Comparing disease screening tests when true disease status is ascertained only for screen positives

MARGARET SULLIVAN PEPE*, TODD A. ALONZO
University of Washington, Department of Biostatistics, Box 357232, Seattle, WA 98195-7232, USA
mspepe@u.washington.edu

SUMMARY

Disease screening is a fundamental part of health care. To evaluate the accuracy of a new screening modality, ideally the results of the screening test are compared with those of a definitive diagnostic test in a set of study subjects. However, definitive diagnostic tests are often invasive and cannot be applied to subjects whose screening tests are negative for disease. For example, in cancer screening, the assessment of true disease status requires a biopsy sample, which for ethical reasons can only be obtained if a subject’s screening test indicates presence of cancer. Although the absolute accuracy of screening tests cannot be evaluated in such circumstances, it is possible to compare the accuracies of screening tests. Specifically, using relative true positive rate (the ratio of the true positive rate of one test to another) and relative false positive rate (the ratio of the false positive rates of two tests) as measures of relative accuracy, we show that inference about relative accuracy can be made from such studies. Analogies with case-control studies can be drawn where inference about absolute risk cannot be made, but inference about relative risk can.

In this paper, we develop a marginal regression analysis framework for making inference about relative accuracy when only screen positives are followed for true disease. In this context factors influencing the relative accuracies of tests can be evaluated. It is important to determine such factors in order to understand circumstances in which one test is preferable to another. The methods are applied to two cancer screening studies, one concerning the effect of race on screening for prostate cancer and the other concerning the effect of tumour grade on the detection of cervical cancer with cytology versus cervicography screening.

1. INTRODUCTION

Screening programmes for cancer are now an established component of health care. Examples of screening tests are mammography for breast cancer, serum PSA levels for prostate cancer, and the Pap smear for cervical cancer. New screening tests and modifications to existing screening tests are sought to improve upon our ability to detect disease early. This paper is concerned with methodology to compare screening tests in regards to their capacities for accurately detecting disease. Although the methods are motivated by the cancer screening problem, they apply equally well to other diseases including, for example, screening for cardiovascular disease.

Let $D$ be a binary indicator of true disease status ($D = 1$ if diseased) and $Y$ be a binary variable denoting the result of the screening test ($Y = 1$ if screen positive for disease and $Y = 0$ if screen negative for disease). The true and false positive rates, $TPR = P(Y = 1 | D = 1)$ and $FPR = P(Y = 1 | D = 0)$, respectively, are used to quantify accuracy. Ideally, a screening test would detect all subjects with disease,
Table 1. Ideal (complete) data from a cohort study. Brackets denote components that are unknown in a screen-positive study. The number of subjects screening negative on both tests \([d]+[h]\) is also known.

<table>
<thead>
<tr>
<th></th>
<th>Diseased (D = 1)</th>
<th>Non-diseased (D = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y_A = 1)</td>
<td>(Y_B = 1)</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>(Y_B = 1)</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>(Y_B = 1)</td>
<td>[d]</td>
</tr>
<tr>
<td>(Y_A = 0)</td>
<td>(Y_B = 1)</td>
<td>e</td>
</tr>
<tr>
<td></td>
<td>(Y_B = 1)</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>(Y_B = 1)</td>
<td>[n(D)]</td>
</tr>
</tbody>
</table>

\(TPR = 1\), and not mistakenly screen positive any non-diseased subjects, \(FPR = 0\). Errors of the latter sort subject individuals to diagnostic work-up that is often invasive and can be a costly burden on public health resources. Most screening tests are unfortunately not completely accurate. The well established Pap smear test for cervical cancer, for example, is thought to have true and false positive rates of about 51% and 2%, respectively (McCormy et al., 1999).

To compare a new screening test, test A, with a standard test, test B, typically a cohort study is undertaken with each subject being screened with both tests. We denote the test results for subject \(i\) by \(Y_{A,i}\) and \(Y_{B,i}\). To evaluate the true and false positive rates for the two tests, true disease status \(D_i\) also needs to be determined. Using the notation of Table 1, estimates are then calculated as

\[
\hat{TPR}_A = \frac{(a + b)}{n_D} \quad \hat{TPR}_B = \frac{(a + c)}{n_D} \\
\hat{FPR}_A = \frac{(e + f)}{n_D} \quad \hat{FPR}_B = \frac{(e + g)}{n_D}.
\]

