Published in final edited form as:

Biometrics. 2016 September ; 72(3): 888–896. doi:10.1111/biom.12480.

Unbiased estimation of biomarker panel performance when combining training and testing data in a group sequential design

Nabihah Tayob*, Kim-Anh Do, and Ziding Feng

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston TX 77030, USA

Summary

Motivated by an ongoing study to develop a screening test able to identify patients with undiagnosed Sjögren's Syndrome in a symptomatic population, we propose methodology to combine multiple biomarkers and evaluate their performance in a two-stage group sequential design that proceeds as follows: biomarker data is collected from first stage samples; the biomarker panel is built and evaluated; if the panel meets pre-specified performance criteria the study continues to the second stage and the remaining samples are assayed. The design allows us to conserve valuable specimens in the case of inadequate biomarker panel performance. We propose a nonparametric conditional resampling algorithm that uses all the study data to provide unbiased estimates of the biomarker combination rule and the sensitivity of the panel corresponding to specificity of 1-t on the receiver operating characteristic curve (ROC). The Copas & Corbett (2002) correction, for bias resulting from using the same data to derive the combination rule and estimate the ROC, was also evaluated and an improved version was incorporated. An extensive simulation study was conducted to evaluate finite sample performance and propose guidelines for designing studies of this type. The methods were implemented in the National Cancer Institutes Early Detection Network Urinary PCA3 Evaluation Trial.

Keywords

Biomarker panel; Conditional estimation; Logistic regression; Shrinkage correction; Two-stage design

1. Introduction

The Salivary Biomarkers for Sjögren's Syndrome Detection- A multi-center study aims to develop a screening test that is able to identify patients with undiagnosed Sjögren's Syndrome in a symptomatic population. Sjögren's Syndrome is an autoimmune disease with symptoms of dry eyes and mouth and occurs most frequently in women. There are seven biomarkers under consideration; three that have shown potential in a preclinical validation

* ntayob@mdanderson.org.

6. Supplementary Materials

Web Appendix A referenced in Section 2.2 and Web Appendix B referenced in Section 3 are available with this paper at the Biometrics website on Wiley Online Library. The R-code and an illustrative example dataset are also available at the Biometrics website on Wiley Online Library.

study (Hu et al., 2010) and four salivary autoantibody biomarkers (Hu et al., 2011). The study consists of two-stages. In the first stage, data from the first 210 recruited patients will be analyzed. A biomarker panel will be built and evaluated. If the specificity of the panel corresponding to 95% sensitivity is not significantly less than 50%, then the samples from the remaining 210 patients are assayed and the completed study data will be used to refine the panel and evaluate its performance. If the panel of biomarkers do not show sufficient performance in the first stage data, the remaining 210 patients are recruited but their samples will not be assayed for this panel of biomarkers but preserved to identify a new set of biomarkers in the future. Currently, diagnosing primary Sjögren's Syndrome is clinically difficult and involves a lip biopsy and measurement of six criteria (Vitali et al., 2002; Shiboski et al., 2012). A biomarker panel that is able to reduce the number of unnecessary diagnostic tests for non-Sjögren's Syndrome patients by 50% while maintaining high sensitivity will be clinically important. This is an ongoing study with the interim analysis expected to take place in 2015 and the study is expected to end in 2016.

This ongoing study, and others like it, provide the motivation for developing statistical methods to combine multiple biomarkers and evaluate their performance in a two-stage group sequential study. Two-stage group sequential studies that allow for early termination for futility are important and widely used in biomarker development because they allow us to conserve valuable sample specimens in the case of inadequate biomarker performance. In completed studies, obtaining an unbiased estimate of the biomarker combination and the performance using all the available data is not trivial.

Conditional unbiased estimation procedures of the performance of a single biomarker in a two-stage study design have been proposed. Pepe et al. (2009) developed bias-adjusted estimates for sensitivity and specificity for dichotomous biomarkers. They make no distributional assumptions about the receiver operating characteristic (ROC) curve. Koopmeiners et al. (2012) developed a general framework for conditional estimation in a two-stage study with continuous biomarkers. The methodology assumes that estimators based on stage 1 data and estimators based on the completed study data have an asymptotically joint bivariate normal distribution with an independent increments covariance structure. Koopmeiners and Feng (2011) showed that this is asymptotically true for $ROC(t)$, the sensitivity of a biomarker corresponding to a specificity of $1-t$ on the ROC curve.

Koopmeiners and Vogel (2013) proposed a two-stage study design that included a stopping rule for futility when studying and evaluating a biomarker panel. The panel is constructed and evaluated for futility using the training data and the biomarker combination rule is evaluated for future application using the validation data. It is customary to use training data to build a model for a biomarker panel and rely on independent validation data to evaluate the model. However, this approach is not efficient since it does not use all the available data to build the model or to evaluate the performance of the biomarker panel. An important reason for this compromise is that there is no existing methodology that can provide efficient, unbiased estimation of the biomarker combination rule and the biomarker panel performance, evaluated via the $ROC(t)$, based on all data generated in a group sequential setting. Our proposed methodology addresses this problem and has potential widespread application. For example, it would be advantageous to use all accumulated data to build a

robust model that will be locked down for an FDA registry trial. Our proposed methods provide efficient and unbiased estimates of biomarker panel performance for planning the FDA registry trial.

We develop a nonparametric conditional resampling algorithm that simultaneously provides unbiased estimates of the biomarker combination rule and $ROC(t)$ using all the data from a completed two-stage study. This methodology extends the work of Pepe et al. (2009) to multiple continuous biomarkers. An additional source of bias results from using the same data to derive the combination rule and estimate $ROC(t)$. This is of particular concern in studies with small sample sizes where the bias is more pronounced. Copas and Corbett (2002) provide a shrinkage correction for this bias and we improve on the correction before incorporating it into our algorithm. Therefore we are able to use all the data from the completed study to estimate the biomarker combination rule and its performance while addressing the two sources of bias.

