI, U.S.A.

Hujoel PP, Moulton LH, Loesche WJ: Estimation of sensitivity and specificity of site-specific diagnostic tests. J Periodont Res 1990; 25: 193–196.

Clinical trials designed to estimate the sensitivity and specificity of periodontal diagnostic tests often use multiple sites per patient as experimental units of analyses. Since site-specific test results within a patient are dependent observations, a correlated binomial model should be employed to estimate the sensitivity and specificity of these diagnostic tests. Ignoring the within-patient correlation can result in an over- or underestimation of the true standard errors.

Introduction

The value of medical diagnostic tests is often assessed according to estimates of their sensitivity and specificity. Sensitivity is the probability that a *person* with a certain condition will be classified by the test as having that condition, and specificity is the probability that a *person* without the condition will be classified by the test as being without the condition. Since both sensitivity and specificity are probabilities, they need to be estimated using certain assumptions. A reasonable assumption for experiments where *persons* are the experimental units of analyses is that responses are independently distributed. Consequently, a probability model which is based on the assumption of independence of observations will be appropriate. However, for site-specific observations the experimental unit of analysis is the *site*, and the definitions of sensitivity and specificity change; sensitivity (specificity) is the probability that a *site* within a patient with(out) the condition will be classified by the test as being with(out) the condition.

Under this definition the assumption of independence becomes questionable since multiple sites within each mouth are the dependent experimental units of analyses (1). Observations on sites within a patient may be correlated and, depending on the magnitude of the correlation and the number of sites per patient sampled, significant under- or overestimation of the true standard errors of sensitivity and specificity can occur. To avoid these pitfalls, site-specific data should be analyzed with a statistical methodology that accounts for the dependence of within-patient observations. We present procedures for estimating sensitivity and speci-ficity under the assumptions: 1) that there are only two outcomes of the diagnostic test: one which is called positive and indicative of the condition; the other called negative and interpreted as not indicative of the condition, and, (b) that site-specific responses within a patient may be correlated. To meet these objectives a modification of the binomial model will be suggested. Methods for estimating sensitivity, specificity, and within-patient correlation coefficients are discussed in general terms and a numerical example is given.

Theory

In clinical experiments designed to investigate the sensitivity and the specificity of diagnostic tests, the response of interest is often binary in nature, namely, the indicated presence or absence of a characteristic. Most commonly these data will be presented in the familiar four-fold table of frequencies resulting from the cross-classification of the test results and the disease status of the patients. The statistical analysis of these types of data requires that the variations in the responses be described by some underlying probability model.

When the observations are independent, the appropriate model for estimating sensitivity and specificity is the one-parameter binomial probability model:

$$P(X = x) = C(n,x) \, p^x q^{n-x} \qquad (1)$$

where $C(n,x)$ represents the number of possible combinations of n persons taken x at a time, x denotes the number of patients with true positive (or true negative) test results, n is the number of diseased (or non-diseased) patients, p is the sensi-

tivity (specificity) parameter, and $q = 1 - p$. A simple estimator of p can be obtained by dividing the number of true positive (or negative) test results x by the number of diseased (or non-diseased) sites n. The variance of the estimated sensitivity (or specificity) is pq/n. For instance, if 85 out of 100 independent tests on diseased persons are true positive results, the sensitivity of the diagnostic test is estimated as $85/100$ or 0.85. The estimated variance of the sensitivity is $(0.85)x(0.15)/100 = 0.001275$.