Unfortunately, in many settings definitive diagnostic testing for disease is invasive. A biopsy sample is typically required for establishing presence or absence of cancer. Angiography is the standard procedure for assessing the extent of coronary artery disease. Such procedures are warranted in subjects who have a positive screening test and hence are considered at high risk of having disease, but they cannot be justified in patients that screen negative. A research study protocol will usually dictate that only those subjects who screen positive on either screening test are referred for definitive diagnostic testing. Data from such a study is represented in Table 1 where the bracketed entries denote unknown values. Studies with this design are frequently reported in the literature. Some recent examples from the cervical cancer screening literature are (De Sutter et al., 1998; Sankaranarayanan et al., 1998; Wright et al., 2000). Table 2 displays data from a study of prostate cancer screening (Smith et al., 1997).

Estimates of the true and false positive rates cannot be determined from such studies, because the number of diseased and non-diseased subjects in the study, i.e. the denominators for \(TPR\) and \(FPR\), are unknown. However, the relative true positive rate, \(rTPR = TPR_A/TPR_B\) and relative false positive rate, \(rFPR = FPR_A/FPR_B\) can be estimated as

\[
rTPR = \frac{(a + b)}{(a + c)} \quad \text{and} \quad rFPR = \frac{(e + f)}{(e + g)}.
\]

Cheng and Macaluso (1997) provide an expression for confidence interval calculation and Schatzkin et al. (1987) note that McNemar’s test can be used to compare true positive or false positive rates. In
Comparing disease screening tests when true disease status is ascertained only for screen positives

Table 2. Distribution of PSA screening test result digital rectal examination (DRE) result for men found to have or not have prostate cancer by biopsy (Smith et al., 1997). Data are stratified by race

<table>
<thead>
<tr>
<th></th>
<th>Black men (n = 949)</th>
<th></th>
<th>White men (n = 18 527)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With prostate cancer</td>
<td>Without prostate cancer</td>
<td>With prostate cancer</td>
</tr>
<tr>
<td></td>
<td>DRE +</td>
<td>DRE−</td>
<td>DRE+</td>
</tr>
<tr>
<td>PSA +</td>
<td>10</td>
<td>28</td>
<td>3</td>
</tr>
<tr>
<td>PSA −</td>
<td>8</td>
<td>?</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In this paper we propose new statistical methodology that allows one to address questions beyond simple estimation and hypothesis testing. We propose a regression framework that can be used to determine factors influencing the relative performance of tests. It is important to determine such factors because different tests may be preferable under different circumstances. For example, Strax et al. (1967) have shown that age affects the relative performance of mammography versus breast self-exam for breast cancer screening presumably because breast tissue is denser in younger women. The data in Table 2 suggest that relative to digital rectal examination (DRE), PSA screening leads to a higher rate of false positives in black men than in white men. Finally, one test might be better than another at detecting certain types of disease such as cancer of a particular histologic type, but worse at detecting other variants of the disease. In order to optimize screening we need to understand the influence of such factors.

The statistical issues that arise here in screen-positive studies as we call them, bear some similarity with issues that arise in case-control studies. In case-control studies, estimation of absolute risk of disease is not possible, but estimation of relative risk is. In screen-positive studies estimation of absolute true and false positive rates is not possible, but relative rates are estimable. Logistic regression methodology allows one to evaluate factors affecting relative risks in case-control studies. The methods we propose here analogously provide a framework for evaluating factors affecting relative rates of true and false positive results. The quantities modelled and the estimation procedures for screen positive studies, however, bear no relation to logistic regression methods for case-control studies. Indeed, odds ratios cannot even be calculated from screen-positive studies because the fourth entries in the $2 \times 2$ tables are missing.

We describe the regression framework and apply it to the data displayed in Table 2, in Sections 2 and 3. In Section 4 we extend the regression models to accommodate disease-specific covariates and show that
the approach can be used to determine if grade of a cancer lesion affects the relative detection rate with data from a cervical cancer study. We next discuss some technical issues in applying the method and contrast the general approach with conditional logistic regression which has previously been proposed for this problem. Some further remarks are provided in Section 8.