An alternative approach was taken by Zhao et al. (2015). They employ an inner bootstrap procedure to estimate both the logistic regression parameters and the performance of the score using all the data, and then an outer resampling procedure, similar to what we consider, to obtain conditional estimates of the $ROC(t)$. They introduced a binormal assumption to reduce computation burden of the double resampling algorithm.

Our proposed methodology focuses on the estimation of $ROC(t)$, and analogously the $ROC^{-1}(t)$, since these measures relate directly to the intrinsic performance of a diagnostic test. The predictive accuracy of a biomarker combination rule can also be evaluated via the positive predictive value (PPV) and negative predictive value (NPV). These quantities are highly valued by clinicians as they relate directly to patient care but can be misleading since they depend on the prevalence of the disease. Our proposed methodology can be translated to the unbiased estimation of the PPV/NPV in a two-stage group sequential design. In the special case when prevalence of disease is fixed, the PPV and NPV can be directly calculated based on results for sensitivity and specificity (Koopmeiners et al., 2012). In the Sjögren's Syndrome study, where prevalence is considered fixed at 35%, the aim is to produce a biomarker panel with PPV of 51% and NPV of 95%. The study requires high NPV since we want to spare non-Sjögren's Syndrome patients from unnecessary and invasive diagnostic procedures, while ensuring most patients with negative tests are non-Sjögren's Syndrome.

In Section 2 we describe the proposed methodology for unbiased estimation of both the combination rule and the performance of a biomarker panel in a two-stage study design. The performance of our method is evaluated under different simulation settings including small sample sizes in Section 3. The methodology we have developed is illustrated using data from a prostate cancer biomarker study evaluating the performance of three validated prostate cancer markers, PSA, PCA3 and T2Erg. A discussion follows in Section 5.

2. Methods

2.1 Estimation of ROC(t) for a single biomarker

The $ROC(t)$ is a clinically meaningful measure of a biomarker's ability to discriminate between cases and controls. It is defined as the true positive rate corresponding to a false positive rate of t on the ROC curve for a single biomarker, i.e. $ROC(t) = S_D\{S_D^{-1}(t)\}$ where $S_D(u)$ is the true positive rate and $S_{\bar{D}}(u)$ is the false positive rate at threshold u . The empirical estimate of $ROC(t)$, in a single stage study design, is obtained by plugging in empirical estimates for the survival functions and therefore avoids distributional assumptions about the biomarker. The variance of the empirical estimate, $\widehat{ROC}(t)$, is

$$\sigma^2_{\widehat{ROC}(t)} = \frac{ROC(t)\{1-ROC(t)\}}{n_D} + \left[\frac{f_D\{S_D^{-1}(t)\}}{f_{\bar{D}}\{S_D^{-1}(t)\}} \right]^2 \frac{t(1-t)}{n_{\bar{D}}},$$

where f_D is the probability density function (pdf) of the biomarker in the cases, $f_{\bar{D}}$ is the pdf in controls and n_D and $n_{\bar{D}}$ are the number of cases and controls, respectively, used to calculate $\widehat{ROC}(t)$.

2.2 Estimation of ROC(t) for a biomarker panel

The simplest estimation of $ROC(t)$ when there are multiple screening markers is a two-step process. The multiple biomarkers are consolidated into a score using a logistic regression model and the empirical estimate of $ROC(t)$ is then calculated for the score. This procedure can be described more formally as follows. Suppose there are q continuous biomarkers under consideration that are denoted by the column vector $\mathbf{Y}_i$, where i indexes the $n = n_D + n_{\bar{D}}$ patients in the study. The disease status is denoted by D_i , where $D_i = 1$ if the patient is a case and $D_i = 0$ if the patient is a control. The logistic regression model to consolidate the multiple markers is given by

$$\text{logit}\{Pr(D_i=1|\mathbf{Y}_i)\} = \beta^T \mathbf{Y}_i. \quad (1)$$

Without loss of generality assume the first biomarker in $\mathbf{Y}_i$ is always 1 for all i , corresponding to the intercept term, so the number of informative biomarkers under consideration is $q-1$. The maximum likelihood estimate of β , a column vector of length q , is given by $\hat{\beta}$. The empirical estimate of $ROC(t)$ is then calculated for the observed score $\hat{u}_i = \hat{\beta}^T \mathbf{Y}_i$. It is well known that this estimate is an over-optimistic estimate of the panel performance with significant bias in small sample sizes.

Copas and Corbett (2002) developed methods to correct the overestimation of the ROC curve that occurs when multiple covariates are combined into a score using logistic regression and the performance of the score is then assessed in the same dataset. We give a brief description of the method here and provide details in Web Appendix A.

Most biomarker studies are interested in the performance of the biomarkers, were they to be used in clinical practice to detect disease. The empirical estimate of $ROC(t)$ for the logistic score, which Copas & Corbett refer to as the retrospective ROC curve, is an overestimate of the performance of the score in a future cohort. The bias results from using the same data to estimate the parameters of the logistic regression model and then the empirical true positive rates and false positive rates of the resultant score. The prospective ROC is defined to be the ROC curve that uses the future proportions of true and false positives expected in the target population (with the same distribution of biomarkers $\mathbf{Y}_i$ as the original sample) assuming β is known. This is the target ROC curve in most biomarker studies. Copas and Corbett (2002) derive a shrinkage correction for the empirical estimate of $ROC(t)$ by examining the expected difference between the retrospective and prospective ROC curves. After a series of asymptotic approximations, a closed form expression for the expected overestimation is approximately

$$F(u, \beta) = \frac{p_u}{n^{3/2}\bar{p}} \sum_{i=1}^n \left\{ d_i - \frac{1}{\bar{p}(1-\bar{p})d_i} \right\} \phi \left\{ \frac{n^{1/2}(u_i - u)}{d_i} \right\}, \quad (2)$$

where $u_i = \beta^T \mathbf{Y}_i$, $p_i = \exp(u_i) / \{1 + \exp(u_i)\}$, $\bar{p} = \sum_{i=1}^n p_i / n$, $p_u = \exp(u) / \{1 + \exp(u)\}$, $d_i^2 = \mathbf{Y}_i^T \Omega^{-1} \mathbf{Y}_i$ and $\Omega = \frac{1}{n} \sum_{i=1}^n p_i(1-p_i) \mathbf{Y}_i \mathbf{Y}_i^T$. The estimate of the correction uses the maximum likelihood estimate of the logistic regression parameters, $\hat{\beta}$, and is denoted by $F(u, \hat{\beta})$. The Copas & Corbett corrected estimate of the $ROC(t)$ is $\hat{S}_D(u) - F(u, \hat{\beta})$ where $u = \hat{S}_D^{-1}(t)$.

