Memorial Sloan-Kettering Cancer Center
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology
& Biostatistics Working Paper Series

Year 2006 *Paper* 5

Comparing the Predictive Values of Diagnostic Tests: Sample Size and Analysis for Paired Study Designs

Chaya S. Moskowitz* Margaret S. Pepe†

*Memorial Sloan-Kettering Cancer Center, moskowc1@mskcc.org

†Fred Hutchinson Cancer Research Center and University of Washington

http://biostats.bepress.com/mskccbiostat/paper5

Comparing the Predictive Values of Diagnostic Tests: Sample Size and Analysis for Paired Study Designs

Chaya S. Moskowitz and Margaret S. Pepe

Abstract

In this paper we consider the design and analysis of studies comparing the positive and negative predictive values of two diagnostic tests that are measured on all subjects. Although statistical methodology is well developed for comparing diagnostic tests in terms of their sensitivities and specificities, comparative inference about predictive values is not. We derive analytic variance expressions for the relative predictive values. Sample size formulas for study design ensue. In addition, two new methods for analyzing the resulting data are presented and compared with an existing marginal regression methodology.

Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs

Chaya S. Moskowitz

Department of Epidemiology and Biostatistics

Memorial Sloan-Kettering Cancer Center

307 East 63rd Street, 3rd Floor, New York, NY 10021

e-mail: moskowc1@mskcc.org

Margaret S. Pepe

Departments of Biostatistics

University of Washington and Fred Hutchinson Cancer Research Center

Box 357232, Seattle, WA 98195-7232

February 17, 2006

Abstract

In this paper we consider the design and analysis of studies comparing the positive and negative predictive values of two diagnostic tests that are measured on all subjects. Although statistical methodology is well developed for comparing diagnostic tests in terms of their sensitivities and specificities, comparative inference about predictive values is not. We derive analytic variance expressions for the relative predictive values. Sample size formulas for study design ensue. In addition, two new methods for analyzing the resulting data are presented and compared with an existing marginal regression methodology.

key words: accuracy, biomarker, screening, sensitivity, ROC curve

1 Introduction

Recent advances in biotechnology are leading to the development of many new medical tests. Tests can be used for various purposes including diagnosis, prognosis, risk prediction, and disease screening. In parallel, statistical methodology for evaluating these tests has received much attention. Two recent books on the subject are references [1] and [2]. Various measures can be used to quantify test accuracy. Frequently used measures for binary tests are the sensitivity and specificity and the positive and negative predictive values. Let $D = I\{\text{diseased}\}$ denote an individual's disease status and $X = I\{\text{test positive}\}$ contain the results of a diagnostic test. Although we use terminology for diagnostic testing where the predicted outcome is 'presence of disease', as noted above other binary outcomes, such as future occurrence of an event, would be relevant for non-diagnostic tests. The sensitivity, $P(X = 1|D = 1)$, and the specificity, $P(X = 0|D = 0)$, assess the probability of a correct test result conditional on disease status. They determine the extent to which the test accurately reflects presence or absence of disease and are often used in the early stages of test development [2, 3, 4]. On the other hand, the positive predictive value, $PPV = P(D = 1|X = 1)$, and negative predictive value, $NPV = P(D = 0|X = 0)$, measure the probability of disease conditional on the test result. They quantify the practical usefulness of the test for assessing disease status. Predictive values are therefore important in the later stages of test development when definitive studies of the test are conducted in large prospective cohorts.

Several authors have argued that the predictive values have greater clinical relevance than sensitivity and specificity and are more directly applicable in patient care [5, 6]. Despite their importance and practical relevance, statistical procedures for comparing predictive values is

surprisingly limited, especially in comparison to methods for comparing sensitivities and specificities.

Let X_1 and X_2 denote binary test results for two tests. In this paper we consider comparison of their positive predictive values, $P(D = 1|X_1 = 1)$ and $P(D = 1|X_2 = 1)$, and comparison of their negative predictive values, $P(D = 0|X_1 = 0)$ and $P(D = 0|X_2 = 0)$. Because the predictive values depend on the prevalence of disease, $P(D = 1)$, naively estimating them from a study where $P(D = 1|\text{sampled}) \neq P(D = 1)$ will result in biased estimates. For this reason, we assume throughout that data are observed from a cohort design where D, X_1, and X_2 are sampled jointly.

Data for comparing two tests often come from a paired study design. In contrast to an unpaired design where each test is applied to a mutually exclusive group of individuals, in a paired study both tests are assessed on all individuals. Pairing has the advantage that it is statistically more efficient [2]. Moreover, applying both tests to the same individual is intuitively appealing and eliminates confounding. Despite these advantages and the fact that paired study designs are often used in practice, inference procedures for comparing predictive values from such designs are not widely available.

Furthermore, we are not aware of any existing sample size calculations for paired designs aimed at comparing the predictive values of two tests. Sample size calculations for comparing predictive values from an unpaired design are available in the literature [2] and sample size calculations for both unpaired and paired designs comparing sensitivity and specificity are also available. (For summaries of these methods see references [1] and [2].) One aim of this paper is to fill in this apparent gap to aid in the design of rigorous prospective cohort studies necessary for definitive evaluation of medical tests.

One promising method for comparing predictive values from a paired study design has been described. Leisenring, Alonzo, and Pepe [7] propose a marginal regression framework for analyzing such data. We review their approach below. In addition, we discuss two new alternate approaches for formally comparing predictive values. The first involves a direct computation of the relative predictive values and their standard errors and is also the basis for the sample size calculations. It is not a regression approach, though, and as such does not allow for adjustment of covariates that might affect the accuracy of the tests. The second approach is a regression framework that uses only individuals for which the two test results disagree (the discordant pairs).

This paper is organized as follows. In the next section we describe data from the National Cystic Fibrosis Patient Registry which will be used to illustrate the methodology. Section 3 suggests metrics for comparing the predictive values, the relative predictive values. Variance estimates for obtaining confidence intervals for the relative predictive values are detailed here as well. In Section 4 we present sample size formulas derived using these variance expressions. Section 5 describes the regression methodology suggested by Leisenring *et al.* while Section 6 contains a new regression framework for data analysis. In Section 7 we compare the three different analysis methods in a simulation study. Conclusions are in Section 8.

2 National Cystic Fibrosis Patient Registry data

For illustrative purposes, the methodology developed in this paper will be applied to data from the 1996 Cystic Fibrosis Foundation National Registry. We have previously analyzed this data and describe it in greater detail in [8]. Briefly, patients with cystic fibrosis (CF) can

have intermittent acute severe respiratory infections which are called pulmonary exacerbations (PExs). While multiple studies have sought to assess the relationship between various potentially predictive factors and PExs using a standard multivariate regression prognostic factor analysis, few have attempted to compare factors in their ability to predict future PExs. Consequently, here we compare two prognostic factors, (a) a positive culture for the bacterium *Pseudomonas Aeruginosa* in 1995 and (b) the occurrence of at least one PEx in 1995. We explore using this information to predict subsequent PExs in 1996. Since CF patients can have several PExs in a single year, we work with dichotomized variables indicating that either a patient had no PExs in the relevant year or the patient had at least one PEx that year. We have data available on 11,960 patients six years of age and older for this analysis.

