Assessing the incremental value of new biomarkers based on OR rules

LU WANG

Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N., Seattle, WA 98109, USA

ALEXANDER R. LUEDTKE

*Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N., Seattle, WA 98109, USA and
Department of Statistics, University of Washington, Seattle, WA 98195, USA*

YING HUANG*

*Fred Hutchinson Cancer Research Center, Seattle, 1100 Fairview Ave N., WA 98109, USA and
Department of Biostatistics, University of Washington, Seattle, WA 98195, USA*

yhuang@fhcrc.org

SUMMARY

In early detection of disease, a single biomarker often has inadequate classification performance, making it important to identify new biomarkers to combine with the existing marker for improved performance. A biologically natural method for combining biomarkers is to use logic rules, e.g., the OR/AND rules. In our motivating example of early detection of pancreatic cancer, the established biomarker CA19-9 is only present in a subclass of cancers; it is of interest to identify new biomarkers present in the other subclasses and declare disease when either marker is positive. While there has been research on developing biomarker combinations using the OR/AND rules, inference regarding the incremental value of the new marker within this framework is lacking and challenging due to statistical non-regularity. In this article, we aim to answer the inferential question of whether combining the new biomarker achieves better classification performance than using the existing biomarker alone, based on a nonparametrically estimated OR rule that maximizes the weighted average of sensitivity and specificity. We propose and compare various procedures for testing the incremental value of the new biomarker and constructing its confidence interval, using bootstrap, cross-validation, and a novel fuzzy p -value-based technique. We compare the performance of different methods via extensive simulation studies and apply them to the pancreatic cancer example.

Keywords: Bootstrap; Combining biomarkers; Cross-validation; Fuzzy p -value; Incremental value; OR/AND rules.

1. INTRODUCTION

In early detection of disease, a single biomarker often has inadequate classification performance. Identifying new biomarkers to combine with established predictors (biomarkers) for improved performance is an

*To whom correspondence should be addressed.

important research goal. For classification of binary diseases, a common modeling approach for combining biomarkers is using a likelihood-based logistic regression model, from which a marker combination score can be derived to subsequently generate a binary test based on a cut-off value. The use of logistic regression models has been well-studied in early detection; it yields optimal marker combination when the underlying risk model is correctly specified (McIntosh and Pepe, 2002), but may otherwise have suboptimal classification performance. Another commonly used approach in the applied literature for combining markers in a binary test is the use of logic rules (Etzioni and others, 2003), e.g., the “OR/AND” rules (Feng, 2010), which consider combination rules to be the set of “or-and” combinations of threshold rules in each biomarker. To declare an individual as disease-positive, the OR rule requires that either one or the other marker passes its individual threshold, while the AND rule requires that both markers pass their thresholds. For example, in early detection of pancreatic cancer, Tang and others (2015) considered a two-marker panel that declares disease if either the established biomarker CA19-9 exceeds a threshold OR a new discovered glycan marker exceeds a threshold. In a prostate cancer screening study, Gann and others (2002) showed that the addition of the ratio of free to total PSA (prostate-specific antigen) within a specific total PSA range with the OR/AND rules could simultaneously improve both specificity and sensitivity relative to the conventional strategy based on total PSA alone.

Logic combination rules are desirable for combining biomarkers mostly because of their simplicity and interpretability. For example, the OR rule is often preferred due to its biological appeal in detecting cancer, which is typically heterogeneous and composed of different subclasses. If biomarkers from each subclass can be identified, an OR rule combining these biomarkers is expected to boost the overall sensitivity without sacrificing much specificity. On the other hand, the AND rule is considered to be useful when individual biomarkers for combination have very high sensitivity and low specificity. Our research in this article is motivated by the development of biomarker combinations to improve early detection of pancreatic cancer. The current best marker for early detection of pancreatic cancer is the CA19-9 test, which detects the sialyl-Lewis A (sLeA) glycan. sLeA levels are not elevated in 25% of pancreatic cancers due to factors such as genetic inability. It is of interest to discover glycans other than sLeA that are overproduced in some pancreatic cancers that are low in sLeA. It is hoped that these glycans, when combined with CA19-9 using the OR rule (i.e., declaring a case if levels of either CA19-9 or the new marker are elevated), can improve the classification performance over CA19-9 alone (Tang and others, 2015). In other words, an important question that needs to be addressed is whether the new biomarker has significant incremental value when combined with an established biomarker, compared with using the established biomarker alone. While several authors have conducted statistical research on biomarker combinations using the OR/AND rules, the focus in the past has been mainly on algorithm development for finding the best combination instead of on making inference about the new biomarker’s incremental value. For example, Baker (2000) proposed a nonparametric multivariate algorithm that extended the idea of receiver-operating characteristic (ROC) cutpoints to multivariate positivity regions in order to find the optimal ROC curve. Etzioni and others (2003) considered classifying prostate cancers using OR/AND rules that combined total PSA with the ratio of free/total PSA; LOGIC regression (Ruczinski and others, 2003) was performed to find the best logic rule that maximizes the cross-validated weighted sum of sensitivity and specificity. Statistical research to answer the inferential question about the incremental value of a new biomarker, however, is lacking. As we will show next, this is a challenging problem due to the non-regularity of the incremental value estimator under the null hypothesis (i.e., when the new biomarker has no incremental value over the established biomarker). In this article, we aim to fill this gap. We will propose and compare various strategies for making inference regarding a new biomarker’s incremental value over an established biomarker. We consider a simple OR rule in this article for combining the established marker with the new marker, motivated by the pancreatic cancer example. However, the technique can be generalized to the AND rule or OR/AND rule combinations.