The first asymptotic approximation used in the Copas & Corbett derivation is that when the difference between the retrospective and prospective false positive rates are asymptotically 0, then $p_i \approx p_u$. This was used to simplify the expression for the difference between the retrospective and prospective true positive rates and was crucial in the derivation. The remainder of the derivation used linearization of complex functions and excluded higher order terms that were argued to be $\mathcal{O}(n^{-3/2})$ or less. The expression was further simplified by using the first approximation, $p_i \approx p_u$.

We examined the Copas & Corbett estimate in a single-stage simulation study design to better understand the finite sample size performance for between three and seven biomarkers. The simulation study design and results are given in Web Appendix A. The Copas & Corbett estimate of $ROC(t)$ underestimated the performance of a biomarker panel and we observed significant negative bias in small sample sizes, up to 10%.

We explored multiple alternative corrections to the ROC curve where we made fewer assumptions and included higher order terms, details given in Web Appendix A. We propose a bias correction $G(u, \beta)$ that has at most 3% bias for all sample sizes and for up to seven biomarkers in single-stage simulation studies. It follows the same derivation as the Copas & Corbett correction but does not use the final simplifying approximation $p_i \approx p_u$ and includes an adjustment as a function of q , the length of β . The other bias corrections considered in

Web Appendix A included higher order terms but did not have sufficient bias reduction in simulations, for vastly increasing complexity, to advocate their use.

We define our proposed bias correction

$$G(u, \beta) = \left(\frac{q-2}{q-1} \right) \frac{p_u}{n^{3/2}\bar{p}} \sum_{i=1}^n \left\{ \frac{d_i p_i (1-p_i)}{p_u (1-p_u)} - \frac{p_i}{d_i p_u \bar{p}} - \frac{1-p_i}{d_i (1-p_u)(1-\bar{p})} \right\} \phi \left\{ \frac{n^{1/2}(u_i - u)}{d_i} \right\}. \quad (3)$$

The estimated performance of a biomarker panel is then

$$\widehat{ROC}(t) = \hat{S}_D \{ \hat{S}_D^{-1}(t) \} - G \{ \hat{S}_D^{-1}(t), \hat{\beta} \}.$$

2.3 Two-stage design and proposed estimator

The two-stage study then proceeds as follows: stage 1 data are used to estimate $ROC(t)$, denoted by $\widehat{ROC}(t)_{stg1}$; if $\widehat{ROC}(t)_{stg1} < h$ then the study is terminated; and if $\widehat{ROC}(t)_{stg1} \geq h$ then the study continues to stage 2. The continuation criterion is denoted by the indicator function $C = I\{\widehat{ROC}(t)_{stg1} \geq h\}$. The stage 1 data is comprised of m_D cases and $m_{\bar{D}}$ controls; and the stage 2 data is comprised of $n_D - m_D$ cases and $n_{\bar{D}} - m_{\bar{D}}$ controls. The estimate of $ROC(t)$ using only stage 2 data is denoted by $\widehat{ROC}(t)_{stg2}$. This estimate is unbiased but inefficient since it does not use all the study data. By contrast, the naïve estimator that uses all the data from the completed study to combine the biomarkers and calculate the empirical estimate of $ROC(t)$, $\widehat{ROC}(t)_{all}$, is biased.

We propose a conditional resampling algorithm to calculate the conditional uniformly minimum-variance unbiased estimator (UMVUE), $E\{\widehat{ROC}(t)_{stg2} | \widehat{ROC}(t)_{all}, C=1\}$, when we have multiple continuous biomarkers. The conditional UMVUE is also called the Rao-Blackwell estimator since the Rao-Blackwell theorem is a key result used to prove that it is the conditional UMVUE by both Pepe et al. (2009) and Koopmeiners et al. (2012). Our estimate of $ROC(t)$ is calculated using all the study data and corrects for bias resulting from the conditional two-stage study design.

For $k = 1, \dots, K$ samples,

Draw stage 1 data: Draw a random sample of size m_D without replacement from the n_D cases and a random sample of size $m_{\bar{D}}$ without replacement from the $n_{\bar{D}}$ controls.

Evaluate stage 1 data: Using stage 1 data:

- Fit logistic regression model, equation (1), to obtain the parameter estimate $\hat{\beta}_{1k}$.
- Calculate the observed score.

- Estimate $ROC(t)$ using bias corrected estimator and denote it by

$$\widehat{ROC}(t)_{stg1,k^*}$$

Interim criterion: Calculate the continuation criterion.

$$C_k = I\{\widehat{ROC}(t)_{stg1,k} \geq h\}$$

Evaluate stage 2 data: Using the stage 2 data, which consists of the remaining $n_D - m_D$ cases and $n_{\bar{D}} - m_{\bar{D}}$ controls:

- Fit logistic regression model, equation (1), to obtain the parameter estimate $\hat{\beta}_{2k}$.
- Calculate the observed score.
- Estimate $ROC(t)$ using bias corrected estimator and denote it by

$$\widehat{ROC}(t)_{stg2,k^*}$$

The final estimates of β and $ROC(t)$ are

$$\hat{\beta}_{RB} = \frac{\sum_{k=1}^K \hat{\beta}_{2k} * C_k}{\sum_{k=1}^K C_k},$$

$$\text{and } \widehat{ROC}(t)_{RB} = \frac{\sum_{k=1}^K \widehat{ROC}(t)_{stg2,k} * C_k}{\sum_{k=1}^K C_k};$$

the averages of the stage 2 estimates for those samples that meet the continuation criterion.