2. REGRESSION MODELS WITH SIMPLE COVARIATES

Let $Z$ denote a vector of covariates with 1 as the first entry and consider initially the relative true positive rate, which at covariate level $Z$ is denoted by $TPR(Z)$. Analogous methods will apply to the relative false positive rates. In order to fit a model to $rTPR(Z)$ we propose to fit a marginal regression model to the disease detection rates, $P[Y_A = 1, D = 1|Z]$ and $P[Y_B = 1, D = 1|Z]$, and postulate that certain parameters in the detection rate model quantify covariate effects on the relative true positive rate. In fact, if we use the models

$$
\log P[D = 1 \text{and } Y_B = 1|Z] = \alpha Z \\
\log P[D = 1 \text{and } Y_A = 1|Z] = \alpha Z + \beta Z
$$

we claim that $\beta$ relates to covariate effects on $rTPR(Z)$.

We write the above models in a single equation using the notation $T$ to denote test type, $T = 1$ for test A and $T = 0$ for test B. We can represent the data $(D, Y_A, Y_B, Z)$ for an individual as two observations $\{(D, Y, Z, T); T = 0, 1\}$, and rewrite the equations above as

$$
\log P[D = 1 \text{ and } Y = 1|Z, T] = \alpha Z + \beta Z T. \tag{2.1}
$$

Observe that this model relates to the marginal disease detection rate frequencies in Table 1, $(a + b)/N$ when $T = 1$ and $(a + c)/N$ when $T = 0$ where $N$ is the total number of screened subjects. It follows from this model form that

$$
\beta Z = \log P[D = 1, Y = 1|Z, T = 1] - \log P[D = 1, Y = 1|Z, T = 0]
= \log \frac{P[D = 1, Y = 1|Z, T = 1]}{P[D = 1, Y = 1|Z, T = 0]}
= \log \frac{P[Y = 1|D = 1, Z, T = 1]}{P[D = 1|Z, T = 0]} + \log \frac{P[D = 1|Z, T = 1]}{P[D = 1|Z, T = 0]}
= \log rTPR(Z) + \log \frac{P[D = 1|Z, T = 1]}{P[D = 1|Z, T = 0]}
= \log rTPR(Z)
$$

if $P[D = 1|Z, T = 1] = P[D = 1|Z, T = 0]$. This last equation states that conditional on the covariates, observations relating to test type A have the same frequency of disease as observations relating to test B. This is trivially true for paired data where the frequencies by design are exactly the same. In more general settings it requires that, conditional on covariates, the two tests are applied in populations with the same disease prevalence. This seems like a sensible requirement of study design in general.

To summarize, the proposal is to fit a marginal regression model with a log link function to the detection rates and to include an interaction term between test type and covariates in the linear predictor. The binary outcome or dependent variable in this model is $U = I[D = 1, Y = 1]$ where $I[\cdot]$ is the
Table 3. Results of prostate cancer data analysis using marginal log-linear models. Also shown are results of conditional logistic regression (CLR) analyses

<table>
<thead>
<tr>
<th>Factor</th>
<th>Estimate</th>
<th>Exponentiated estimate</th>
<th>95% CI</th>
<th>p-value</th>
<th>CLR estimate</th>
<th>CLR p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>( \hat{\alpha}_1 = -4.07 )</td>
<td>-</td>
<td>(0.69, 1.78)</td>
<td>0.66</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Race (white = 0, black = 1)</td>
<td>( \hat{\alpha}_2 = 0.11 )</td>
<td>1.11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Test type (DRE = 0, PSA = 1)</td>
<td>( \hat{\beta}_1 = 0.34 )</td>
<td>1.40</td>
<td>(1.26, 1.56)</td>
<td>&lt;0.001</td>
<td>( \hat{\beta}^*_1 = 0.66 )</td>
<td>&lt;0.000</td>
</tr>
<tr>
<td>Race × test type</td>
<td>( \hat{\beta}_2 = 0.41 )</td>
<td>1.51</td>
<td>(0.95, 2.39)</td>
<td>0.08</td>
<td>( \hat{\beta}^*_2 = 0.60 )</td>
<td>0.15</td>
</tr>
<tr>
<td>False positive rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>( \hat{\eta}_1 = -2.81 )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Race (white = 0, black = 1)</td>
<td>( \hat{\eta}_2 = -0.68 )</td>
<td>0.51</td>
<td>(0.35, 0.73)</td>
<td>&lt;0.001</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Test type (DRE = 0, PSA = 1)</td>
<td>( \hat{\theta}_1 = -0.26 )</td>
<td>0.77</td>
<td>(0.71, 0.83)</td>
<td>&lt;0.001</td>
<td>( \hat{\theta}^*_1 = -0.31 )</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Race × test type</td>
<td>( \hat{\theta}_2 = 0.61 )</td>
<td>1.84</td>
<td>(1.16, 2.92)</td>
<td>0.01</td>
<td>( \hat{\theta}^*_2 = 0.69 )</td>
<td>0.01</td>
</tr>
</tbody>
</table>