This article is organized as follows: In Section 2, we present an estimator for an OR rule that maximizes the weighted average of sensitivity and specificity, based on which the incremental value of the new marker is estimated. We develop procedures for testing the significance of the incremental value and for constructing its confidence interval (CI) utilizing the bootstrap, cross-validation, and a novel fuzzy p -value technique. In Section 3, we conduct extensive simulation studies to compare the performance of different methods. The application of the developed methods to the pancreatic cancer example is illustrated in Section 4. We finally make concluding remarks in Section 5.

2. METHODOLOGY

Let D be a binary disease outcome, with value 1 for diseased and 0 for non-diseased. Let X be an established biomarker (predictor) for predicting D and let Y be a new biomarker that we are interested in evaluating. The objective is to test whether combination of Y with X based on the OR rule offers any incremental value in classification performance over X alone and to estimate the incremental value. Without loss of generality, we assume larger marker values are associated with higher risk of disease. Suppose a case is declared if either X or Y is elevated, i.e., $X > c_1$ or $Y > c_2$, for some thresholds c_1 and c_2 . We define sensitivity as $Sen(c_1, c_2) = P(X > c_1 \text{ OR } Y > c_2 | D = 1)$ and specificity as $Spe(c_1, c_2) = P(X \leq c_1 \text{ AND } Y \leq c_2 | D = 0)$. For a test based on the biomarker X alone, sensitivity and specificity are defined as $Sen(c) = P(X > c | D = 1)$ and $Spe(c) = P(X \leq c | D = 0)$, respectively, for some threshold c . To characterize the incremental performance of the new biomarker Y , we consider the weighted average of sensitivity and specificity as an overall summary measure of a model's performance (Han and others, 2011), for pre-specified weight $w \in (0, 1)$, for either the model based on X alone or the model based on the combination of X and Y . The most common special case of the weighted average of sensitivity and specificity is the Youden's index (Youden, 1950), which weights a model's sensitivity and specificity equally. The equal weighting of sensitivity and specificity is oftentimes adopted in early stages of biomarker development when the objective is to discover a biomarker with good potential for future application based on a convenient case/control sample. This index will be adopted in our numerical studies in this article. In late phases of biomarker development, depending on the clinical practice where the biomarker will be utilized, weights for sensitivity/specificity can be chosen to reflect the relative importance of not missing the detection of a case versus not making false positive detection of a control, through a cost-benefit analysis. We define the incremental value of the new marker Y as the increase in the maximum value of the weighted average of sensitivity and specificity using the OR rule combining X and Y compared with the rule using X alone, i.e.,

$$\theta = \max_{c_1, c_2} [\omega Sen(c_1, c_2) + (1 - \omega) Spe(c_1, c_2)] - \max_c [\omega Sen(c) + (1 - \omega) Spe(c)], \quad (2.1)$$

with some weight $\omega \in (0, 1)$. For the rule based on X alone, the rule that maximizes the summary measure of performance, i.e., the weighted average of sensitivity and specificity, is searched in a one-dimensional biomarker space. When using the OR rule combining Y with X , the search is expanded to a two-dimensional biomarker space for the maximum performance measure. Since the optimization is performed over a larger space, the maximum possible value of the performance measure obtained using the OR rule is always greater than or equal to that using the rule based on X alone. That is, the rule based on $X > c$ can be represented as a special case of the OR rule with $c_1 = c$ and $c_2 = \infty$ (or the largest value for a bounded Y). Thus the incremental value θ (2.1) is always non-negative.

We note a connection between the maximizer of the weighted average of sensitivity and specificity and a logistic regression risk model (Ruczinski and others, 2003). Specifically, suppose the risk of the disease

conditional on the established biomarker X follows a logic regression model

$$P(D = 1 | X) = \alpha_0 + \alpha_1 I(X > c). \quad (2.2)$$

When conditions in **Result 1** below are satisfied, the threshold c in (2.2) will be the one that maximizes the weighted average of sensitivity and specificity in a binary classification rule based on X alone. Similarly, suppose the risk of disease conditional on X and Y follows a logic regression model

$$P(D = 1 | X, Y) = \beta_0 + \beta_1 I(X > c_1 \text{ OR } Y > c_2). \quad (2.3)$$

When conditions in **Result 2** below are satisfied, the thresholds c_1 and c_2 in (2.3) will be the corresponding thresholds for X and Y that maximize the weighted average of sensitivity and specificity in an OR rule.