The distributions of $\widehat{ROC}(t)_{RB}$ and $\hat{\beta}_{RB}$ are estimated using a bootstrap procedure. Each bootstrap sample is a random sample of size n_D drawn with replacement from the cases and a random sample of size $n_{\bar{D}}$ drawn with replacement from the controls. The first m_D and $m_{\bar{D}}$ form the stage 1 data for the cases and controls respectively. For each of the B bootstrap samples, we calculate the continuation criterion C_b and our estimator $\widehat{ROC}(t)_{RB,b}$, where $b = 1, \dots, B$. The estimates $\widehat{ROC}(t)_{RB,b}$ and $\hat{\beta}_{RB,b}$ from bootstrap samples with $C_b = 1$ depict the empirical distribution of $\widehat{ROC}(t)_{RB}$ and $\hat{\beta}_{RB}$ thus allowing the construction of bootstrap confidence intervals.

3. Simulation Results

Simulation studies were used to evaluate the performance of our method for a range of sample sizes. We begin by generating case-control data with three biomarkers. The biomarker measures in controls were generated from a multivariate normal distribution with mean $\mathbf{0} = [0, 0, 0]$ and covariance matrix Σ . The biomarker measures in cases were generated from a multivariate normal distribution with mean $\mu = [\mu_1, \mu_2, \mu_3]$ and covariance matrix Σ . The parameters for μ are chosen to give the required $ROC(0.2)$ and Σ is constructed so that each biomarker has a variance of 1, the correlation between the first and second biomarker is

set to be 0.7, between the first and third is set to be 0.5 and between the second and third is set to be 0.4. The biomarkers in the simulation are set to be moderately/strongly correlated with each other. We did not include higher correlations, since if the biomarkers were very strongly correlated then model selection would be necessary to avoid problems of severe multicollinearity. We did not observe much difference in the results when biomarkers were mildly correlated.

The simulation studies considered a wide range of sample sizes. The total number of cases and controls are assumed to be the same and $n_D = n_{\bar{D}} = 400, 200, 100$ or 50. The number of cases and controls in stage 1 are $m_D = m_{\bar{D}} = 200, 100, 50$ or 25, respectively. The interim criterion used was $C = 1$ if the upper limit of the 95% Wald confidence interval of $ROC(0.2)$ was greater than 0.7. For each sample size, the true $ROC(0.2)$ is varied between 0.55 and 0.7. When the true $ROC(0.2) = 0.55$, then it is less likely that a particular study will meet the interim criterion and we have a large percent of studies that terminate early for futility.

In Table 1 we compare the point estimates of $ROC(0.2)$. Three estimators are considered: the naïve estimator $\widehat{ROC}(t)_{all}$, the unbiased but inefficient $\widehat{ROC}(t)_{stg2}$ and our proposed Rao-Blackwell estimator $\widehat{ROC}(t)_{RB}$. The percent stopping early in Table 1 is the proportion of the 1000 simulation studies performed that did not meet the interim criterion. Empirical estimates in the table are based on those studies that did meet the interim criterion and were completed.

When there are moderate sample sizes, 200 cases and 200 controls in the completed study or larger, we observe that our proposed Rao-Blackwell estimator has minimal bias — comparable to the unbiased stage 2 estimator. In scenarios where the true $ROC(0.2) = 0.55$ and early termination for futility is likely, the naïve estimator $\widehat{ROC}(t)_{all}$ has the largest bias. The bias of $\widehat{ROC}(t)_{all}$ decreases as the percentage of simulation studies that terminate early decreases. When $ROC(0.2) = 0.70$ we observe the three estimators all have minimal bias for moderate study sizes. For all scenarios, the proposed estimator $\widehat{ROC}(t)_{RB}$ has smaller empirical standard error than the stage 2 estimator $\widehat{ROC}(t)_{stg2}$ indicating that it is more efficient. When $ROC(0.2) = 0.7$ the proposed estimator and the naïve estimator are almost equally efficient.

For small sample sizes, we observe that there is a slight overcorrection (negative bias) for both the proposed estimator $\widehat{ROC}(t)_{RB}$ and the stage 2 estimator $\widehat{ROC}(t)_{stg2}$. In Web Appendix B, we include simulation results for the $ROC(t)$ estimators that use the empirical estimate with no bias correction (Web Appendix Table 5), and the Copas & Corbett estimator (Web Appendix Table 6). When $n_D = n_{\bar{D}} = 50$, the Rao-Blackwell estimator with no bias correction has a positive bias between 4 and 10 percent. When we incorporate the Copas & Corbett correction, the negative bias is about 8% for all $ROC(0.2)$. Using our proposed bias correction the negative bias reduces to between 1 and 3 percent.

Table 2 compares point estimates of β from our proposed method to the true values. $\hat{\beta}_{RB}$ has minimal bias in moderate sample sizes, even when a large proportion of studies terminate

early. For small sample sizes, there is greater bias observed for some of the model coefficients. We observe up to 25% bias in our estimates of the model coefficients when the total number of cases and controls are 100 or fewer.

Web Appendix Table 7 contains the simulation results of the inefficient stage 2 estimates of the model parameters and Web Appendix Table 8 summarizes the biased estimates of the model parameters based on all the data. The estimates of β based on all the data are only slightly more biased than those using only stage 2 data suggesting that the parameter estimates of the combination score are less affected by the two-stage design since the conditions for continuing to stage 2 are directly based on $ROC(t)$ but indirectly related to the β s.