Here the two "diagnostic tests" we wish to compare are actually risk factors. In this context, presence of a risk factor is equated to testing positive. Further, these prognostic factors are sought to predict a future event, not to diagnose a current condition. As noted earlier, although the context differs from that of traditional diagnostic testing we see that the statistical framework is the same.

3 Estimating the relative predictive values

Let X_j hold the result of test j, $j = 1, 2$. Its positive predictive value is $PPV_{X_j} = P(D = 1|X_j = 1)$ and negative predictive value is $NPV_{X_j} = P(D = 0|X_j = 0)$. There are a number of ways to quantify differences in the predictive values. We focus here on relative predictive values. The relative positive predictive value is defined as $rPPV = \frac{PPV_{X_1}}{PPV_{X_2}}$ and the relative negative predictive value is $rNPV = \frac{NPV_{X_1}}{NPV_{X_2}}$.

Note that an ideal study will evaluate positive and negative predictive values together, because taken separately they do not present a complete picture of the accuracy of the test. Consider that it is possible to artificially inflate the PPV simply by declaring that everyone has the disease regardless of the test result. While the PPV would be equal to one suggesting a perfect test, the NPV would be zero correctly reflecting that the test is useless.

Data from a paired study can be summarized by the two tables shown in Table 1. There the $i = 1, ..., N$ individuals are classified into $k = 1, ..., 8$ cells. Corresponding to each of the eight cells shown in the table, we assume that there is a true unobserved probability p_k, where $\sum_{k=1}^{8} p_k = 1$. In practice these probabilities are estimated by their empirical estimates, $\frac{n_k}{N}$. Thus $PPV_{X_1} = \frac{p_5+p_6}{p_1+p_2+p_5+p_6}$ is estimated by $\widehat{PPV}_{X_1} = \frac{n_5+n_6}{n_1+n_2+n_5+n_6}$, $PPV_{X_2} = \frac{p_5+p_7}{p_1+p_3+p_5+p_7}$ is estimated by $\widehat{PPV}_{X_2} = \frac{n_5+n_7}{n_1+n_3+n_5+n_7}$, and similar expressions can be written for NPV_{X_1} and NPV_{X_2}. The relative predictive values can be then estimated as

$$r\widehat{PPV} = \frac{\widehat{PPV}_{X_1}}{\widehat{PPV}_{X_2}} = \frac{(n_5 + n_6)(n_1 + n_3 + n_5 + n_7)}{(n_5 + n_7)(n_1 + n_2 + n_5 + n_6)} \quad (1)$$

$$r\widehat{NPV} = \frac{\widehat{NPV}_{X_1}}{\widehat{NPV}_{X_2}} = \frac{(n_3 + n_4)(n_2 + n_4 + n_6 + n_8)}{(n_2 + n_4)(n_3 + n_4 + n_7 + n_8)}. \quad (2)$$

Starting with the observation that the joint distribution of $\{D, X_1, X_2\}$ is multinomial with probabilities p_k, we can apply the multivariate central limit theorem together with the delta method to show that

$$\frac{1}{\sqrt{n}} \begin{bmatrix} log\ r\widehat{PPV} - log\ rPPV \\ log\ r\widehat{NPV} - log\ rNPV \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{\Sigma} \right)$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_P^2 & \sigma_{PN} \\ \sigma_{PN} & \sigma_N^2 \end{bmatrix}.$$

The components of $\mathbf{\Sigma}$ are lengthy and complicated. This is particularly true of the covariance σ_{PN}. To obtain simpler expressions for these variance components, we used the multinomial-Poisson transformation which transforms the likelihood of the data into a Poisson likelihood with additional parameters ([9]). We provide the resulting expressions in the appendix.

We can estimate the variance of $\log r\widehat{PPV}$ and $\log r\widehat{NPV}$ with $\frac{\hat{\sigma}_P^2}{N}$ and $\frac{\hat{\sigma}_N^2}{N}$, respectively, replacing each of the components in the expressions for σ_P^2 and σ_N^2 with their empirical estimates. These estimates yield $100(1-\alpha)\%$ confidence intervals:

$$\log r\widehat{PPV} \quad \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_P^2}{N}} \tag{3}$$

$$\log r\widehat{NPV} \quad \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_N^2}{N}}. \tag{4}$$

By exponentiating (3) and (4) we obtain upper and lower confidence limits for $r\widehat{PPV}$ and $r\widehat{NPV}$.

A $100(1-\alpha)\%$ elliptical confidence region for $(rPPV, rNPV)$ is the set of $(rPPV,$ $rNPV)$ such that

$$\begin{bmatrix} \log r\widehat{PPV} - \log rPPV \\ \log r\widehat{NPV} - \log rNPV \end{bmatrix}^T \widehat{\sum}^{-1} \begin{bmatrix} \log r\widehat{PPV} - \log rPPV \\ \log r\widehat{NPV} - \log rNPV \end{bmatrix} \leq \chi^2_{2,\alpha}$$

where $\widehat{\sum}$ is the estimate of $\mathbf{\Sigma}$.

In the CF data there are 5054 (42%) patients who had at least one PEx in 1996 ($D = 1$). Further, 4972 patients had a PEx in 1995 ($X_1 = 1$) and 308 patients were positive for

P. Aeruginosa in 1995 ($X_2 = 1$). Measuring the predictive accuracy of 1995 PEx, we estimate $PPV = 0.73$ and $NPV = 0.80$. For *P. Aeruginosa*, we estimate $PPV = 0.60$ and $NPV = 0.58$. We find $rPPV = 1.22$ and $rNPV = 1.37$. Separate 90% confidence intervals for the $rPPV$ and $rNPV$ are (1.14, 1.34) and (1.35, 1.38) respectively. In contrast, Figure 1 shows the joint 90% confidence region for ($rPPV$, $rNPV$). In both cases the $rPPV$ and $rNPV$ are bounded well away from one indicating that 1995 PEx does a significantly better job of predicting 1996 PEx.

4 Sample size formulas

The confidence intervals in (3) and (4) can be used to test hypotheses about the $rPPV$ and $rNPV$. Concentrating first on the $rPPV$, write the null hypothesis as $H_{0(P)} : rPPV \leq \delta$ where $\delta = 1$ for a superiority study testing whether PPV_{X_1} is larger than PPV_{X_2}. For a non-inferiority study, to assess if PPV_{X_1} is not substantially less than PPV_{X_2} we take δ less than but close to 1 . H_0 is rejected if the lower confidence limit for $rPPV$ is larger than δ. Suppose we wish to design a study to have power $1 - \beta$ under an alternative hypothesis $H_1 : rPPV = \gamma$. That is, we seek to choose the sample size so that with probability $1 - \beta$ the lower confidence limit for $rPPV$ will exceed δ. Then

$$1 - \beta = P\left(\log \widehat{rPPV} - z_{1-\alpha}\sqrt{\frac{\sigma_P^2}{N}} > \log \delta \mid rPPV = \gamma\right) \tag{5}$$