Result 1 For a binary rule based on X that classifies an observation as diseased if $X > \delta$, suppose the threshold value c^* is the maximizer of the weighted average of sensitivity and specificity, i.e., $c^* = \arg \max_{\delta} [\omega \text{Sen}(\delta) + (1 - \omega) \text{Spe}(\delta)]$. If (2.2) holds, the CDF (cumulative distribution function) of X is not flat in a neighborhood of X , $\omega \frac{\alpha_0 + \alpha_1}{P(D=1)} - (1 - \omega) \frac{1 - \alpha_0 - \alpha_1}{P(D=0)} > 0$ and $\omega \frac{-\alpha_0}{P(D=1)} + (1 - \omega) \frac{1 - \alpha_0}{P(D=0)} > 0$, then c^* equals the parameter c indexing (2.2).

The proof of **Result 1** is given in Appendix A of the [supplementary material](#) available at *Biostatistics* online. When the risk model (2.2) holds, the weighted average of sensitivity and specificity in the rule based on $X > \delta$, i.e., $\omega \text{Sen}(\delta) + (1 - \omega) \text{Spe}(\delta)$, can be represented as a linear function of the CDF of the biomarker X at δ , for $\delta \geq c$ or $\delta \leq c$ separately. The coefficient of the CDF of X equals $-\omega \frac{\alpha_0 + \alpha_1}{P(D=1)} + (1 - \omega) \frac{1 - \alpha_0 - \alpha_1}{P(D=0)}$ for $\delta \geq c$, and equals $\omega \frac{-\alpha_0}{P(D=1)} + (1 - \omega) \frac{1 - \alpha_0}{P(D=0)}$ for $\delta \leq c$. As a result, the positivity condition of the coefficients stated in **Result 1** guarantees that the weighted average of sensitivity and specificity reaches the maximum at $\delta = c$. Assuming larger marker value is associated with higher risk, i.e., $\alpha_1 > 0$, then for a given risk model (2.2), the positivity condition is equivalent to

$$w \in \left(\frac{1 - \alpha_0 - \alpha_1}{1 - (\alpha_0 + \alpha_1) \left(1 - \frac{P(D=0)}{P(D=1)}\right)}, \frac{1 - \alpha_0}{1 - \alpha_0 \left(1 - \frac{P(D=0)}{P(D=1)}\right)} \right).$$

An example for a given model (2.2) is presented in Figure 1 of the [supplementary material](#) available at *Biostatistics* online, which demonstrates how weighted sensitivity and specificity changes with δ as w varies. Interestingly, the condition is always satisfied when $w = 1/2$, i.e., equal weight is given to sensitivity and specificity.

Result 2 For an OR rule based on X and Y that classifies an observation as diseased if $X > \delta_1$ or $Y > \delta_2$, suppose the threshold value (c_1^*, c_2^*) for X and Y is the maximizer of the weighted average of sensitivity and specificity, i.e., $(c_1^*, c_2^*) = \arg \max_{\delta_1, \delta_2} [\omega \text{Sen}(\delta_1, \delta_2) + (1 - \omega) \text{Spe}(\delta_1, \delta_2)]$. If (2.3) holds, the CDF of (X, Y) is not flat in the neighborhood of c_1, c_2 , $\omega \frac{\beta_0 + \beta_1}{P(D=1)} - (1 - \omega) \frac{1 - \beta_0 - \beta_1}{P(D=0)} > 0$ and $\omega \frac{-\beta_0}{P(D=1)} + (1 - \omega) \frac{1 - \beta_0}{P(D=0)} > 0$, then (c_1^*, c_2^*) equals parameters (c_1, c_2) indexing (2.3).

The proof of **Result 2** is given in Appendix B of the [supplementary material](#) available at *Biostatistics* online. Similarly as that in **Result 1**, we show that when the risk model (2.3) holds, the weighted average of sensitivity and specificity in an OR rule based on X and Y can be written as a linear function of the

joint CDF of biomarkers X and Y at (δ_1, δ_2) , for $\delta_1 \leq c_1$ and $\delta_2 \leq c_2$, $\delta_1 \leq c_1$ and $\delta_2 \geq c_2$, $\delta_1 \geq c_1$ and $\delta_2 \leq c_2$, or $\delta_1 \geq c_1$ and $\delta_2 \geq c_2$ separately. The positivity condition of the coefficients in **Result 2** ensures that the weighted average of sensitivity and specificity achieves its maximum at $\delta_1 = c_1$ and $\delta_2 = c_2$. Again assuming $\beta_1 > 0$, for a given model (2.3), the positive condition results in

$$w \in \left(\frac{1 - \beta_0 - \beta_1}{1 - (\beta_0 + \beta_1) \left(1 - \frac{P(D=0)}{P(D=1)}\right)}, \frac{1 - \beta_0}{1 - \beta_0 \left(1 - \frac{P(D=0)}{P(D=1)}\right)} \right),$$

which is always true for $w = 1/2$.