indicator function, and rather than using a typical binary regression link function such as logistic or probit, we use a log link function because the scale for analysis involves ratios. In this model, the coefficients relating to the interaction are the coefficients in a regression model for the relative true positive rates:

\[
\log rTPR(Z) = \beta Z
\]

this induces a regression model for the relative false positive rates:

\[
\log rFPR(Z) = \theta Z.
\]

3. ILLUSTRATION WITH PROSTATE CANCER DATA

Smith et al. (1997) screened 18,527 white men and 949 black men for prostate cancer using DRE and serum PSA. The PSA level was considered suspicious for cancer if it exceeded 4.0 ng ml\(^{-1}\). Subjects with positive screening test results on either DRE or PSA were referred for ultrasound guided needle biopsy. Although biopsy is prone to error because it can sometimes miss the disease site, it will be considered a gold standard definitive test here. Data from this study are displayed in Table 2. Results of fitting the regression models (2.1) and (2.2) are shown in Table 3 where \( Z = (1, \text{ race}) \) and \( T = \text{ test type} \) with the race and test type variables coded as shown in the table. The results indicate that PSA has a higher true positive rate than DRE (\( \hat{\beta}_1 = 0.34, p < 0.001 \)) and that the relative performance appears to be even higher in blacks than in whites (\( \hat{\beta}_2 = 0.41, p = 0.08 \)). When the main effect for race on the detection rate was dropped from the model because of its apparent lack of effect, i.e. we set \( \alpha_2 = 0 \), the \( p \)-value for \( \hat{\beta}_2 \), the race effect on \( rTPR \), became 0.002. In Table 3 the estimated \( rTPR \) for PSA versus DRE in whites is \( 1.40 \), while that for blacks is \( 1.40 \times 1.51 = 2.11 \).

Turning now to the false positive rates, it appears that at least in whites, PSA has a lower false positive rate than does DRE (\( \hat{\theta}_1 = -0.26, p < 0.001 \)). However, the relative performance in blacks is not as good as it is in whites, the \( rFPR \) among blacks being about 1.84 times that in whites (\( \hat{\theta}_2 = 0.61, p = 0.01 \)).
Table 4. Comparison of cervicography and cytology in the detection of high- and low-grade cervical lesions. Shown are data for 228 subjects who tested positive with one or more of the screening tests. A total of 4964 subjects had negative screening test results with both cervicography and cytology.

<table>
<thead>
<tr>
<th></th>
<th>High-grade lesions</th>
<th>Low-grade lesions</th>
<th>No lesions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cervicography</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>−</td>
<td>4</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td><strong>Cytology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>6</td>
<td>29</td>
<td>11</td>
</tr>
<tr>
<td>−</td>
<td>48</td>
<td>?</td>
<td>81</td>
</tr>
</tbody>
</table>

Indeed, testing the hypothesis that $\theta_1 + \theta_2 = 0$ yielded a $p$-value of 0.14 and an estimated $rFPR$ of $\exp(\hat{\theta}_1 + \hat{\theta}_2) = 1.41$ in blacks. This suggests that PSA is associated with a higher false positive rate than DRE in black men. Since both the $rTPR$ and the $rFPR$ are higher in blacks than in whites, this suggests that perhaps a threshold value higher than the standard 4.0 ng ml$^{-1}$ should be used in blacks. A definitive answer to this question cannot be given based on these data, however, because it is not known if the baseline test, DRE, has similar performance in the two races. Although the detection rates with DRE are similar for the two races ($\hat{\alpha}_2 = 0.11$, $p = 0.66$), the prevalences may be different and if so, the relative performance of DRE in the two races would be different.

Although our main focus is on the interaction terms in the models, it is interesting to consider the interpretation of main effect parameters $\alpha$ and $\eta$, and what factors can affect them. The parameter $\alpha$ relates to the detection rate, $P[D = 1, Y = 1]$, with the baseline test. Thus factors that influence disease prevalence will affect $\alpha$, as will factors that influence the true positive rate of the baseline test. The study design does not allow one to separately identify effects on these two components, but only allows one to assess the compound effects on detection rate. Similar considerations apply to $\eta$, which relates to effects on the false referral rate, $P[D = 0, Y = 1]$, with the baseline test.