Since $\frac{\log \widehat{rPPV} - \log rPPV}{\sqrt{\sigma_P^2/N}} \sim N(0,1)$ and $\log rPPV = \log \gamma$ under the alternative hypothesis, we can rewrite (5) as

$$1 - \beta = P\left(\frac{\log \widehat{rPPV}}{\sqrt{\frac{\sigma_P^2}{N}}} > \frac{\log \delta - \log \gamma}{\sqrt{\frac{\sigma_P^2}{N}}} + z_{1-\alpha}\right)$$

Solving for N we find

$$N = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{(\log(\gamma/\delta))^2}\sigma_P^2$$

We now need to insert a value for σ_P^2 to calculate N. Under $H_1,\ PPV_{X_1} = \gamma PPV_{X_2}$ yielding

$$\sigma_P^2 = \frac{2(p_7+p_3)\gamma PPV_{X_2}^2 + (-p_6 + p_5(1-\gamma))PPV_{X_2} + p_6 + p_7(1-3\gamma PPV_{X_2})}{(p_5+p_6)(p_5+p_7)}$$

Thus the required sample size is

$$\begin{aligned} N &= \left(\frac{(z_{1-\alpha} + z_{1-\beta})}{\log(\gamma/\delta)}\right)^2 \times \\ &\left(\frac{2(p_7+p_3)\gamma PPV_{X_2}^2 + (-p_6 + p_5(1-\gamma))PPV_{X_2} + p_6 + p_7(1-3\gamma PPV_{X_2})}{(p_5+p_6)(p_5+p_7)}\right) \end{aligned} \tag{6}$$

For testing $H_{0(N)} : rNPV \leq \delta$ similar calculations result in the sample size formula

$$\begin{aligned} N &= \left(\frac{(z_{1-\alpha} + z_{1-\beta})^2}{log\ (\gamma/\delta)}\right)^2 \times \\ &\left(\frac{-2(p_4+p_8)\gamma NPV_{X_2}^2 + (-p_3 + p_4 - \gamma(p_2-p_4))NPV_{X_2} + p_2 + p_3}{(p_2+p_4)(p_3+p_4)}\right) \end{aligned} \tag{7}$$

Both of these sample size formulas require specification not only of a predictive value for the second test, a threshold value δ under the null hypothesis and an assumed value for the $rPPV$ under the alternative hypothesis, but of some additional probabilities as well. Intuitively, the probabilities of test positivity, $P(X_1 = 1)$ and $P(X_2 = 1)$, and correlation between the tests enter into the variances of $r\widehat{PPV}$ and $r\widehat{NPV}$.

As an example, suppose we had two tests and were interested in testing whether the positive predictive value of the first test was superior to the positive predictive value of the second factor. We have previous data indicating that $PPV_{X_2} = 0.7$ and wish to conclude

that the first test is superior if the $rPPV = 1.2$. Here $\delta = 1$ and $\gamma = 1.2$. We estimate, from prior data say, that $p_3 = 0.07$, $p_5 = 0.2$, p_6=0, and $p_7 = 0.05$. For significance level $\alpha = 0.05$ with power=$1 - \beta = 0.90$, we calculate from (6) that 192 subjects are needed.

We conducted a small simulation study to determine how well the formula given in (6) worked in this example. Data was first simulated for 192 subjects under the null hypothesis, rejecting H_0 if the lower limit of the confidence interval for log $rPPV$ was greater than 0, and this process was repeated 1000 times. We then generated data under the alternative hypothesis, fixing $\gamma = 1.2$, and repeated the same steps. (More details on how we generated data can be found in Section 7.) Under both the null and alternative hypotheses, we explored varying the number of positive test results for each test while holding the sample size and PPVs fixed. In all situations we studied, the empirical size was always close to the nominal 0.05 level. As expected we found 90% power when the probabilities $\{p_3, p_5, p_6, p_7\}$ were chosen correctly (top row of Table 2). However the power was less in other scenarios. In particular, when the number of subjects positive on both tests (N_{++}) or on at least one test (N_P) decreased, the power decreased.

If paired pilot data are available then such would be used to estimate the entries in Table 1 for substitution into the sample size formulas. Often, however, only data for studies of single tests will be available, yielding tentative values for the marginal probabilities $P(X_j = 1)$, $P(D = 1|X_j = 1)$, $P(D = 1|X_j = 0)$, for $j = 1, 2$. Equivalently, these provide values for: $\{p_1+p_2,\ p_1+p_3,\ p_2+p_4,\ p_3+p_4\}$ and $\{p_5+p_6,\ p_5+p_7,\ p_6+p_8,\ p_7+p_8\}$. To complete the table and hence the sample size calculations, some measures relating to the joint distribution of the two tests need to be stipulated. For example $P(X_1 = 1, X_2 = 1|D = 0) = p_1/(p_1+p_2+p_3+p_4)$ and $P(X_1 = 1, X_2 = 1|D = 1) = p_5/(p_5 + p_6 + p_7 + p_8)$ would suffice. A small paired case-

control pilot study could yield estimates of these parameters.

5 Marginal regression models

As seen in Section 3, the relative positive and negative predictive values can be estimated directly from the data. Leisenring, Alonzo, and Pepe (2000) take a different approach using generalized linear models to compare predictive values. They reorganize their data into long form, which is to say that each subject has two data records, one record for each test. They define an indicator variable $Z = I\{\text{Test 1}\}$ denoting to which test the record belongs. Thus, a subject with data $\{D, X_1, X_2\}$ has two records, $\{D, X, Z = 1\}$ with $X = X_1$ and $\{D, X, Z = 0\}$ with $X = X_2$. They propose the positive predictive value model

$$g(P[D = 1|Z, X = 1]) = \alpha_P + \beta_P Z \tag{8}$$

and negative predictive value model

$$g(P[D = 1|Z, X = 0]) = \alpha_N + \beta_N Z. \tag{9}$$

Records with positive test results, i.e. for which $X = 1$, are used to fit model (8). Since $PPV_{X_1} = P(D = 1|Z = 1, X = 1)$ and $PPV_{X_2} = P(D = 1|Z = 0, X = 1)$, model (8) compares PPV_{X_1} and PPV_{X_2} with β_P quantifying the difference between them on the g scale. In contrast, by conditioning on records with negative test results, model (9) compares NPV_{X_1} and NPV_{X_2} with β_N quantifying the difference between them. The models can be fit separately or simultaneously using GEE. Note that subjects can contribute none, one, or two records to fitting the predictive value model. For example, if a subject tests positive on

both tests, he will contribute two observations to fitting model (8), while if he tests positive on only test 1 he will contribute only one record.

The interpretation of the βs depends upon the choice of the link function, g. Using the logit link function as Leisenring *et al.* do, makes comparisons in terms of the odds ratios of the predictive values. Then e^{β_P} is the ratio of the odds of disease given a positive result for the first test compared to the odds of disease given a positive result for the second test. If instead we use the natural logarithm as the link function, e^{β_P} is the relative positive predictive value. Although both scales of comparison are valid, our focus is on the relative predictive values in part because we feel they are more easily interpreted than odds ratios. Note also that the estimates of the $rPPV$ and $rNPV$ obtained from these models are identical to the estimates obtained using the formulas in Section 3. The standard error estimates can differ between the two methods, at least in finite samples. We discuss and explore this point further in Section 7.