These calculations need to be adapted when several sites are sampled within a patient. In this sampling scheme, it can no longer be assumed that the true negative or true positive test results are independent observations. Responses on sites within a patient may be correlated and this assumption needs to be incorporated in the statistical probability model. One model that could be selected is the correlated binomial model presented by Bahadur (2). For the ith person of the sample the correlated binomial model (CBM) is written as:

$$P(X = x_i) = C(n_i, x_i)\, p^{x_i}\, q^{n_i - x_i} \qquad (2)$$
$$\{1 + [\rho/(2p(1-p))][(x_i - n_i p)^2 + x_i(2p - 1) - n_i p^2]\}$$

where x_i denotes the number of true positive (or true negative) test results in patient i, n_i is the total number of diseased (or non-diseased) sites in patient i, and ρ is the correlation of the responses within a patient. This model consists of two parts: the first part is the binomial probability model as presented in (1), the second part is a correction factor which adjusts for the dependence of sites within a patient. The result is a two-parameter model: one parameter, p, is the sensitivity or the specificity of the model; the other parameter, ρ, is the within-patient correlation coefficient of sites. For a sample of N persons, the likelihood will be written as $L = \Pi P(x_i)$ (3). Estimates of p and ρ are obtained by maximizing the likelihood function. This maximization procedure is not as straightforward as for the binomial probability model, where a maximum is obtained when $\hat{p} = x/n$. Function maximization computer routines need to be employed for simultaneous estimation of p and ρ. (A copy of a program written in SAS code (3) can be obtained from the authors). For a more in-depth discussion of this probability model the reader is referred to Bahadur (2) and Kupper and Haseman (4).

An applied aspect of this model is that it offers an additional parameter, ρ, which may provide a quantitative measure of the patient-factor effect on the outcome of the diagnostic test. A positive correlation coefficient indicates that if one site tested as a true positive (negative) test result in a patient, the probability that other sites in the same

patient will also test as true positives (negatives) will be increased. A zero within-patient correlation coeficient indicates that one site testing positive (negative) does not affect the probabilities of other sites testing true positives (negatives).

The dependence of site-specific test readings within a patient can result from host-factor effects (intrinsic or extrinsic) or can be due to processing characteristics of the diagnostic test itself. For instance, it may indicate the presence of a patient-specific factor (e.g. colonization with bacteria which interact with the test) which makes it more likely for sites within patients to respond as true positives (negatives), as opposed to false negatives (positives). Or it may also result from a property of the diagnostic test itself. Multiple sites per patient are often tested and processed simultaneously. If the reagents of the test are inactive or improperly handled in some patients, the false negative (positive) results will tend to be correlated resulting in a significant within-patient correlation coefficient.

Example

Consider the following set of data where an enzymatic diagnostic test (5) was employed to determine whether a site was infected by 2 specific organisms. The reference test used to evaluate the enzymatic diagnostic test was an antibody assay against 2 organisms: *Treponema denticola* and *Bacteroides gingivalis*. For the purpose of this paper it will be assumed that the reference test is a gold standard. The number of true positive test results divided by the number of infected sites in a sample of 29 patients is:

3/6, 2/6, 2/4, 5/6, 4/5, 5/5, 4/6, 3/4, 2/4,
3/4, 5/5, 4/4, 6/6, 3/3, 5/6, 1/2, 4/6, 0/4,
5/6, 4/5, 4/6, 0/6, 4/5, 3/5, 0/2, 2/6, 2/4.

The number of true negative test results divided by the number of non-infected sites in a sample of 21 patients is:

0/1, 3/3, 1/2, 3/3, 1/1, 2/3, 3/3, 1/1, 0/1, 2/3, 2/3,
1/1, 0/1, 1/3, 1/1, 2/2, 4/4, 3/3, 5/5, 1/1, 3/3.

The sample sizes in both groups will rarely coincide, since in some patients all sites may be infected or non-infected sites.

If responses on sites within a patient were assumed to be independent observations, these data could be summarized in a 2×2 table such as shown in Table 1. The binomial probability distribution would be selected as the underlying probability model, and the sensitivity and specificity of the diagnostic test would be respectively determined by $\hat{p} = 94/142 = 0.662$ and $\hat{p} = 39/48 = 0.813$. The estimated variances of these estimates would be $\hat{p}\hat{q}/n$.

Table 1. Summary table of the diagnostic test results: 94 out of 142 sites are true positive results and 39 out of 48 are true negative results. Note that with the information of this table alone, the sample size of patients cannot be determined

	Sites	Test result positive	Test result negative	
Infection	present	94	48	142
	absent	9	39	48

Thus the standard error of the sensitivity estimate would be 0.040 and the standard error of the specificity would be 0.056.