4. REGRESSION WITH DISEASE-SPECIFIC COVARIATES

Consider a covariate specific to diseased individuals, some feature of the disease such as severity, histology or grade of a cancer lesion, say. One might ask if the relative rates at which two tests detect disease depend on such features. In Table 4 we display data from De Sutter et al. (1998) relating to a comparison of cervicography and cervical cytology for cervical cancer screening. The data suggest that cytology is better than cervicography at detecting high-grade lesions but that low-grade lesions are better detected with cervicography.

To analyse such data we propose that the features relating to disease be quantified in a discrete fashion, so that disease can be considered to be one of $K$ possible types. In our example, grade of lesion defines disease type and $K = 2$. For each individual we define $K$ separate disease indicators, $D_k = I$ (diseased with $k$th set of features). Thus the data for an individual in the cohort can be represented as

$$\{(D_k, Y, Z, T); k = 1, \ldots, K \text{ and } T = 0, 1\}.$$ Only one of the $K$ disease indicators will equal 1 for an individual known to be diseased and none will
equal 1 for an individual known to not have disease.

The detection rate model we propose to fit with these data is

\[ \log P[D_k = 1, Y = 1 | Z, T] = \alpha Z + \beta Z T + \gamma X_k + \delta X_k T \]  

(4.1)

where as before Z is a simple covariate defined for all study subjects with first element equal to 1, and \( X_k \) is a covariate vector defined from the \( k \)th set of disease features. It can be shown that this model induces the following model for the relative true positive rate:

\[ \log \frac{P[Y_A = 1 | Z, D_k = 1]}{P[Y_B = 1 | Z, D_k = 1]} = \log rTPR_k(Z) = \beta Z + \delta X_k. \]

Thus the parameter \( \delta \) quantifies how the relative true positive rate varies with features of disease as defined by \( X_k \). For example, suppose that \( X \) is a set of \( K - 1 \) dummy variables for disease type with disease type 1 being baseline, \( X_1 = (0, \ldots, 0) \), and \( X_k \) having a single entry of 1 at the \( k - 1 \)th dummy variable position. Then the \( k \)th element of \( \delta \) quantifies the difference between \( \log(rTPR) \) for test A versus B for detecting disease type \( k \) in comparison to the \( \log(rTPR) \) for detecting disease type 1. In general, interactions between simple covariates and disease-specific covariates can also be incorporated into the model by including such terms and their interactions with \( T \) in the detection rate model. This allows for the possibility that covariates will modify the differences in relative performance across disease types.

5. ILLUSTRATION WITH CERVICAL CANCER DATA

De Sutter et al. (1998) conducted a multicenter study to compare cervicography with the standard Pap smear cytology test for detecting cervical cancer. Subjects who were positive for either test were referred for colposcopy with directed biopsy which is again considered a gold standard in this example. A total of 5192 women completed the protocol of whom 228 were biopsied. Histologic examination of the biopsy was used to classify disease as low grade (condyloma or CIN 1) or high grade (CIN 2 or invasive cancer) or not present.

In this example, there are no simple covariates so that \( Z = 1 \). The covariate \( X \) is an indicator variable for grade of disease, \( X = 1 \) for high grade and \( X = 0 \) for low grade. Each individual thus contributed four observations to the analysis, \((D,Y,T,Z)\) one for each disease type and test type combination. Results of the analysis are shown in Table 5.

With regard to the analysis of true positive rates, the parameters \( \beta \) and \( \delta \) are of most interest. Since \( \beta < 0 \) \((p = 0.032)\) we conclude that for low-grade disease, cytology screening has a lower true positive rate than does cervicography. Since \( \delta > 0 \) \((p = 0.001)\) the data also show that the relative performance of cytology versus cervicography is different for high- and low-grade disease. Indeed \( \beta + \delta = 0.48 \) and a significance test for \( \beta + \delta = 0 \) yields a \( p \)-value of 0.01. Hence we conclude that for high-grade disease, cytology has the higher true positive rate.