These are marginal models, although not in the usual sense. The more familiar marginal models developed by Liang and Zeger (1986) [10] are appropriate in the situation where there are paired or multiple outcomes. In working with the predictive values, however, there is a single outcome, D, while the diagnostic tests, X_1 and X_2, are paired. A naive application of the Liang and Zeger marginal models to this data would involve modeling $P(X_j|D)$. Notice that this approach reverses the roles of the tests and D. It yields models appropriate for the sensitivity and specificity, but not for the predictive values. Using D as the outcome, the standard multivariate regression analysis would model $P(D|X_1, X_2)$. This approach does not allow direct comparison of $P(D|X_1)$ and $P(D|X_2)$. The models proposed by Leisenring *et al.* are instead marginal with respect to the covariate, facilitating comparison of these two

quantities by using the tests as covariates.

Fitting models (8) and (9) to the CF data using a log link function, we estimate the $rPPV$ comparing 1995 PEx to *P. Aeruginosa* is 1.22 with a 90% confidence interval of (1.15, 1.34). We estimate the $rNPV$ to be 1.37 with a 90% confidence interval of (1.35, 1.38). The estimates of the $rPPV$ and the $rNPV$ are the same as those presented in Section 3, as we expected. Also notice that the confidence intervals obtained here using the marginal regression approach differ only sightly from the confidence intervals obtained using the analytic variance estimates derived above. Testing for differences using the generalized score statistics described by Leisenring *et al.* indicates that both the PPV and NPV of 1995 PEx are significantly higher than the PPV and NPV of *P. Aeruginosa* with $p < 0.001$ in both cases.

6 Regression models using discordant pairs

McNemar's test is a standard way of analyzing paired (or matched) binary data. It is used to compare the sensitivities and specificities of two binary tests. McNemar's test is based on the idea that concordant pairs, pairs where both tests yield the same result (or where both subjects have the same exposure in a matched case-control study) contain no information, while discordant pairs, pairs where the two diagnostic tests yield different results, contain all the vital information. The test statistic uses only the discordant pairs. McNemar's test is not directly applicable to compare the predictive values of two tests, because the predictive values condition on the test result rather than disease status. The basic idea of using the discordant pairs, however, is a natural way of analyzing paired binary data. Here we explore

this idea and develop a new test statistic for comparing the predictive values based in a regression framework.

For discordant pairs we define a new variable W

$$W = \begin{cases} 0 & \text{if } X_1 = 1 \text{ and } X_2 = 0 \\ 1 & \text{if } X_1 = 0 \text{ and } X_2 = 1 \end{cases}$$

Using W, the model

$$g(P[D = 1 \mid W, X_1 \neq X_2]) = \gamma_0 + \gamma_1 W$$

provides a regression framework for testing for differences between the two tests. If the natural logarithm is used as the link function, $e^{\gamma_1} = \frac{P(D=1|X_1=0,X_2=1)}{P(D=1|X_1=1,X_2=0)}$ is the ratio of the probability of disease given a negative result on the first test and positive result on the second test to the probability of disease given a positive result on the first test and a negative result on the second. In general, a hypothesis test based on this model answers a different question than the one in which we are primarily interested. That is, we are interested in testing $H_{0(P)} : P(D = 1|X_1 = 1) = P(D = 1|X_2 = 1)$ (and similarly $H_{0(N)}$), but this new approach tests the null hypothesis that $H_D : \gamma_1 = 0$, i.e. $H_D : P(D = 1|X_1 = 1 \text{ and } X_2 = 0) = P(D = 1|X_1 = 0 \text{ and } X_2 = 1)$. Interestingly, however, we can show that when the marginal probabilities of the two tests are equal, when $P(X_1 = 1) = P(X_2 = 1)$, testing H_D is equivalent to testing $H_{0(P)}$ and $H_{0(N)}$. See the proof in Appendix B. Moreover, we can show that $H_{0(P)}$ and $H_{0(N)}$ are equivalent so there is only one hypothesis to test. (This proof is also included in Appendix B.)

Furthermore, when $P(X_1 = 1) = P(X_2 = 1)$ the unstandardized score statistics for testing the two hypotheses $H_{0(P)}$ and H_D are also equivalent. To see this, let n_{disc} denote the

number of discordant pairs and define $\bar{W} = \frac{1}{n_{disc}} \sum_{i=1}^{n_{disc}} W_i$, $m_i = X_{1_i} + X_{2_i} = 0,\ 1,$ or 2, the number of positive test results for the i^{th} individual, $\bar{Z} = \left(\sum_{i=1}^{N} I\{m_i > 0\} X_{2_i}\right) / \left(\sum_{i=1}^{N} m_i\right)$, the proportion of all positive test results contributed by X_2, and $\bar{D}^{(\beta_P)} = \left(\sum_{i=1}^{N} m_i D_i\right) / \left(\sum_{i=1}^{N} m_i\right)$. The standardized score statistic for testing H_D is

$$S(\gamma_1) = \frac{\left(\sum_{i=1}^{n_{disc}} \{W_i(D_i - \bar{D})\}\right)^2}{\bar{D}(1-\bar{D}) \sum_{i=1}^{n_{disc}} \{(W_i - \bar{W})^2\}}.$$

and the generalized score statistic for $H_{0(P)}$ derived by Leisenring *et al.*, written in different notation than presented in their paper, is

$$S(\beta_P) = \frac{\left(\sum_{i=1}^{N} \{I\{m_i > 0\} D_i (X_{2_i} - m_i \bar{Z})\}\right)^2}{\sum_{i=1}^{N} \{I\{m_i > 0\} (D_i - \bar{D}^{(\beta_P)})^2 (X_{2_i} - m_i \bar{Z})^2\}}$$

The unstandardized score statistic for testing $H_{0(P)}$, i.e. the square root of the numerator of $S(\beta_P)$, is the sum of two components:

$$\sum_{i=1}^{N} I\{m_i > 0\} D_i (X_{2_i} - m_i \bar{Z}) =$$

$$\sum_{i=1}^{N} I\{m_i = 1\} D_i (X_{2_i} - \bar{Z}) + \sum_{i=1}^{N} I\{m_i = 2\} D_i (1 - 2\bar{Z}) \tag{10}$$

where we have simply broken $I\{m_i > 0\}$ into the two possibilities $I\{m_i = 1\}$ and $I\{m_i = 2\}$ and then substituted the respective values for m_i into the formula. The first term contains only those individuals who have a positive result for a single test, i.e. the discordant pairs. The second term contains those individuals who have positive results for both tests. To see that this second term is equal to zero when the estimated marginal probabilities of the two tests are equal, notice that in this situation $\bar{Z}$ will always equal 0.5 because each test will

contribute the same number of positive test results ($\bar{Z} = \frac{\sum X_{2_i}}{\sum X_{1_i}+X_{2_i}}$). This result leaves only the first term containing the discordant pairs. Further manipulation of this first term, which we show in Appendix C, reveals that it can be rewritten as $\sum_{i=1}^{n_{disc}}\{W_i(D_i - \bar{D})\}$ which is the unstandardized score statistic for testing H_D. Thus, when the marginal probabilities of the two factors are equal in a particular data application, the unstandardized score statistics from the marginal GEE approach and the discordant pairs approach are exactly equal. Their variances, however, are not.