In general, even when the actual disease risk model conditional on biomarker(s) may not follow the conditions specified in Results 1 and 2, it is still appealing to identify classification rules based on X alone or an OR combination of X and Y by maximizing the weighted average of sensitivity and specificity, given that the weighted average of sensitivity and specificity is a clinically meaningful operational criterion of practical interest. So is the estimation of the incremental value of Y based on difference in model performance, as defined in equation (2.1).

2.1. Inference

2.1.1. Estimation. To estimate the incremental value θ , we consider nonparametric estimators of the classification rule based on either X alone or combinations of X and Y . In particular, we estimate threshold(s) in the corresponding rules by maximizing the weighted average of nonparametric estimates of sensitivity and specificity. Let subscripts D and $\bar{D}$ indicate case and control status, respectively, such that X_D and Y_D indicate biomarker measurements among cases and $X_{\bar{D}}$ and $Y_{\bar{D}}$ indicate biomarker measurements among controls. Let n_D and $n_{\bar{D}}$ be sample sizes for cases and controls, respectively. We compute $\hat{c}$ as the maximizer in c of $w \sum_{i=1}^{n_D} I(X_{Di} > c)/n_D + (1-w) \sum_{j=1}^{n_{\bar{D}}} I(X_{\bar{D}j} \leq c)/n_{\bar{D}}$. Similarly, we compute $\hat{c}_1$ and $\hat{c}_2$ as the maximizers in c_1 and c_2 of $w \sum_{i=1}^{n_D} I(X_{Di} > c_1 \text{ OR } Y_{Di} > c_2)/n_D + (1-w) \sum_{j=1}^{n_{\bar{D}}} I(X_{\bar{D}j} \leq c_1 \text{ AND } Y_{\bar{D}j} \leq c_2)/n_{\bar{D}}$. In our simulation studies and real data example, we adopt the grid-search method to obtain the maximizers $\hat{c}$, $\hat{c}_1$, and $\hat{c}_2$. For the rule based on X alone, we take the grid points to include all unique values of X from the sample; for the OR rule combining Y with X , the search space is set to include all combinations of unique values of X and Y from the sample. These grid searches are guaranteed to find the thresholds maximizing the empirical weighted average of sensitivity and specificity. Based on $\hat{c}$, $\hat{c}_1$, and $\hat{c}_2$, we then estimate θ nonparametrically as

$$\begin{aligned} \hat{\theta} = & \left[\frac{w}{n_D} \sum_{i=1}^{n_D} I(X_{Di} > \hat{c}_1 \text{ OR } Y_{Di} > \hat{c}_2) + \frac{1-w}{n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} I(X_{\bar{D}j} \leq \hat{c}_1 \text{ AND } Y_{\bar{D}j} \leq \hat{c}_2) \right] \\ & - \left[\frac{w}{n_D} \sum_{i=1}^{n_D} I(X_{Di} > \hat{c}) + \frac{1-w}{n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} I(X_{\bar{D}j} \leq \hat{c}) \right]. \end{aligned}$$

Note however this “naïve” estimator estimates the rule and its performance from the same dataset and thus is subject to overfitting bias. To reduce overfitting, a K-fold cross-validation method can be adopted instead. In performing cross-validation, first the dataset is split into K mutually exclusive and exhaustive subsets stratified on case/control status. Each time, one of the K subsets is used as the test set and the remaining $K - 1$ subsets are combined together to form a training set. The thresholds are estimated based on the training set and then they are used to obtain the incremental value estimator based on the k^{th} test

Fig. 1. Distributions of $\hat{\theta}$, $\hat{\theta} - \theta$, and $\hat{\theta}_{CV}$ based on 1000 simulated datasets, and distributions of $\hat{\theta}^*$, $\hat{\theta}^* - \hat{\theta}$, and $\hat{\theta}_{CV}^*$ over 1000 bootstrap samples, for $\theta = 0$ and 0.05 .

set, denoted by $\hat{\theta}_k$. The cross-validated estimator of incremental value is produced by taking average of the resulting K estimators, i.e.,

$$\hat{\theta}_{CV} = \frac{1}{K} \sum_{k=1}^K \hat{\theta}_k.$$

Next we investigate approaches to test the hypothesis that Y has significant incremental value when combined with X through an OR rule, i.e., to test

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0,$$

as well as approaches to construct the CI of the incremental value. We will propose a novel fuzzy p -value-based testing procedure and investigate various bootstrap-based approaches for inference.