The previous simulations focused on a case-control study design with equal number of cases and controls. The proposed methodology is also evaluated in a prospective study design with unbalanced numbers of cases and controls. Specifically, where the probability of being a case is set to be 0.35, the proportion of cases expected in the Sjögren's Syndrome study. The total number of patients in the study is varied between 100 and 800, and as before we assume that 50% of patients are included in stage 1. The results are presented in Table 9 and Table 10 of Web Appendix B and are comparable to those observed in the 1:1 case-control study design. We observe similar levels of bias reduction in our proposed estimator compared to the nave estimator for the different sample sizes considered. The proposed estimator is also observed to be more efficient than the unbiased stage 2 estimator in this simulation setting. The negative bias of the proposed estimator is between 1 and 4 percent in simulations with small sample sizes (100 patients) and we conclude that the bias correction performs well in this scenario of unequal numbers of cases and controls.

The last set of simulations explore the validity of our proposed methodology in the Sjögren's Syndrome study. Here, the simulation study design assumes we have 210 patients in each stage, where the probability of being a case is set to be 0.35. The 7 biomarkers are generated from multivariate normal distributions with parameters set to ensure $ROC^{-1}(0.95)$ ranges between 0.50 and 0.65, i.e at 95% sensitivity, the false positive rate is varied between 50% and 65%. The interim criterion used is $C=1$ if the lower bound of the one-sided 95% confidence interval for $ROC^{-1}(0.95)$ is less than or equal to 50%. In Table 3, we observe that when the true specificity is 50%, 7% of the studies terminate at stage 1 and when the true specificity is 35%, 53.7% of the studies terminate at stage 1. The interim analysis criterion was chosen to be a relatively conservative stopping decision. In other words, a study would terminate early only when the interim data clearly indicates that the biomarker panel would not have the ability to meet the pre-specified performance criteria, similar to the rationale for early stopping in therapeutic trials. We show both the unadjusted and biased corrected estimate of $ROC^{-1}(0.95)$ using stage 1 data from the 1000 studies to evaluate the performance of our proposed bias correction. The unadjusted estimator induces between 9.1 and 11.4% negative bias while our bias corrected estimator has at most 1% negative bias. For simulation studies that are completed, our proposed estimator has less bias than the nave estimator and is more efficient than the unbiased stage 2 estimator for all true $ROC^{-1}(0.95)$. The results from this simulation study indicate it will be reasonable to apply our proposed methodology to the Sjögren's Syndrome study.

4. Data Application

4.1 Motivation

The development of the proposed statistical methods was motivated by the ongoing Sjögren's Syndrome Detection Trial. Since the data is not yet available, we use data from a completed prostate cancer trial (Wei et al., 2014) to illustrate the proposed method.

Prostate cancer is the second most common cancer among men, and mainly occurs in older men. Gleason score (range of 2–10) is a system of grading prostate cancer tissue samples from biopsy. Gleason scores of 6 and lower have low long-term mortality and therefore active surveillance is often recommended rather than anti-cancer therapies that have associated side effects. It could be argued that patients with low-risk cancer should not be detected by a surveillance strategy so as to reduce the harm due to over-treatment. Prostate cancer with Gleason scores of 7 and higher are more likely to spread outside the prostate and cause health problems and are usually treated with anti-cancer therapies or surgery.

Current surveillance strategies include digital rectal exams and measuring the amount of prostate-specific antigen (PSA) in the blood. Currently, prostate biopsies are often performed in patients with a PSA value of more than 4ng/ml. However PSA has poor specificity since there are many causes of high PSA, including an enlarged prostate and prostatitis. Surveillance using PSA has been criticized as leading to unnecessary biopsies in patients with no prostate cancer and over-treatment of low-risk prostate cancers resulting in excess morbidity. The use of multiple biomarkers to detect aggressive prostate cancer and determine whether a biopsy is necessary could potentially reduce the number of unnecessary prostate biopsies and improve the detection of aggressive prostate cancer.

4.2 Data

The National Cancer Institutes Early Detection Network Urinary PCA3 Evaluation Trial was a multicenter trial to evaluate the performance of prostate cancer antigen 3 (PCA3) for detection of prostate cancer. Biorepository samples collected were also available to assay urinary TMPRSS2:Erg (T2Erg) gene fusion. These two biomarkers measure abnormal genomic changes by prostate cancer cells and we study their performance when added to PSA. The trial, which included an interim analysis for PCA3 evaluation in its original study design, included 561 patients undergoing a prostate biopsy for the first time. Wei et al. (2014) examined the performance of PCA3 in both the initial and repeat biopsy patients and found improved diagnostic performance. Our attention is restricted to the initial biopsy patients for whom we have pre-biopsy measurements of PSA, PCA3 and T2Erg. 148 patients had a Gleason score ≥ 7 , 116 patients with Gleason score of 6 and 297 control patients with no prostate cancer found in biopsies.

4.3 Analysis

Our aim is to detect aggressive prostate cancer and at the same time to spare unnecessary biopsies, therefore we define cases as patients with Gleason score ≥ 7 . The control group consists of patients with no prostate cancer or with low-risk cancer, Gleason score of 6. We estimate the false positive rate of the biomarker panel associated with 95% sensitivity

($ROC^{-1}(0.95)$) to evaluate the performance. Although the continuation criteria of the original study was not based on the biomarker panel, we simulate a two-stage study design with an interim criterion based on a biomarker panel for illustration purposes. The continuation criterion used in our analysis is that the lower limit of the 95% confidence interval of $ROC^{-1}(0.95)$ must be less than 45%, reflecting our intention to detect aggressive cancers at a high rate and simultaneously reduce the number of unnecessary biopsies.

Samples from 62 cases and 157 controls were used in the interim analysis and form our stage 1 data. In these samples, we identify the following score for combining the three markers: $2.03 * \log(PSA) + 0.70 * \log(PCA3) + 0.22 * \log(T2Erg + 1)$. The estimate of $\widehat{ROC}^{-1}(0.95)$, which includes the shrinkage correction, is 0.55 (95% CI: 0.39 – 0.71) indicating sufficient performance of the biomarker panel and our study proceeds to stage 2. The remaining samples form our stage 2 data and consist of 86 cases and 256 controls.