Equation (10) implies that

$$var\left(\sum_{i=1}^{n} I\{m_i > 0\}D_i(X_{2_i} - m_i\bar{Z})\right) =$$

$$var\left(\sum_{i=1}^{n} I\{m_i = 1\}D_i(X_{2_i} - \bar{Z})\right) + var\left(\sum_{i=1}^{n} I\{m_i = 2\}D_i(1 - 2\bar{Z})\right) +$$

$$2cov\left(\sum_{i=1}^{n} I\{m_i = 1\}D_i(X_{2_i} - \bar{Z}),\ \sum_{i=1}^{n} I\{m_i = 2\}D_i(1 - 2\bar{Z})\right) \qquad (11)$$

In Appendix D we show that the covariance term is zero leaving only the first two terms on the right side of (11). The implication of this result is that the variance of the score statistic from the marginal GEE approach will always be at least as large as the variance of the score statistic from the discordant pairs approach. Thus, $S(\gamma_1)$ is more efficient than $S(\beta_P)$ in this setting. Intuitively, when we use the discordant pairs score statistic to test $H_{0(P)}$ we assume that the marginal probabilities of the two diagnostic tests are equal and hence variability in the second component of (10) is not at issue. Therefore, $S(\gamma_1)$ has a narrower set of alternatives and is more powerful for these alternatives. Below we explore this idea further in a simulation study.

The gain in power when using this approach comes at the expense that it is only a valid way for testing for differences in the positive and negative predictive values in the restricted situations when $P(X_1 = 1) = P(X_2 = 1)$. This situation, however, can occur fairly often in practice. One example is when the two underlying tests are continuous and thresholds are chosen so that the same percentage of people in the population test positive on both tests. For instance, when comparing two screening tests that are inherently continuous in nature (such as cancer biomarker levels), we may be willing to devote resources to further testing on a certain percentage of individuals. This situation is not uncommon in a health policy analysis framework. We can set the thresholds for positivity on the two tests to yield this same percentage of "positive" individuals and compare the accuracy of the two tests using this methodology. In Reference [8] we discuss analyzing the predictive values of continuous diagnostic tests using this idea in more detail.

7 Simulation study

We conducted a simulation study to compare the three approaches described above. Simulating data to have specified marginal distributions, $P(D = 1|X_1)$ and $P(D = 1|X_2)$, while allowing each individual to have two binary tests but only a single outcome is not straightforward. We refer the interested reader to Leisenring *et al.* for more discussion of this point as well as a detailed description of how data can be generated in this circumstance. We simulated data in exactly the same way described there.

We had two aims in conducting this simulation study. Our first was to compare the estimates of the variances of $rPPV$ and $rNPV$ using the expressions derived in Section 3 with

those obtained from the regression approach of Leisenring *et al.* Although asymptotically the variances are the same, there is the possibility that the estimates may differ in smaller sample sizes.

For this purpose we fixed $P(D = 1) = .2$, $PPV_{X_2} = .7$, $NPV_{X_2} = .9$ and varied the $rPPV$ and $rNPV$. For some choices of PPV_{X_1} and NPV_{X_1} the simulated data resulted in $n_1 + n_2 = 0$, $n_1 + n_3 = 0$, $n_6 + n_8 = 0$, or $n_7 + n_8 = 0$. While we could still calculate the variance expressions in Σ for these situations, the GEE algorithm used to implement Leisenring *et al.*'s approach failed to converge. In these situations we employed a small sample correction, adding one observation to each cell of Table 1. We also slightly modified the approach suggested by Leisenring *et al.* by creating a new variable $\bar{D} = 1 - D$ and modeling $P(\bar{D}|X = 0, Z)$ in order to estimate the $rNPV$.

Our second aim in this simulation study was to compare the properties of the score statistic derived by Leisenring *et al.*, $S(\beta_P)$, with the score statistic based upon the discordant pairs approach, $S(\gamma_1)$. For the reasons discussed in Section 6 we expected that using $S(\gamma_1)$ would result in a more powerful test when the marginal distributions of the two tests are the same. For this aim in addition to specifying parameters described in the preceding paragraph, we also specified $P(X_j = 1)$ and set $P(X_1 = 1) = P(X_2 = 1)$.

Tables 3 and 4 show the results from the first aim. In general we found a very high degree of correlation between the standard errors produced by the two methods with the analytic expressions yielding slightly smaller estimates (Table 3). By comparing these standard errors with the standard error from the simulated distribution of the statistic (column titled "Actual") we see that for large sample sizes, both the analytic and GEE standard errors provide equally good estimates of the standard errors. In the smaller sample sizes, both the analytic

and GEE standard errors may over- or under-estimate the true standard error. For each scenario considered, the direction of this difference is the same for both estimates with the magnitude of the difference usually smaller for the estimates from the analytic expressions.

In addition, the coverage probabilities resulting from the two approaches are very close (Table 4) . For the larger sample sizes, both approaches yield coverage probabilities close to the nominal 90% level. For the smaller sample sizes, the coverage probabilities appear to deviate from the nominal level. In some situations the coverage probabilities are too high while in other situations they are too low. Generally the analytic variance expressions yield coverage probabilities slightly closer to the nominal level than do the variance expressions from GEE, but the difference is often minimal. Across all sample sizes, Tables 3 and 4 indicate that both variance estimates have similar properties. One does not appear to offer any real advantage over the other.

The simulation results for our second aim are contained in Tables 5 and 6. We see that the performance of the two statistics under the null hypothesis (Table 5) depends not so much on the overall sample size, but on the number of subjects that are ultimately included in the analysis denoted by N_P and N_{disc}. Observe that data for examining $\gamma_1 = 0$ is always a subset of that for examining $\beta_P = 0$, $N_{\text{disc}} \leq N_P$. For the larger sample sizes, both statistics maintain the 0.05 level. As the sample size decreases, though, $S(\beta_P)$ does a better job of maintaining Type I error closer to the nominal rate.

As we expected, it appears that $S(\gamma_1)$ is the more powerful statistic (Table 6). Despite the fact that it is based on a smaller effective sample size than is $S(\beta_P)$, it still outperforms $S(\beta_P)$. We conclude that in a given application, if (i) one is relatively certain that the marginal probabilities of the two factors are equal and (ii) numbers of discordant pairs are

large enough for large sample distribution theory approximations to apply, basing a test for equality of the two positive predictive values on the discordant pairs is preferable.

8 Discussion

In this paper we have developed two new ways of comparing the predictive values of two tests that are assessed using a paired study design and compared them with a previously proposed method by Leisenring *et al.* The first method involves directly estimating the $rPPV$ and $rNPV$ and using analytic variance estimates. It results in estimates that are the same as those of Leisenring *et al.* when a log link function g is used. The advantage of our approach is that analytic variance expressions give rise to sample size formulas for study design. For analysis, however, the marginal regression method has the advantage that it can easily accommodate adjusting for covariates. When there are factors that might affect the predictive accuracy of the two tests, we can include these factors and their interactions terms with the tests in our model. In this way one can test whether the factors significantly affect the predictive values of one or both of the tests and can estimate the predictive values for different scenarios defined by these factors. Leisenring *et al.* discuss this point further in their paper. A reasonable strategy in practice would be to design a study using our sample size formulas but to use the regression framework for analysis.