The analysis of false positive rates does not include grade of disease as a covariate because it is irrelevant. The parameter \( \theta < 0 \) \((p < 0.001)\) indicates that the false positive rate associated with cytology is less than that for cervicography. The same \( p \)-value is obtained with McNemar’s test for paired binary data, as was suggested by Schatzkin et al. (1987). Moreover, the estimated relative false positive rate from the model is the empirical estimate from the table 0.34 = 31/92 with a confidence interval that corresponds to that proposed by Cheng and Macaluso (1997). We thus illustrate that the regression framework introduced here reduces to the current standard-of-practice statistical methods in the simple setting where there are no covariates.
Table 5. Results of cervical cancer data analyses

<table>
<thead>
<tr>
<th>Factor</th>
<th>Parameter estimate</th>
<th>Exponentiated estimate</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>( \hat{a} = -4.57 )</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Test type (cytology = 1)</td>
<td>( \hat{\beta} = -0.43 )</td>
<td>0.65</td>
<td>(0.44, 0.96)</td>
<td>0.032</td>
</tr>
<tr>
<td>Grade (high = 1)</td>
<td>( \hat{y} = -1.10 )</td>
<td>0.33</td>
<td>(0.20, 0.57)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Test ( \times ) grade</td>
<td>( \hat{\delta} = 0.91 )</td>
<td>2.49</td>
<td>(1.44, 4.28)</td>
<td>0.001</td>
</tr>
<tr>
<td>False positive rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>( \hat{\eta} = -4.03 )</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Test type (cytology = 1)</td>
<td>( \hat{\theta} = -1.09 )</td>
<td>0.34</td>
<td>(0.23, 0.49)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

6. FITTING THE MODELS

To fit models (2.1) or (4.1) to data we note that these are marginal generalized linear models (in fact they are marginal log–linear models) and hence generalized estimating equation (GEE) methods as described by Zeger and Liang (1986) were used to fit them. The link function is a log link. This is not the canonical link for binomial data. Moreover, unlike the logit link function, it can allow fitted probabilities to exceed 1. Nevertheless, we use the log link because, as we have shown, this yields an important interpretation for coefficients associated with the test type covariate \( T \) as pertaining to relative true and false positive rates. Moreover, for datasets that we have used, all fitted probabilities were in the range (0, 1).

We used an independence working covariance matrix in applying GEE, thus treating observations as independent in the estimation procedure but using a robust estimator of the variance–covariance matrix of estimates. A GEE procedure that uses the actual variance–covariance of the subject observations rather than an independence model for the working covariance matrix would presumably be more efficient (Zeger and Liang, 1986). The within-subject variance–covariance can often be easily calculated. Consider the setting of Section 2 and the analysis of true positive rates. Let \( U_{ij} = I[D_i = 1, Y_{ij} = 1] \) with \( T_{i1} = 0 \) and \( T_{i2} = 1 \). These binary variables are the response variables used to fit the model. Observe that

\[
\text{cov}(U_{i1}, U_{i2}|Z_i, T_{i1}, T_{i2}) = E(U_{i1}, U_{i2}|Z_i, T_{i1}, T_{i2}) - E(U_{i1}|Z_i, T_{i1})E(U_{i2}|Z_i, T_{i2})
\]

\[
= P(Y_{iA} = 1, Y_{iB} = 1, D_i = 1|Z_i) - P(Y_{iA} = 1, D_i = 1|Z_i)P(Y_{iB} = 1, D_i = 1|Z_i).
\]

The components of the second term are estimated from the model (2.1). When \( Z \) is discrete the first component can be estimated empirically by the observed frequency in the upper left-hand corner of the table for diseased subjects among all subjects with covariate \( Z = Z_i \).

We re-estimated the regression coefficients for the models fit to the prostate cancer data using this working variance–covariance matrix. For the full model displayed in Table 3, the estimates and standard errors were unchanged. For the model that restricted \( \alpha_2 \) to equal 0, i.e. excluded race as a main effect, the standard error was reduced by only 5%.
7. CONDITIONAL LOGISTIC REGRESSION

As an alternative to the marginal regression approach that we have used, one could consider applying conditional logistic regression methods for matched case-control data to this problem as was proposed by Cheng and Macaluso (1997). The two test results for a subject constitute a matched set. Only subjects for whom one test is positive and one test is negative (i.e. the discordant pairs) enter into the analysis. Let $Z$ be a simple covariate, again with first element equal to 1, and consider the analysis of true positive results, that is data for subjects known to be diseased.