However, since sites within a patient cannot be assumed to be independent, a different underlying probability model should be selected and the within-patient correlation coefficient should be estimated. With the use of the correlated binomial model one can estimate these parameters. A preliminary likelihood ratio test to investigate if the within-patient correlation of the test responses is significantly different from zero yields a $\chi^2 = 0.57$ for the specificity and $\chi^2 = 4.73$ for the sensitivity. The presence of a significant subject-effect (p = 0.03) in the infected sites precludes the use of the ordinary binomial distribution model for these data. In general, a test of significance of the within-patient correlation coefficients should not be used to decide on the choice of the probability model. If several sites within a patient are sampled, the observations on those sites are by definition dependent observations and a correlated binomial model or other statistical model which accounts for this dependence should be employed. The within-patient correlation coefficients of dependent observations can be zero and/or insignificant due to small sample sizes.

By maximizing the likelihood function of the correlated binomial model (3) we obtain the maximum likelihood estimate (MLE) of the sensitivity to be 0.645. The MLE of the within-patient correlation coefficient is 0.121. The standard error of the sensitivity is 0.052. The MLE of specificity 0.798 is with a standard error of 0.067 and a within-patient correlation coefficient of 0.148. Table 2 shows a comparison of the estimates obtained by the binomial model and the correlated binomial model. Note that there are only small differences in the sensitivity and specificity estimates between the two models. Standard errors of these estimates,

however, are rather different, with the binomial model giving smaller standard errors than the correlated binomial model. In the presence of positive correlation, the incorrect binomial model indicates that the sample contains more information, and hence less variance, than is warranted.

Discussion

Several individuals have recommended the use of diagnostic tests in periodontal disease (6, 7). This search for effective diagnostic tests emphasizes the need for objective procedures to evaluate and compare these tests. Ignoring the correlation between sites can significantly affect the variance estimates and thereby distort differences between diagnostic tests. It can be shown that the variance of the ordinary binomial estimator of sensitivity and specificity, $\hat{p} = x/n$, is not equal to pq/n but to:

$$\text{Var}(\hat{p}) = pq/n + \Sigma n_i(n_i - 1)/n^2.p.q.\rho \qquad (4)$$

where n is the total number of infected (or non-infected) sites and n_i is the number of sites within patient x_i. This formula illustrates the effect of the within-patient correlation coefficient on the variance estimates. If the within-patient correlation coefficient is zero, all results will be identical to results obtained by the ordinary binomial probability model. However, the larger the correlation coefficient and/or the more sites sampled per patient, the greater the effect on the variance. A positive correlation coefficient will inflate the variance, a negative correlation coefficient will reduce the variance. It is apparent from this formula that even a very small within-patient correlation coefficient can have a sizeable effect on the variance estimate if the number of sites sampled per patient is large. For instance, studies of periodontal disease activity may measure well over 100 sites per patient and therefore produce significant under-estimation of the standard errors of their estimates if the ordinary binomial probability model is used.

Use of the correlated binomial model offers several advantages: 1) it allows for a correct estimation of the sensitivity and specificity and their variances, 2) the significance of the within-patient correlation coefficient may provide a criterion for assessing the effect of patient-related factors on the performance of the diagnostic tests, and, 3) the

Table 2. Comparison of sensitivity and specificity estimates obtained with the binomial model and the correlated binomial model. Observe that the standard errors are underestimated by using the binomial model

Model	Sensitivity	S.E.(a)	$\hat{\rho}$	Specificity	S.E.	$\hat{\rho}$
Binomial model	0.662	0.040	*	0.813	0.056	*
Correlated binomial model	0.645	0.052	0.121	0.798	0.067	0.148

(a) Standard error

model is flexible and can easily be expanded to provide likelihood ratio tests to compare different diagnostic tests. The model has the disadvantage that, for some problems, boundary conditions may be met and that, under these conditions, asymptotic theory of MLE breaks down. For these situations we recommend the use of jackknifing methodology (8).