The second method developed here is motivated by the standard way of analyzing paired binary data and uses only the discordant pairs in the spirit of McNemar's Test. This method is easily implemented using simple widely available statistical procedures. In many situations, though, i.e. when $P(X_1 = 1) \neq P(X_2 = 1)$, the hypothesis being tested in this

method is different than the one in which we are interested. It does not directly pertain to the predictive values nor is it entirely clear that the question that is being answered by this approach is even relevant. In contrast, the direct approach and the approach of Leisenring *et al.* provide general valid methods for comparing predictive values in a paired study design.

We have shown that when the marginal probabilities of a positive result on the two diagnostic tests are equal, the discordant pairs approach does test for differences between the predictive values and does so more efficiently. A disadvantage of the discordant pairs approach, however, is that it cannot accommodate missing data; the values of both X_1 and X_2 must be known for a subject to be included in the analysis. In contrast, with the other two approaches if the result of one diagnostic test is missing for a given subject, information on their known test result can still be included in the analysis. These methods fundamentally are based on estimation of the individual predictive values, the marginal probabilities $P(D = 1|X_j)$ $j = 1, 2$, which utilize information from only a single diagnostic test (i.e. estimating PPV_{X_j} requires knowledge of only the result of the j^{th} test).

The biostatistical literature on test accuracy has focused on the diagnostic setting. But as we have emphasized, testing is done more broadly. In the cystic fibrosis data, for example, the purpose is prognostic, to predict occurrence of a future event. Predictive values are ultimately most important for quantifying the practical usefulness of a test. Moreover, statistical methodology is the same regardless of the application. However, the context for application does play a role in considering how and why predictive values of two tests may differ. In the diagnostic setting, where the purpose is to detect presence or absence of a condition, predictive values may differ because the tests differ in their sensitivities and specificities to the condition. In the prognostic context on the other hand, one thinks simply

in terms of risk inferred by factors yielding a positive test result.

Acknowledgements

This research was supported in part by the NIAID Clinical Research on AIDS Training Grant (5-T32-A17450-08) and by Grant R01 GM54438 from the National Institutes of Health.

References

[1] Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc.: New York, 2002.

[2] Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: New York, 2003.

[3] Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* 2001; **93**:1054–1061.

[4] Moons KGM, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clinical Chemistry* 2004; **50**(3):473–476.

[5] Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. *British Medical Journal* 1994; **308**:1552.

[6] Guggenmoos-Holzman I, van Houwelingen HC. The (in)validity of sensitivity and specificity. *Statistics in Medicine* 2000; **19**:1783–1792.

[7] Leisenring W, Alonzo TA, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* 2000; **56**:345–351.

[8] Moskowitz CS, Pepe MS. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics* 2004; **5**:113–127.

[9] Baker SG. The multinomial-poisson transformation. *The Statistician* 1994; **43**:495–504.

[10] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.

Appendix A

$$\begin{aligned}\sigma_P^2 &= \frac{1}{(p_5+p_7)(p_5+p_6)} \times \{p_6(1-PPV_{X_2}) + p_5(PPV_{X_2} - PPV_{X_1}) + \\ &\quad 2(p_7+p_3)PPV_{X_1}PPV_{X_2} + p_7(1-3PPV_{X_1})\} \\ \sigma_N^2 &= \frac{NPV_{X_2}(-p_3+p_4-2(p_4+p_8)NPV_{X_1}) + (p_2+p_3) - NPV_{X_1}(p_2-p_4)}{(p_2+p_4)(p_3+p_4)}\end{aligned}$$

Appendix B

Here we show that when $P(X_1 = 1) = P(X_2 = 1)$ testing H_D is equivalent to testing $H_{0(P)}$ and $H_{0(N)}$. That $H_{0(P)}$ and $H_{0(N)}$ are equivalent in this situation can be seen by

$$\begin{aligned}P(D=1|X_1=1) &= P(D=1|X_2=1) \\ \Rightarrow P(D=0|X_1=1) &= P(D=0|X_2=1)\end{aligned}$$

$$\Rightarrow \quad \frac{P(D=0)-P(D=0,X_1=0)}{P(X_1=1)} = \frac{P(D=0)-P(D=0,X_2=0)}{P(X_2=1)}$$

$$\Rightarrow \quad \frac{P(D=0|X_1=0)P(X_1=0)}{P(X_1=1)} = \frac{P(D=0|X_2=0)P(X_2=0)}{P(X_2=1)}$$

$$\Rightarrow \quad P(D=0|X_1=0) = P(D=0|X_2=0).$$

Proof of the equivalence of H_D and $H_{0(P)}$ is simplified by noticing that $P(X_2=1|X_1=1)=P(X_1=1|X_2=1)$ if and only if $P(X_1=1)=P(X_2=1)$. To verify this statement begin by supposing that $P(X_2=1|X_1=1)=P(X_1=1|X_2=1)$. Then $P(X_1=1,X_2=1)/P(X_1=1)=P(X_1=1,X_2=1)/P(X_2=1)$ implying that $P(X_1=1)=P(X_2=1)$. If instead we begin by supposing that $P(X_1=1)=P(X_2=1)$, then $P(X_1=1,X_2=1)/P(X_1=1)=P(X_1=1,X_2=1)/P(X_2=1)$ implying that $P(X_2=1|X_1=1)=P(X_1=1|X_2=1)$. Also notice that $P(X_2=1|X_1=1)=P(X_1=1|X_2=1)$ implies that $P(X_2=0|X_1=1)=P(X_1=0|X_2=1)$.

To show the equivalence of H_D and $H_{0(P)}$ we write,

$$P(D=1|X_1=1)= \tag{12}$$

$$P(D=1|X_1=1,X_2=1)P(X_2=1|X_1=1)+$$

$$P(D=1|X_1=1,X_2=0)P(X_2=0|X_1=1)$$

$$P(D=1|X_2=1)= \tag{13}$$

$$P(D=1|X_1=1,X_2=1)P(X_1=1|X_2=1)+$$

$$P(D = 1|X_1 = 0, X_2 = 1)P(X_1 = 0|X_2 = 1)$$

The first probability, $P(D = 1|X_1 = 1, X_2 = 1)$, on the right side of the equation is the same in both (12) and (13). Given the equivalence of $P(X_2 = 1|X_1 = 1)$ and $P(X_1 = 1|X_2 = 1)$ and of $P(X_2 = 0|X_1 = 1)$ and $P(X_1 = 0|X_2 = 1)$, if H_D is true it must also be that $H_{0(P)}$ is true. Conversely, if $H_{0(P)}$ is true it must also be that H_D is true.