2.1.2. Hypothesis testing. To perform the one-sided hypothesis test, we consider both a fuzzy p -value based approach and the bootstrap based approaches as described below.

A challenge with the test of incremental value in this problem setting is the non-regularity of the incremental value estimator under the null hypothesis. In other words, the naïve nonparametric estimator $\hat{\theta}$ is not asymptotically normal, so the standard testing procedure based on asymptotic normality of the test statistics is not applicable here. Figure 1 presents numerical examples of distribution of $\hat{\theta}$ for various

θ values. When the null hypothesis is true ($\theta = 0$), $\hat{\theta}$ is heavily right-skewed with a peak at zero. The distribution of $\hat{\theta}$ approaches normality as θ moves away from zero.

Fuzzy p -value Approach In this section, we propose a novel test for the incremental value of Y that leverages the fact that $n^{1/2}\hat{\theta}$ will converge to zero under the null given some regularity conditions. Let $\psi = \max_{c_1, c_2} [\omega \text{Sen}(c_1, c_2) + (1 - \omega) \text{Spe}(c_1, c_2)]$ and $\phi = \max_c [\omega \text{Sen}(c) + (1 - \omega) \text{Spe}(c)]$. Under some regularity conditions, the key of which is that the c maximizing $\omega \text{Sen}(c) + (1 - \omega) \text{Spe}(c)$ is unique, one can show that the nonparametric maximum likelihood estimator (NPMLE) $\hat{\phi}$ satisfies the following asymptotically linear expansion

$$\begin{aligned} \hat{\phi} - \phi = & \frac{1}{n} \sum_{i=1}^n \left(\omega \frac{I\{D_i = 1\}}{P(D = 1)} [I\{X_i > c^*\} - P(X > c^* | D = 1)] \right. \\ & \left. + (1 - \omega) \frac{I\{D_i = 0\}}{P(D = 0)} [I\{X_i \leq c^*\} - P(X \leq c^* | D = 0)] \right) + o_P(n^{-1/2}), \end{aligned} \quad (2.4)$$

where above $o_P(n^{-1/2})$ represents a term that converges to zero in probability once multiplied by $n^{1/2}$. A sketch of the argument for the above expansion is provided in Appendix C of the [supplementary material](#) available at *Biostatistics* online.

The term in the sum on the right-hand side of (2.4) represents the canonical gradient of the parameter ϕ (Bickel and others, 1998). Under the independent and identically distributed (i.i.d.) assumption, $\sqrt{n}(\hat{\phi} - \phi) \rightsquigarrow N(0, \sigma_\phi^2)$ as the sample size goes to infinity, where σ_ϕ^2 is the efficiency bound for regular and asymptotically linear estimators for ϕ within the nonparametric model. Here, we have made the simplifying assumption that the data are a sample of n i.i.d. observations, so that standard efficiency theory can be applied. Nonetheless, if a fixed number of cases and controls are sampled, then the dominant term above breaks into the sum of an empirical mean over cases and an empirical mean over controls, and the remainder term will remain negligible. Hence, central limit theorem results can be obtained in that case as well. Under similar regularity conditions to those needed for (2.4), the key of which is that the c_1, c_2 in the closure of the support of X, Y maximizing $\omega \text{Sen}(c_1, c_2) + (1 - \omega) \text{Spe}(c_1, c_2)$ are unique,

$$\begin{aligned} \hat{\psi} - \psi = & \frac{1}{n} \sum_{i=1}^n \left(\omega \frac{I\{D_i = 1\}}{P(D = 1)} [I\{X_i > c_1^* \text{ OR } Y_i > c_2^*\} - P(X > c_1^* \text{ OR } Y > c_2^* | D = 1)] \right. \\ & \left. + (1 - \omega) \frac{I\{D_i = 0\}}{P(D = 0)} [I\{X_i \leq c_1^* \text{ AND } Y_i \leq c_2^*\} - P(X \leq c_1^* \text{ AND } Y \leq c_2^* | D = 0)] \right) \\ & + o_P(n^{-1/2}). \end{aligned} \quad (2.5)$$

Hence, $\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N(0, \sigma_\psi^2)$. Under the assumptions needed for (2.4) and (2.5) to hold, the null hypothesis that $\theta = 0$ implies that c_2^* falls at the upper edge of the support for Y , in the sense that it is equal to the smallest number c_2 such that $P(Y > c_2) = 0$. In this case, the summation on the right-hand side above is equal to the summation on the right-hand side of (2.4), i.e. the right-hand sides of (2.4) and (2.5) are equivalent up to an $o_P(n^{-1/2})$ term. Because $\theta = 0$ under the null, we see that

$$\text{Under the null:} \quad n^{1/2}\hat{\theta} = n^{1/2}[\hat{\theta} - \theta] = n^{1/2}[\hat{\psi} - \psi] - n^{1/2}[\hat{\phi} - \phi] = o_P(1).$$

Now, using that $\hat{\psi}$ and $\hat{\phi}$ are consistent estimators, we also have that

Under a fixed alternative: $n^{1/2}\hat{\theta}$ diverges in probability,

i.e., for any fixed $t > 0$, $P(n^{1/2}\hat{\theta} > t) \rightarrow 1$ as $n \rightarrow \infty$.