In Table 4 we report the estimates and associated 95% confidence intervals obtained naïvely using all the data, using only the stage 2 data and the estimates from our proposed algorithm that uses all the data. We report the estimates of the false positive rate in the entire cohort of controls and separately for those with low-risk prostate cancer (Gleason score of 6) and those with no prostate cancer. All the estimates included the shrinkage corrections; however, due to the large sample size, we did not observe much difference in the estimates obtained with or without the correction. The estimate of $\widehat{ROC}^{-1}(0.95)$ applying our proposed methodology is 0.67. Note that the proposed estimate is slightly more efficient than the estimate obtained using the stage 2 data indicated by the narrower confidence interval but larger than the naïve estimator, which is biased. The estimated rule for combining the three markers is $1.35 * \log(PSA) + 0.54 * \log(PCA3) + 0.23 * \log(T2Erg + 1)$. The biomarker panel is able to avoid unnecessary biopsies in 33% of control patients since all patients in the study would be indicated to undergo a prostate biopsy based on current guidelines, such as elevated PSA or other indications that do not include the two additional biomarkers included in the panel. Among patients with no prostate cancer, we will avoid 43% of unnecessary biopsies and among patients with Gleason score of 6, we will avoid 9%.

5. Discussion

We have proposed a nonparametric conditional resampling algorithm for investigating the performance of a biomarker panel in a group sequential design. This work was motivated by the widespread use of two-stage designs, which allow early termination for futility, in biomarker development. Two-stage designs allow us to conserve valuable specimens when interim analyses indicate the biomarker panel does not have required levels of clinical utility. The five phases of biomarker development for early detection of cancer outlined by Pepe et al. (2001) are (1) biomarker discovery; (2) validating the biomarker can distinguish patients with and without cancer; (3) use of retrospective preclinical samples to determine if the biomarker can distinguish who will and will not develop cancer; (4) prospective studies where decision rule determined in previous phases is implemented; and (5) determine if screening using the biomarker reduces the burden of cancer on the population. Our methodology can be used in phase (2) & (3) studies where the aim is to evaluate the

performance of the biomarker panel and prepare for phase (4) trials that require a decision rule.

The development of the biomarker combination rule and evaluation of its performance are often dual goals in biomarker studies for practical reasons. Patient specimens are valuable resources and in each study the aim is to extract as much information as possible from the data. Our proposed methodology for two-stage designs is able to accomplish both goals with negligible bias and efficient estimation when we have moderate to large sample sizes, at least 100 cases and 100 controls in each stage. With large sample sizes the naïve approach, which ignores the two-stage design, gives positively biased estimates of the biomarker panel performance and is not recommended.

In small sample sizes our methodology gives a conservative estimate of the performance, i.e. underestimates the true $ROC(t)$, so it may conclude a biomarker panel does not have clinical utility to move to the next phase, when in fact it does. The naïve approach occasionally has negligible bias when the two sources of biases, which have different directions, are of the same magnitude but in general it is also not appropriate for small sample sizes. Even if the biomarker panel achieves clinically useful performance criteria, the parameter estimates for the biomarker combination rule will have large bias in small sample sizes and may not be the optimal rule for validation studies. It is well known that the maximum likelihood estimates from logistic regression can have appreciable bias in small sample sizes Cordeiro and McCullagh (1991).

Based on our simulation results, we caution against developing the biomarker combination rule and evaluating its performance in a single study when the number of available samples is very small in each stage. In our single stage simulation study (Web Appendix A) we found that for 3 biomarkers we require more than 50 cases and 50 controls to get reasonable estimates for the parameters, and for 7 biomarkers we require at least 200 cases and 200 controls. These results assume biomarkers are multivariate normally distributed and as such can be used as guidelines of the sample size required at each stage when the biomarkers are either multivariate normal or can be considered multivariate normal after a suitable transformation. We recommend further simulation studies to assess the sample size required at each stage in studies where assumptions of multivariate normality are not reasonable.

While we have developed methodology that allows for simultaneous estimation of the biomarker combination rule and the performance of the panel in a two-stage design, the methods should be implemented with caution in small sample sizes. When designing phase (2)–(4) studies for a biomarker panel, it is important to ensure sufficient sample sizes are available to obtain credible results.

We briefly discuss the choice of stopping rule here; full consideration of study design when using group sequential methods for biomarker evaluation is an area of future research. The stopping rule for the two-stage design should be based on the clinical study under consideration and the rationale that one should not stop easily, i.e. to be conservative. In these retrospective studies, where the biomarker assays do not affect patient management, there is no ethical motivation to stop the trial early. Rather, biomarkers with true clinical

utilities are rare and we do not want to prematurely reject these biomarkers. Early stopping at the interim analysis should be to conserve the remaining valuable sample specimens when the biomarker panel demonstratively does not meet performance criteria. Early termination is desirable under the null hypothesis (biomarker panel has no clinical value) but undesirable under the alternative hypothesis (biomarker panel has true clinical utility). The proposed Rao-Blackwell estimator has been evaluated in a range of simulation studies and we have found that it performs well regardless of the proportion of studies terminating early.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the National Cancer Institutes Early Detection Network Urinary PCA3 Evaluation Trial for use of their data. Kim-Anh Do is partially supported by a Cancer Center Support Grant (NCI Grant P30CA016672). Ziding Feng is partially supported by EDRN grant (NCI Grant U24086368). Nabihah Tayob is partially supported by start-up and incentive funds to Kim-Anh Do and Ziding Feng.