Appendix C

To see that the first term on the right hand side of equation (10) is equivalent to the unstandardized score statistics for testing H_D when $P(X_1 = 1) = P(X_2 = 1)$, first note that $\bar{Z} = .5 = \frac{\sum_{i=1}^{n_{disc}} w_i}{n_{disc}}$ since $\frac{\sum_{i=1}^{n_{disc}} w_i}{n_{disc}}$ is just the fraction of positive results for X_2 among all discordant test results. This equality allows us to write

$$\begin{aligned}
&\sum_{i=1}^{n}\{I\{m_i > 0\}D_i(X_{2_i} - m_i\bar{Z})\} \\
&= \sum_{i=1}^{n_{disc}}\{D_i(X_{2_i} - \bar{Z})\} \\
&= \sum_{i=1}^{n_{disc}} D_i X_{2_i} - \sum_{i=1}^{n_{disc}} D_i\bar{Z} \\
&= \sum_{i=1}^{n_{disc}} D_i w_i - \left(\sum_{i=1}^{n_{disc}} D_i\right)\left(\frac{\sum_{i=1}^{n_{disc}} w_i}{n_{disc}}\right) \\
&= \sum_{i=1}^{n_{disc}} D_i w_i - \left(\frac{\sum_{i=1}^{n_{disc}} D_i}{n_{disc}}\right)\left(\sum_{i=1}^{n_{disc}} w_i\right) \\
&= \sum_{i=1}^{n_{disc}}\{w_i(D_i - \bar{D})\}
\end{aligned}$$

which is the unstandardized score statistic from the discordant pairs approach.

Appendix D

In this Appendix we show that the covariance term in (11) is zero. This task is greatly simplified by first proving that the expected value of the second term in (10) is zero when we assume that $P(X_1 = 1) = P(X_2 = 1)$.

Assume that $P(X_1 = 1) = P(X_2 = 1)$. Then

$$\sum_{i=1}^{n} E\left[I\{m_i = 2\}D_i(1 - 2\bar{Z})\right] = 0$$

Proof

$$\begin{aligned}
&\sum_{i=1}^{n} E\left[I\{m_i = 2\}D_i(1 - 2\bar{Z})\right] \\
&= \sum_{i=1}^{n} \Big\{P(m_i = 2)\ E\left[I\{m_i = 2\}D_i(1 - 2\bar{Z}) \mid m_i = 2\right] + \\
&\qquad P(m_i \neq 2)\ E\left[I\{m_i = 2\}D_i(1 - 2\bar{Z}) \mid m_i \neq 2\right]\Big\} \\
&= \sum_{i=1}^{n} P(m_i = 2)\ E\left[D_i(1 - 2\bar{Z}) \mid m_i = 2\right] + 0 \\
&= \sum_{i=1}^{n} P(m_i = 2)\ E\left[(1 - 2\bar{Z})\ E(D_i \mid \bar{Z}, m_i = 2) \mid m_i = 2\right] \\
&= \sum_{i=1}^{n} P(m_i = 2)\ P(D \mid m_i = 2)\ E\left[(1 - 2\bar{Z}) \mid m_i = 2\right] \\
&= \sum_{i=1}^{n} P(m_i = 2)\ P(D \mid m_i = 2) \times 0
\end{aligned}$$

The next to last equality follows by assuming that $P(X_1 = 1) = P(X_2 = 1)$ once we know that an individual is positive for both diagnostic tests ($m_i = 2$), knowing the proportion of all positive results that belong to the second test does not add any additional information in determining and individual's outcome, D. Hence, conditional on $m_i = 2$, D and $\bar{Z}$ are independent. The last equality follows because $\bar{Z} = .5$ when $P(X_1 = 1) = P(X_2 = 1)$.

This result seems rather intuitive when we think of the second term in (10) as the part of the marginal regression score statistic that tests for a difference in the marginal probabilities of the two diagnostic tests. The expected value of this term is zero under its particular portion of the null hypothesis. Now we can easily show that the covariance term in (11) is zero by writing

$$
\begin{aligned}
& cov\left(\sum_{i=1}^{n} I\{m_i = 1\} D_i (X_{2_i} - \bar{Z}),\ \sum_{i=1}^{n} I\{m_i = 2\} D_i (1 - 2\bar{Z})\right) \\
= & \sum_{i=1}^{n} cov\left(I\{m_i = 1\} D_i (X_{2_i} - \bar{Z}),\ I\{m_i = 2\} D_i (1 - 2\bar{Z})\right) \\
= & \sum_{i=1}^{n} \Big\{ E\left[\left(I\{m_i = 1\} D_i (X_{2_i} - \bar{Z})\right)\left(I\{m_i = 2\} D_i (1 - 2\bar{Z})\right)\right] \\
& \qquad - \left(E\left[I\{m_i = 1\} D_i (X_{2_i} - \bar{Z})\right]\right)\left(E\left[I\{m_i = 2\} D_i (1 - 2\bar{Z})\right]\right) \Big\}
\end{aligned}
$$

An individual cannot have a positive result for exactly one factor and positive results for both factors at the same time, so $E\left[\left(I\{m_i = 1\} D_i (X_{2_i} - \bar{Z})\right)\left(I\{m_i = 2\} D_i (1 - 2\bar{Z})\right)\right] = 0$. This fact together with the above claim shows that the covariance term is zero.

Table 1: Data from a paired study design

$D = 0$

	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	n_1	n_2
$X_1 = 0$	n_3	n_4

$D = 1$

	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	n_5	n_6
$X_1 = 0$	n_7	n_8

Table 2: Empirical power for testing $H_0 : rPPV \leq 1$ with $N = 192$ and $\alpha = 0.05$. $PPV_{X_2} = .7$ and $\gamma = 1.2$ are fixed across all simulations while the average number of subjects positive on both tests, N_{++}, and the average number of subjects positive on at least one test, N_P, vary. Results shown are averages across 1000 simulations.

N_{++}	N_P	Power
61	87	0.90
57	85	0.89
52	83	0.85
49	82	0.82
37	72	0.72

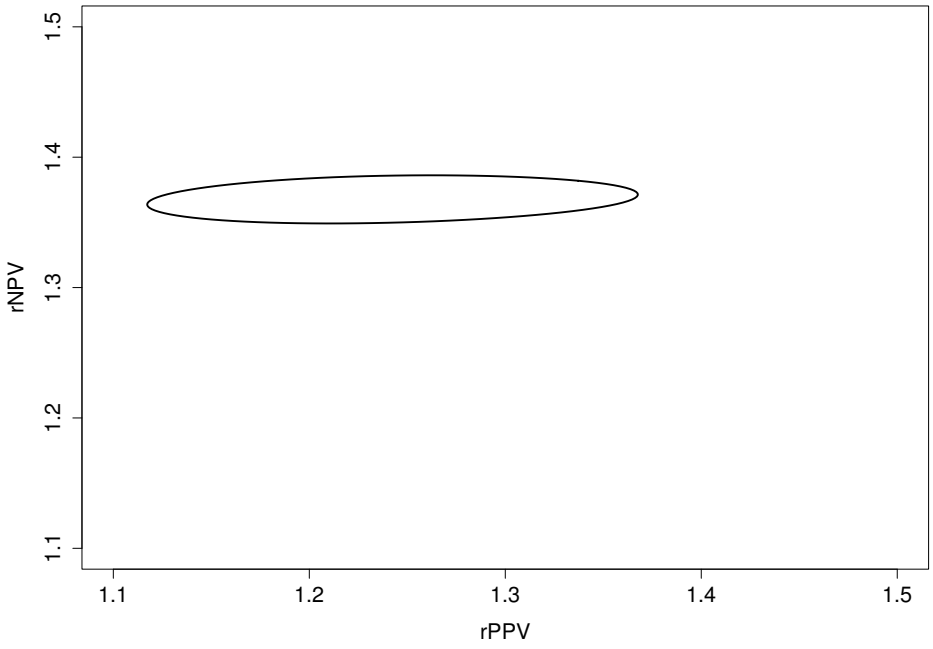

Figure 1: Predicting 1996 pulmonary exacerbations: elliptical confidence region comparing information on 1995 pulmonary exacerbations vs. a culture for *P. Aeruginosa*

Table 3: Comparison of the estimated standard errors of log $\widehat{rPPV}$ from the analytic variance expressions in Appendix A and the Leisenring *et al.* GEE approach. Results are based on 1000 simulations with $P(D=1)=.2,\ PPV_{X_2}=.7,$ and $NPV_{X_2}=.9$.