We now use these facts to introduce a fuzzy p -value. Let F denote a cumulative distribution function for a continuous random variable on $\mathbb{R}$. By the above two facts, if Z is a random variable with cumulative distribution function F , under the null $p(Z) \equiv F(Z - n^{1/2}\hat{\theta})$ converges to a standard uniform random variable, whereas, under a fixed alternative, $p(Z)$ converges in probability to 0. Note that $p(Z)$ is a valid fuzzy p -value according to the definition given in Geyer and Meeden (2005). To generate a concrete decision, one could sample Z from F and reject if $p(Z) \leq \alpha$. Under the null hypothesis, the null will reject with probability approaching α . In our simulation, we use $\hat{\theta}_{CV}$ in computation of the fuzzy p -value to minimize over-fitting bias in small sample size; we let F equal to the CDF of the normal distribution $N(0, \sigma_{\psi}^2)$, where σ_{ψ}^2 is estimated with $\hat{\sigma}_{\psi}^2$ via the bootstrap. Alternatively, a Wald-type estimate of σ_{ψ}^2 could be obtained by computing the empirical variance of the term in the sum in (2.5), where here unknown probabilities and maximizing thresholds would be replaced by estimates.

Bootstrap Approach A commonly used approach for performing hypothesis tests is to construct bootstrap (Efron and Tibshirani, 1994) CIs for an estimand and evaluate whether the CI covers the parameter value specified in the null hypothesis. Here, we investigate different bootstrap methods to perform the hypothesis test about the incremental value raised by Y . We first consider bootstrap procedures based on the naive estimate of θ , because they are computationally simple and commonly used in practice. Both empirical and percentile bootstrap methods are considered. Suppose we have a data set of size n , from which we draw B random samples of size n with replacement, stratified on case/control status. Let $\hat{\theta}$ and $\hat{\theta}^*$ be the nonparametric incremental value estimates based on the original data set and the bootstrap samples, respectively. The one-sided $(1 - \alpha) \times 100\%$ empirical bootstrap CIs are constructed as $[2\hat{\theta} - \hat{\theta}_{1-\alpha}^*, \infty)$, where $\hat{\theta}_{1-\alpha}^*$ denotes the $1 - \alpha$ quantile of $\hat{\theta}^*$. The one-sided $(1 - \alpha) \times 100\%$ percentile bootstrap CIs are constructed as $[\hat{\theta}_{\alpha}^*, \infty)$. The one-sided test for the incremental value being greater than zero can be based on whether the lower bound of the $(1 - \alpha) \times 100\%$ one-sided bootstrap CI is above zero. Percentile bootstrap CI has been widely used in biomarker research for characterizing and comparing biomarker performances. However, its validity requires symmetry in the distribution of the estimator (Van der Vaart, 1998), which is clearly violated under the null hypothesis in our problem setting based on $\hat{\theta}$. In contrast, the rationale behind the empirical bootstrap is to approximate the distribution of $\hat{\theta} - \theta$ by the distribution of $\hat{\theta}^* - \hat{\theta}$. From Figure 1, the right tails of the distributions for $\hat{\theta}$ and $\hat{\theta}^*$ agree reasonably well, suggesting the potential of testing the incremental value based on the lower confidence limit of the one-sided empirical bootstrap CI based on $\hat{\theta}$. Nonetheless, we emphasize that we do not currently have theory supporting the validity of the bootstrap under the null, and therefore our simulation will serve as preliminary evidence for or against its validity. Hereafter, we refer to the approaches based on empirical bootstrap CI or percentile bootstrap CI of $\hat{\theta}$ as EB and PB, respectively.

From Figure 1, the distribution of the cross-validated estimate $\hat{\theta}_{CV}$ is close to normal under both the null and the alternative hypotheses, suggesting the potential of using the Wald-test for incremental value based on the cross-validated incremental estimate. Thus we also consider a one-sided Wald test for the incremental value. Let $SD(\hat{\theta}_{CV}^*)$ be the standard deviation of cross-validated incremental value estimate obtained using bootstrap resampling. The Wald test rejects the null if $\hat{\theta}_{CV}/SD(\hat{\theta}_{CV}^*) > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution. We refer to this approach as Wald.CV. Note that one might also consider combining other bootstrap procedures, e.g., percentile or empirical bootstrap, with cross-validated estimate of θ . Through numerical explorations, we found that when cross-validated

estimates are used, performance of those alternative bootstrap procedures is either comparable or inferior to Wald.CV (results omitted). Thus we choose to focus on presenting Wald.CV results in the rest of the article.