References

- Copas JB, Corbett P. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*. 2002; 89:315–331.
- Cordeiro GM, McCullagh P. Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1991; 53:629–643.
- Hu S, Gao K, Pollard R, Arellano-Garcia M, Zhou H, Zhang L, Elashoff D, Kallenberg CG, Vissink A, Wong DT. Preclinical validation of salivary biomarkers for primary Sjögren's syndrome. *Arthritis Care Res (Hoboken)*. 2010; 62:1633–1638. [PubMed: 20617533]
- Hu S, Vissink A, Arellano M, Roozendaal C, Zhou H, Kallenberg CG, Wong DT. Identification of autoantibody biomarkers for primary Sjögren's syndrome using protein microarrays. *Proteomics*. 2011; 11:1499–1507. [PubMed: 21413148]
- Koopmeiners JS, Feng Z. Asymptotic properties of the sequential empirical roc, ppv and npv curves under case-control sampling. *The Annals of Statistics*. 2011; 39:3234–3261. [PubMed: 24039313]
- Koopmeiners JS, Feng Z, Pepe MS. Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. *Statistics in Medicine*. 2012; 31:420–435. [PubMed: 22238117]
- Koopmeiners JS, Vogel RI. Early termination of a two-stage study to develop and validate a panel of biomarkers. *Statistics in Medicine*. 2013; 32:1027–1037. [PubMed: 23413213]
- 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. [PubMed: 11459866]
- Pepe MS, Feng Z, Longton G, Koopmeiners J. Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. *Statistics in Medicine*. 2009; 28:762–779. [PubMed: 19097251]
- Shiboski SC, Shiboski CH, Criswell L, Baer A, Challacombe S, Lanfranchi H, Schiodt M, Umehara H, Vivino F, Zhao Y, Dong Y, Greenspan D, Heidenreich AM, Helin P, Kirkham B, Kitagawa K, Larkin G, Li M, Lietman T, Lindegaard J, McNamara N, Sack K, et al. American college of rheumatology classification criteria for sjögren's syndrome: a data-driven, expert consensus approach in the sjögren's international collaborative clinical alliance cohort. *Arthritis Care Res (Hoboken)*. 2012; 64:475–487. [PubMed: 22563590]
- Vitali C, Bombardieri S, Jonsson R, Moutsopoulos HM, Alexander EL, Carsons SE, Daniels TE, Fox PC, Fox RI, Kassan SS, Pillemer SR, Talal N, Weisman MH. Classification criteria for sjgren's

syndrome: a revised version of the european criteria proposed by the american-european consensus group. *Annals of the Rheumatic Diseases*. 2002; 61:554–558. [PubMed: 12006334]

Wei JT, Feng Z, Partin AW, Brown E, Thompson I, Sokoll L, Chan DW, Lotan Y, Kibel AS, Busby JE, Bidair M, Lin DW, Taneja SS, Viterbo R, Joon AY, Dahlgren J, Kagan J, Srivastava S, Sanda MGA. Can urinary pca3 supplement psa in the early detection of prostate cancer? *Journal of Clinical Oncology*. 2014; 32:4066–4072. [PubMed: 25385735]

Zhao S, Zheng Y, Prentice RL, Feng Z. Two-stage biomarker panel study and estimation allowing early termination for futility. *Biostatistics*. 2015 accepted.

Table 1

Simulation studies for three biomarkers. Results are based on the completed studies from 1000 simulated datasets. The empirical mean (empirical standard error) are given for each estimator of the $ROC(0.2)$.

True $ROC(0.2)$	% Early stopping	$\widehat{ROC}(0.2)_{all}$	$\widehat{ROC}(0.2)_{stg2}$	$\widehat{ROC}(0.2)_{RB}$
$n_D = n_{\bar{D}} = 400, m_D = m_{\bar{D}} = 200$				
0.55	81.7	0.5841 (0.0315)	0.5483 (0.0517)	0.5489 (0.0433)
0.60	53.7	0.6205 (0.0313)	0.6001 (0.0537)	0.6016 (0.0417)
0.65	15.2	0.6574 (0.0333)	0.6515 (0.0514)	0.6494 (0.0381)
0.70	1.8	0.7016 (0.0323)	0.6988 (0.0479)	0.6997 (0.0324)
$n_D = n_{\bar{D}} = 200, m_D = m_{\bar{D}} = 100$				
0.55	53.6	0.5772 (0.0459)	0.5480 (0.0735)	0.5475 (0.0569)
0.60	26.9	0.6129 (0.0468)	0.5959 (0.0736)	0.5957 (0.0552)
0.65	10.8	0.6535 (0.0462)	0.6455 (0.0701)	0.6459 (0.0498)
0.70	1.6	0.6986 (0.0447)	0.6957 (0.0650)	0.6960 (0.0443)
$n_D = n_{\bar{D}} = 100, m_D = m_{\bar{D}} = 50$				
0.55	32.7	0.5711 (0.0666)	0.5443 (0.1088)	0.5465 (0.0748)
0.60	17.1	0.6092 (0.0670)	0.5967 (0.1057)	0.5930 (0.0733)
0.65	6.9	0.6517 (0.0664)	0.6445 (0.0996)	0.6428 (0.0686)
0.70	2.2	0.6976 (0.0651)	0.6943 (0.0937)	0.6936 (0.0622)
$n_D = n_{\bar{D}} = 50, m_D = m_{\bar{D}} = 25$				
0.55	20.1	0.5640 (0.1005)	0.5430 (0.1465)	0.5470 (0.0981)
0.60	10.9	0.6008 (0.0958)	0.5912 (0.1420)	0.5899 (0.0931)
0.65	5.9	0.6430 (0.0910)	0.6363 (0.1369)	0.6356 (0.0864)
0.70	3	0.6837 (0.0850)	0.6801 (0.1301)	0.6795 (0.0799)

Table 2

Simulation studies for three biomarkers. Results are based on the completed studies from 1000 simulated datasets. The empirical mean (empirical standard error) are given for each estimator of the β , excluding the intercept.