							Standard Error†		
N	$rPPV$	PPV_{X_1}	$rNPV$	NPV_{X_1}	$P(X_1=1)$	$P(X_2=1)$	Analytic	GEE	Actual
500	1.1	.77	1.0	.90	0.15	0.17	.0770	.0773	.0712
	1.1	.77	1.1	.99	0.25	0.17	.0696	.0699	.0692
	1.2	.84	1.0	.90	0.14	0.17	.0750	.0754	.0743
	1.2	.84	1.1	.99	0.23	0.17	.0690	.0693	.0703
100	1.1	.77	1.0	.90	0.15	0.17	.1806	.1851	.1802
	1.1	.77	1.1	.99	0.25	0.17	.1604	.1635	.1596
	1.2	.84	1.0	.90	0.14	0.17	.1764	.1810	.1724
	1.2	.84	1.1	.99	0.23	0.17	.1592	.1624	.1647
50	1.1	.77	1.0	.90	0.15	0.17	.2604	.2791	.2355
	1.1	.77	1.1	.99	0.25	0.17	.2034	.2094	.1579
	1.2	.84	1.0	.90	0.14	0.17	.2276	.2356	.1749
	1.2	.84	1.1	.99	0.23	0.17	.2053	.2111	.1662

† Presented is the average across simulations of the standard error estimates obtained from first the analytic variance and then the GEE approach, and the actual standard error of log $\widehat{rPPV}$ across the simulations.

Table 4: Comparison of the coverage probabilities for 90% confidence intervals obtained by using the variance estimates in Appendix A with those estimates obtained from the Leisenring *et al.* marginal GEE approach. Results are based on 1000 simulations with $P(D = 1) = .2,\ PPV_{X_2} = .7,$ and $NPV_{X_2} = .9$.

					$rPPV$		$rNPV$		(Correct†)
N	$rPPV$	PPV_{X_1}	$rNPV$	NPV_{X_1}	Analytic	GEE	Analytic	GEE	
500	1.1	.77	1.0	.90	89.3	89.7	90.2	90.9	(.00)
	1.1	.77	1.1	.99	89.8	90.0	90.0	90.0	(.02)
	1.2	.84	1.0	.90	90.9	91.0	89.5	89.4	(.00)
	1.2	.84	1.1	.99	89.4	89.5	90.1	90.1	(.02)
100	1.1	.77	1.0	.90	92.3	92.8	89.2	89.3	(.07)
	1.1	.77	1.1	.99	91.4	91.9	88.3	88.3	(.47)
	1.2	.84	1.0	.90	89.7	90.7	90.3	90.3	(.11)
	1.2	.84	1.1	.99	90.3	90.5	87.1	87.5	(.49)
50	1.1	.77	1.0	.90	93.1	93.8	90.1	91.6	(.43)
	1.1	.77	1.1	.99	96.8	96.9	89.4	89.7	(.73)
	1.2	.84	1.0	.90	95.2	95.9	94.2	94.5	(.40)
	1.2	.84	1.1	.99	95.3	96.0	89.4	89.6	(.76)

† Fraction of the simulations in which a small sample correction was made.

Table 5: Comparison of the empirical size of the score statistic, $S(\beta_P)$, derived by Leisenring *et al.*, and the score statistic from the discordant pairs approach, $S(\gamma_1)$. Results are based on 1000 simulations with $P(D = 1) = .2$.

N	$P(X_j = 1)$	PPV_{X_j}	$PDV_{1,0}$ [a]	$PDV_{0,1}$ [b]	N_P [c]	N_{disc} [d]	$P(S(\cdot) > \chi^2_{1,.95})$	
							$S(\beta_P)$	$S(\gamma_1)$
500	.1	.8	.66	.66	75	48	.0420	.0460
	.2	.8	.42	.42	126	53	.0420	.0460
	.3	.6	.16	.15	193	87	.0460	.0550
100	.1	.8	.67	.67	15	10	.0920	.0940
	.2	.8	.42	.41	25	11	.0620	.0750
	.3	.6	.15	.16	39	17	.0610	.1180
50	.1	.8	.66	.65	7	5	.2110	.3400
	.2	.8	.43	.40	13	5	.0830	.2710
	.3	.6	.16	.16	19	9	.0560	.2830

[a] Average $PDV_{0,1} = P(D = 1|X_1 = 1, X_2 = 0)$ across the simulations.

[b] Average $PDV_{1,0} = P(D = 1|X_1 = 0, X_2 = 1)$ across the simulations.

[c] Average number of subjects with at least one positive test result.

[d] Average number of subjects with discordant test results.

Table 6: Comparison of the empirical power of the score statistic, $S(\beta_P)$, derived by Leisenring *et al.*, and the score statistic from the discordant pairs approach, $S(\gamma_1)$. Results are based on 1000 simulations with $P(D = 1) = .2$.

N	$P(X_j = 1)$	PPV_{X_1}	PPV_{X_2}	$PDV_{1,0}$ [a]	$PDV_{0,1}$ [b]	N_P [c]	N_{disc} [d]	$P(S(\cdot) > \chi^2_{1,.95})$	
								$S(\beta_P)$	$S(\gamma_1)$
500	.1	.8	.5	.78	.28	80	60	.9660	.9750
	.2	.8	.5	.76	.10	145	90	1.0000	1.0000
	.2	.8	.7	.58	.27	132	64	.5540	.7090
100	.1	.8	.5	.77	.28	16	12	.3580	.4040
	.2	.8	.5	.75	.10	29	18	.7600	.8400
	.2	.8	.7	.56	.27	27	13	.1460	.1920
50	.1	.8	.5	.78	.28	8	6	.2270	.3610
	.2	.8	.5	.77	.10	14	9	.4510	.5700
	.2	.8	.7	.57	.29	13	6	.1330	.2270

[a] Average $PDV_{0,1} = P(D = 1|X_1 = 1, X_2 = 0)$ across the simulations.

[b] Average $PDV_{1,0} = P(D = 1|X_1 = 0, X_2 = 1)$ across the simulations.

[c] Average number of subjects with at least one positive test result.

[d] Average number of subjects with discordant test results.