The conditional logistic regression model presented by Cheng and Macaluso (1997) is expressed as

$$\logit P(Y_A = 1 | Y_A + Y_B = 1, D = 1, Z) = \beta^* Z.$$ (7.1)

This model pertains to the probability amongst diseased subjects with discordant results, that $Y_A$ rather than $Y_B$ is positive. Having expressed the interpretation of the regression parameter $\beta^*$ in this fashion, it seems of minor practical interest.

The model (7.1) can be derived from a subject-specific model for the true positive rates

$$\logit P[Y_{iB} = 1 | D_i = 1, Z_i] = \alpha_i^*$$
$$\logit P[Y_{iA} = 1 | D_i = 1, Z_i] = \alpha_i^* + \beta^* Z_i$$

or equivalently

$$\logit P[Y_{ij} = 1 | D_i = 1, Z_i, T_{ij}] = \alpha_i^* + \beta^* T_{ij} Z_i$$ (7.2)

where $T_{ij}$ denotes test type as before. Written in this fashion $\beta^*$ relates to the odds of a positive result with test type A versus test B for a diseased individual. The odds in this model is interpreted at a subject-specific level. This contrasts with the population level interpretation for the relative true positive rates discussed earlier in this paper.

We applied this modelling approach to the data in Table 2, with covariate vector $Z = (1, \text{race})$ as before. The estimates are shown in Table 3 under the CLR column. Qualitatively, the results are similar to those using the marginal log-linear regression approach. However, we find the CLR estimates less appealing because: (i) odds ratios rather than rate ratios are estimable with the CLR approach. In contrast to epidemiologic studies of low incidence disease, in our context rate ratios are not approximated by odds ratios. We believe that relative rates like relative risks have a more appealing interpretation than odds ratios in practice; (ii) although the CLR parameters are often interpreted relative to the subject-specific models (7.2), the validity of the subject-specific model cannot be assessed. In fact, one fits the marginal model (7.1) to the data and hence we consider it more appropriate to interpret $\beta^*$ (and $\alpha^*$) relative to that marginal model. Indeed it is quite possible that subject-specific models other than (7.2) give rise to (7.1), in which case interpretation of $\beta^*$ relative to (7.2) is invalid; (iii) most importantly, even if (7.2) is the true subject-specific model, we contend that it is the population level relative performance of the two tests that holds relevance for public health purposes, rather than the subject-specific relative performance. The population level relative performance is quantified by the population average models that we have proposed in Sections 2 and 4 rather than by (7.2).

8. LATENT CLASS ANALYSIS

We have focused on relative true and false positive rates as measures of relative accuracy in this paper. Although estimation of absolute values for true and false positive rates would be more useful, they are essentially not identifiable from screen-positive studies. It has been suggested, however, that when several
tests are performed and if some statistical structure is assumed for the joint error rates, then absolute values can be estimated. Walter (1999) discussed inference under the conditional independence assumption

\[ P[Y_A = 1, Y_B = 1|D] = P[Y_A = 1|D]P[Y_B = 1|D] \]  

(8.1)

and showed that latent class models yield maximum likelihood estimates of true and false positive rates and of disease prevalence. Cheng et al. (1999) raised significant concerns about this approach. They showed that the validity of this approach relied heavily on assumption (8.1), by showing that in the absence of (8.1) very different combinations of (prevalence, TPR_A, TPR_B, FPR_A, FPR_B) would give rise to the same expected frequencies of observed data (i.e. Table 1) and hence by implication that serious biases would occur even in large samples when (8.1) fails. Walter acknowledged that assumption (8.1) can often fail in practice. Consider as an example, if both test A and test B are more likely to be positive when a subject has severe disease than if he has mild disease. Indeed, using goodness-of-fit statistics, the assumption was shown to fail in Walter’s example.

Walter (1999) suggested that when the conditional independence is in doubt, inference can in fact be made under less restrictive conditions. We now argue, albeit informally, that even under minimal assumptions, latent class analysis can yield invalid estimates of test accuracy. Consider the setting without covariates. Since there are six degrees of freedom in the data (Table 1), and five parameters of key interest (prevalence, TPR_A, TPR_B, FPR_A, FPR_B), as suggested by Walter (1999) the data allow estimation of one parameter that quantifies correlation. Without loss of generality, we choose that parameter to quantify correlation between the false positive errors on the right-hand side of Table 1. Some assumption concerning the correlation of true positive test results on the left-hand side of Table 1 is required in order to complete specification of the likelihood. Equivalently, some assumption about the probability frequency in the \( d \) cell is required.