In some dental studies, the number of sites sampled does not vary greatly across patient, and differences between the point estimates of the ordinary binomial model and the correlated binomial model will be slight. Greater differences will arise with larger variations in the number of sites examined in each patient and small patient sample size, since the contribution of each patient is effectively weighted differently by the methods according to the binomial denominators. More importantly, and the primary motivation for using a correlated binomial model, is that the variances of these estimates can differ considerably. Ignoring the patient-effect on the outcome of the diagnostic tests can considerably reduce the variance of the estimates when the within-patient correlation coefficient is positive and, as a result, can greatly inflate the type I error rate when diagnostic tests are evaluated or compared. For example, suppose that the true standard error of the sensitivity estimate in Table 2 is 0.054 (as estimated by the correlated binomial model), then calculation of a nominal 95% level confidence interval using the ordinary binomial estimate of 0.04 would yield an interval with only 87% actual coverage. (Note: $1.96 \times (0.040/0.052) = 1.51$, the z score corresponding to an 87% confidence interval).

The error rates of the reference tests used for evaluating the diagnostic test are of considerable importance for the estimation of sensitivity and specificity. Three different possibilities exist: (1) the reference test is a gold standard (no misclassifications), (2) the reference test has known error rates, and, (3) the reference test has unknown error rates. If a reference test with known or unknown error rates is used, more complicated estimation procedures are necessary if one wants to obtain unbiased estimates of sensitivity and specificity (9, 10). Some of these estimation methods have been subject to criticism (11). The methods described in this paper provide an unbiased estimate when the reference test is a gold standard. When the methods are used for situations were the reference test has error rates, the obtained estimates of sensitivity and specificity will be biased and dependent on the prevalence of the condition under investigation.

Since the search for site-specific diagnostic tests has been given a high priority by the National Institute of Dental Research (12), it is important that the obtained estimates of sensitivity and specificity be accurate. The clinical value of a diagnostic test may be misjudged and/or comparisons between different diagnostic tests may yield misleading conclusions when large within-patient correlation coefficients are present. We conclude that the correlated binomial probability model represents a realistic and useful model for investigating the sensitivity and the specificity of site-specific diagnostic tests.

Acknowledgments

References

1. Fleiss JL, Wallenstein S, Chilton NW, et al. A re-examination of within-mouth correlations of attachment level and of change in attachment level. *J Clin Perio* 1988; **15**: 411.
2. Bahadur RR. A representation of the joint distribution of responses to n dichotomous items. In: Solomon H, ed. *Studies in Item Analysis and Prediction*. Stanford: Stanford University Press, 1961.
3. SAS Institute Inc. SAS Language guide for personal computers, Version 6 Edition. Cary, NC: SAS Institute Inc. 1985.
4. Kupper LL, Haseman KJ. The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* 1978; **34**: 69.
5. Loesche WJ, Syed SA, Stoll J. Trypsin-like activity in subgingival plaque. A diagnostic marker for spirochetes and periodontal disease. *J Periodontol* 1987; **58**: 266.
6. Polson AM, Goodson JM. Periodontal diagnosis. Current status and future needs. *J Periodontol* 1985; **56**: 25.
7. Hancock EB. Determination of periodontal disease activity. *J Periodontol* 1981; **52**: 492.
8. Gladen B. The use of the jackknife to estimate proportions from toxicological data in the presence of litter effects. *J A Stat Ass* 1979; **74**: 278.
9. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1980; **36**: 167.
10. Buck AA, Gart JJ. Comparison of a screening test and a reference test in epidemiologic studies. *J Epidemiol* 1966; **83**: 586.
11. Begg CB. Biases in the assessment of diagnostic tests. *Stat in Med* 1987; **6**: 411.
12. Kakehashi S, Parakall P. Proceedings from state of art workshop. *J Periodontol* 1982; **53**: 475.

Address:

Philippe P. Hujoel
Department of Dental Public Health Sciences, SM-35
School of Dentistry
University of Washington
Seattle, WA 98195-9950
U.S.A.