2.1.3. Estimation of two-sided CI. In practice, besides testing the incremental value of a new biomarker, it is also of interest to understand the uncertainty of the incremental value estimate. We consider the construction of a two-sided CI for increment value using the Wald.CV approach mentioned before. Specifically, the two-sided $(1 - \alpha) \times 100\%$ Wald CIs for θ are constructed as $\left[\hat{\theta}_{CV} - z_{1-\alpha/2} SD(\hat{\theta}_{CV}^*), \hat{\theta}_{CV} + z_{1-\alpha/2} SD(\hat{\theta}_{CV}^*) \right]$.

3. SIMULATION STUDY

In this section, we conduct simulation studies to compare the performance of the methods described in Section 2 for testing and making inference about a new biomarker's incremental value. Here, we consider equally weighted sensitivity and specificity as the classification performance measure and define ϕ and ψ as the optimal average sensitivity and specificity based on an established marker alone or based on its combination with a new marker with an OR rule. The incremental value θ is then defined as the difference between ψ and ϕ .

Let D be a binary disease outcome. Let X and Y be two biomarkers that are independently distributed; each follows the standard normal distribution. We consider two types of scenarios where the underlying true risk model is (i) a logic model for the risk of D conditional on X and Y : $P(D = 1 | X, Y) = \beta_0 + \beta_1 I(X > c_1 \text{ OR } Y > c_2)$ with thresholds c_1, c_2 and parameters β_0, β_1 and (ii) a logistic risk model $\text{logit}(P(D = 1 | X, Y)) = \beta_0^* + \beta_1^* X + \beta_2^* Y$, yet investigators have adopted the simple OR rule to combine X and Y . In both scenarios, we set disease prevalence $P(D = 1) = 0.1$ with appropriate selection of $c_1, c_2, \beta_0, \beta_1, \beta_0^*, \beta_1^*$, and β_2^* . For the logic model, the true incremental value of Y is 0 when $c_2 = \infty$, otherwise it is greater than 0. We consider five different values of the threshold for biomarker X , $c_1 = 0, 0.67, 0.84, 1.04$, and 1.34 , which correspond to the 50th, 75th, 80th, 85th, and 91th percentiles of the standard normal distribution, respectively. In the following, we select different values of β_0, β_1 or $\beta_0^*, \beta_1^*, \beta_2^*$ to achieve a wide variety of classification performance based on X alone (ϕ) or based on the OR combination of X and Y (ψ). We consider scenarios with classification performance of the established biomarker X equal to 0.6, 0.7, 0.8, or 0.9. A range of incremental values $\{\theta = 0, 0.05, \text{ or } 0.1\}$ are considered in our simulation studies.

We consider case/control samples with equal numbers of cases and controls $n_D = n_{\bar{D}} = 100, 250$, or 500 randomly sampled from the population, based on which we compute the naïve performance estimate $\hat{\theta}$ and the cross-validated estimate $\hat{\theta}_{CV}$ based on 10-fold cross-validation. For each setting, we evaluate bias of $\hat{\theta}$ and $\hat{\theta}_{CV}$. We compare four different methods to perform the one-sided test at significance level $\alpha = 0.05$ for incremental value with respect to Type I error rate and power: (i) EB: the method based on empirical bootstrap CI using $\hat{\theta}$, (ii) PB: the method based on percentile bootstrap CI using $\hat{\theta}$, (iii) Wald.CV, Wald test using $\hat{\theta}_{CV}$, and (iv) the fuzzy p -value approach. In addition, we examine coverage of the 95% two-sided CI of incremental value using the Wald.CV method. In each setting, 1000 Monte-Carlo simulations are conducted with 1000 bootstrap replicates constructed stratified on case/control status. The simulation results for various scenarios and sampling size at each fixed θ are summarized in Tables 1, 2, and 3. Corresponding results under the local alternative where $\theta \propto n_D^{-1/2}$ are presented in Table 1 in the [supplementary material](#) available at *Biostatistics* online.

From Table 1 below and Table 1 in the [supplementary material](#) available at *Biostatistics* online, under both correctly specified and mis-specified underlying models, we see that the naïve estimator could overestimate the new biomarker's performance when sample size is small. When the true incremental

value is small, the overestimation issue becomes less severe as sample size increases, as well as for settings with better performance of X alone (i.e., settings with larger ϕ value). Using cross-validation in general corrects this overestimation problem and can lead to small attenuation in some settings.