Under a conditional independence assumption of true positive tests results, \( p_d = (p_b, p_c)/p_a \), where \( p_k \) denotes the probability frequency in the \( k \)th cell, \( k = a, b, c, d \). Under this assumption, the data in Table 2 for black men yield \( \hat{d} = 22 \) and test accuracy estimates \( \hat{TPR}_A = 56\% \) and \( \hat{TPR}_B = 26\% \). On the other hand, if one assumes that the nature of disease is such that it is detectable by at least one of the two tests, i.e. \( p_d = 0 \), then \( \hat{d} = 0 \) and \( \hat{TPR}_A = 83\% \) and \( \hat{TPR}_B = 39\% \). If one assumes that a significant proportion of disease is not detected by either test (e.g. \( p_d = p_a + p_b + p_c \)) then \( \hat{d} = 46 \) and \( \hat{TPR}_A = 41\% \) and \( \hat{TPR}_B = 20\% \). Clearly, the assumption concerning the structure of the true positive correlation parameter has an overwhelming impact on the estimates of absolute true positive rates. Moreover, the data do not allow one to investigate the validity of these assumptions. In simple terms, there is nothing in the data to distinguish between \( \hat{d} = 22, \hat{d} = 0 \) or \( \hat{d} = 46 \). Therefore, it is our opinion that inference about absolute measures of test accuracy from a screen-positive study should be viewed with caution. Without external knowledge of disease prevalence or test performance or such, absolute accuracy is simply not identifiable from the data. Inference about relative accuracy on the other hand, though not as informative, is straightforward.

9. DISCUSSION

The marginal log-linear modelling approach that we have suggested here is a flexible method that allows one to make inference about relative true and false positive rates. It is easy to implement with existing software packages. Code and data for our examples can be found at http://www.biostat.washington.edu/~mspepe/work.html. Although the covariates in our two illustrations were discrete, we note that the regression model formulation allows general types of covariates including covariates with continuous values. Observe that in the latter case empirical values for \( rTPR(Z) \) and \( rFPR(Z) \) cannot be calculated from the data but can be estimated based on the regression model. An important application for
Comparing disease screening tests when true disease status is ascertained only for screen positives

the regression models might be in analysing data from a study conducted at multiple sites. Since disease prevalence, spectrum of disease, as well as implicit or explicit criteria for defining test positivity might vary across sites, detection rates and false referral rates are also likely to vary across sites. Rather than naively pooling data across sites, one could model the detection rates and false referral rates as functions of study site. Using the models one would determine if \( rTPR \) and \( rFPR \) are similar across sites in which case the common values would be estimated again in the context of the regression model.

Although typical screen-positive study designs give rise to paired test result data, in some settings data might not be paired. This occurs, for example, if the implementation of one screening test interferes with the implementation of another, and also in observational studies. Our methodology is valid in these settings too. Important issues for further research include methods for developing study designs, including sample size calculations and exploring the impact of imperfect reference tests on inference. Our results in these regards will be reported elsewhere.

Although our review of recent literature indicates that \( rTPR \) and \( rFPR \) are almost exclusively used to compare tests in screen-positive studies, relative performance of two tests can be quantified with parameters other than relative true and false positive rates. One could, for example, use the difference in detection rates and the difference in false referral rates, which are estimated by \((b-c)/N\) and \((f-g)/N\), respectively. These measures depend on disease prevalence and hence are less transportable in this sense across populations, than are relative true and false positive rates. They need to be calibrated by disease prevalence in order for them to have meaning in terms of test performance amongst diseased and non-diseased populations. Nevertheless, these measures have a certain appeal for public health applications, are estimable from screen-positive data, and it would be valuable to calculate them in analyses. Chock et al. (1997) suggest the ‘\( FP: TP \) ratio’ which is estimated by \((f-g)/(b-c)\) for an overall summary measure and shows how to estimate this measure for varying values of disease prevalence. Regression methods for evaluating covariate effects on these parameters might also be of interest.

ACKNOWLEDGEMENT

This research was supported by grant R01 GM54438 from the National Institutes of Health.

REFERENCES


[Received April 6, 2000; revised August 18, 2000; accepted for publication August 31, 2000]