True ROC(0.2)	True β	$\hat{\beta}_{RB}$
$n_D = n_{\bar{D}} = 400, m_D = m_{\bar{D}} = 200$		
0.55	$\begin{bmatrix} 0.13 & 0.56 & 0.48 \end{bmatrix}$	$\begin{bmatrix} 0.14(0.11) & 0.57(0.12) & 0.49(0.09) \end{bmatrix}$
0.60	$\begin{bmatrix} 0.77 & 0.18 & 0.31 \end{bmatrix}$	$\begin{bmatrix} 0.79(0.13) & 0.18(0.12) & 0.31(0.10) \end{bmatrix}$
0.65	$\begin{bmatrix} 1.13 & -0.04 & 0.22 \end{bmatrix}$	$\begin{bmatrix} 1.14(0.14) & -0.03(0.12) & 0.22(0.10) \end{bmatrix}$
0.70	$\begin{bmatrix} 1.44 & -0.22 & 0.13 \end{bmatrix}$	$\begin{bmatrix} 1.46(0.14) & -0.22(0.12) & 0.14(0.10) \end{bmatrix}$
$n_D = n_{\bar{D}} = 200, m_D = m_{\bar{D}} = 100$		
0.55	$\begin{bmatrix} 0.13 & 0.56 & 0.48 \end{bmatrix}$	$\begin{bmatrix} 0.12(0.17) & 0.58(0.17) & 0.49(0.13) \end{bmatrix}$
0.60	$\begin{bmatrix} 0.77 & 0.18 & 0.31 \end{bmatrix}$	$\begin{bmatrix} 0.79(0.18) & 0.19(0.16) & 0.32(0.14) \end{bmatrix}$
0.65	$\begin{bmatrix} 1.13 & -0.04 & 0.22 \end{bmatrix}$	$\begin{bmatrix} 1.16(0.20) & -0.03(0.17) & 0.22(0.14) \end{bmatrix}$
0.70	$\begin{bmatrix} 1.44 & -0.22 & 0.13 \end{bmatrix}$	$\begin{bmatrix} 1.48(0.21) & -0.23(0.18) & 0.14(0.15) \end{bmatrix}$
$n_D = n_{\bar{D}} = 100, m_D = m_{\bar{D}} = 50$		
0.55	$\begin{bmatrix} 0.13 & 0.56 & 0.48 \end{bmatrix}$	$\begin{bmatrix} 0.12(0.26) & 0.61(0.25) & 0.50(0.19) \end{bmatrix}$
0.60	$\begin{bmatrix} 0.77 & 0.18 & 0.31 \end{bmatrix}$	$\begin{bmatrix} 0.81(0.27) & 0.20(0.25) & 0.32(0.20) \end{bmatrix}$
0.65	$\begin{bmatrix} 1.13 & -0.04 & 0.22 \end{bmatrix}$	$\begin{bmatrix} 1.19(0.29) & -0.03(0.25) & 0.23(0.20) \end{bmatrix}$
0.70	$\begin{bmatrix} 1.44 & -0.22 & 0.13 \end{bmatrix}$	$\begin{bmatrix} 1.54(0.32) & -0.23(0.27) & 0.14(0.21) \end{bmatrix}$
$n_D = n_{\bar{D}} = 50, m_D = m_{\bar{D}} = 25$		
0.55	$\begin{bmatrix} 0.13 & 0.56 & 0.48 \end{bmatrix}$	$\begin{bmatrix} 0.14(0.38) & 0.65(0.39) & 0.54(0.31) \end{bmatrix}$

True $ROC(0.2)$	True β	$\hat{\beta}_{RB}$
0.60	$\begin{bmatrix} 0.77 & 0.18 & 0.31 \end{bmatrix}$	$\begin{bmatrix} 0.86(0.42) & 0.21(0.38) & 0.35(0.30) \end{bmatrix}$
0.65	$\begin{bmatrix} 1.13 & -0.04 & 0.22 \end{bmatrix}$	$\begin{bmatrix} 1.27(0.45) & -0.04(0.40) & 0.24(0.32) \end{bmatrix}$
0.70	$\begin{bmatrix} 1.44 & -0.22 & 0.13 \end{bmatrix}$	$\begin{bmatrix} 1.63(0.75) & -0.23(0.62) & 0.13(0.50) \end{bmatrix}$

Table 3

Simulation studies for seven biomarkers as in the Sjogren Syndrome Study, 210 subjects recruited in each stage. Probability of being a case is 0.35. Results are based on the completed studies from 1000 simulated datasets. The empirical mean (empirical standard error) are given for each estimator of the $ROC^{-1}(0.95)$. Estimators without * used bias correction.

True $ROC^{-1}(0.95)$	$\widehat{ROC}^{-1}(0.95)_{sig1*}$	$\widehat{ROC}^{-1}(0.95)_{sig1}$	% Early stopping	$\widehat{ROC}^{-1}(0.95)_{all}$	$\widehat{ROC}^{-1}(0.95)_{sig2}$	$\widehat{ROC}^{-1}(0.95)_{RB}$
0.50	0.4431 (0.1006)	0.4981 (0.1067)	7.0	0.4938 (0.0721)	0.5046 (0.1052)	0.5018 (0.0733)
0.55	0.4917 (0.1017)	0.5478 (0.1068)	16.5	0.5363 (0.0705)	0.5532 (0.1046)	0.5512 (0.0742)
0.60	0.5398 (0.1013)	0.5967 (0.1054)	32.1	0.5782 (0.0694)	0.6044 (0.1032)	0.6017 (0.0771)
0.65	0.5904 (0.0989)	0.6480 (0.1019)	53.7	0.6171 (0.0678)	0.6557 (0.0985)	0.6519 (0.0768)

Table 4
Results from the example data analysis. The estimate and 95% confidence interval (CI) are reported.

Subgroup	All data (naïve)		Stage 2 data		All data	
	$\widehat{ROC}_{all}^{-1}(0.95)$	95% CI	$\widehat{ROC}_{stage2}^{-1}(0.95)$	95% CI	$\widehat{ROC}_{RB}^{-1}(0.95)$	95% CI
All controls	0.62	0.53–0.71	0.67	0.57–0.77	0.67	0.57–0.76
Gleason Score=6	0.89	0.82–0.97	0.88	0.78–0.98	0.91	0.85–0.98
No Cancer	0.52	0.42–0.63	0.60	0.48–0.71	0.57	0.46–0.69