From Table 2, when the null hypothesis is true, the test based on PB often has inflated Type I error, whereas the corresponding test based on EB typically has Type I error rate smaller than the nominal level (e.g., when biomarker X itself has good performance). The test based on Wald.CV in general tends to be more conservative than EB. The fuzzy p -value method works reasonably well with Type I error fairly close to the nominal level for all settings considered. When the alternative hypothesis is true, the test based on EB generally has better or comparable power compared with other tests; the performance of Wald.CV and the fuzzy p -value method are more or less comparable to each other and their relative performance varies across settings.

From Table 3 and Table 1 in the [supplementary material](#) available at *Biostatistics* online, for the purpose of constructing two-sided CI for the incremental value, the Wald.CV approach is satisfactory for both underlying models. It clearly shows that the Wald.CV two-sided CIs have coverage either close to or slightly larger than the nominal level. Under the mis-specified model and alternative hypothesis, although the two-sided CIs can have slight undercoverage for the smaller sample size, their coverage approaches the nominal level as when the sample size is large enough, i.e., $n_D = n_{\bar{D}} = 500$.

Overall, we observe a similar pattern on performance comparison among different approaches under the (correctly specified) logic and (mis-specified) logistic risk models. For testing the significant incremental value of a new biomarker, the one-sided test based on empirical bootstrap CI is recommended; the fuzzy p -value approach is also desirable, given its theoretical foundation and reasonable performance. For making inference about the uncertainty of θ estimator, the Wald CI based on the cross-validated estimator is desired for constructing two-sided CI about θ .

4. PANCREATIC CANCER STUDY

In this section, we apply the proposed methods to a real data example from a pancreatic cancer study aimed at identifying biomarkers for early detection of pancreatic cancer. In this study, plasma samples were collected from $n_D = 109$ patients with pancreatic cancer and $n_{\bar{D}} = 91$ healthy individuals for biomarker measurement ([Tang and others, 2015](#)). The sLeA glycan, on which the CA19-9 assay is based, is currently the only established biomarker for pancreatic cancer detection. However, its performance for early detection of pancreatic cancer is not satisfactory given that it is not elevated in about 25% of pancreatic cancers. [Tang and others \(2015\)](#) found that sialyl-Lewis X (sLeX), a structural isomer of sLeA, was elevated in the plasma of 14–19% of patients with low sLeA. Thus, a biomarker panel combining sLeA and sLeX can potentially be useful in the clinical detection of pancreatic cancer. In this study, the estimated optimal average sensitivity and specificity based on sLeA alone is 0.683 (0.636 after cross-validation). Here, we estimate the incremental value of sLeX when combined with sLeA using an OR rule and test the hypothesis that a strategy combining the two biomarkers performs better than using the sLeA biomarker alone. The estimated naïve and cross-validated incremental values of sLeX are 0.079 and 0.062, respectively. We apply the EB, PB, and Wald.CV methods to conduct a one-sided test for incremental value of sLeX. We also apply the fuzzy p -value approach to the data and compute the average rejection rate over 1000 random draws. In addition, the two-sided CI is constructed based on Wald.CV.

When using the original data set, the one-sided tests based on EB and PB both reject the null hypothesis with lower limits of one-sided CI for incremental value of sLeX being 0.022 and 0.042, respectively, while the Wald.CV method fails to do so (p -value = 0.093). This finding is not surprising, given that EB and PB have been shown to be more powerful compared with Wald.CV. The fuzzy p -value approach rejects the null hypothesis with probability 0.67. The two-sided CI derived using Wald.CV is $[-0.030, 0.153]$.

Table 1. *Naïve and cross-validated (CV) estimates of incremental value and corresponding standard deviation (SD) in the parenthesis based on 1000 Monte Carlo simulations under different underlying models and scenarios. ϕ indicates the performance of biomarker X alone and θ indicates the incremental value*

Correctly specified model (logic model)											
θ	ϕ	c_1	c_2	β_0	β_1	Estimate (SD)			CV Estimate (SD)		
						$n_D = n_{\bar{D}}$			$n_D = n_{\bar{D}}$		
						100	250	500	100	250	500
0	0.6	0	∞	0.064	0.072	0.008 (0.010)	0.003 (0.004)	0.004 (0.004)	-0.012 (0.025)	-0.006 (0.010)	-0.005 (0.011)
0	0.6	0.84	∞	0.078	0.113	0.012 (0.005)	0.005 (0.006)	0.003 (0.003)	-0.009 (0.028)	-0.005 (0.012)	-0.003 (0.007)
0	0.7	0	∞	0.028	0.144	0.002 (0.004)	0.001 (0.001)	0.000 (0.001)	-0.008 (0.010)	-0.003 (0.003)	-0.002 (0.002)
0	0.7	0.84	∞	0.055	0.225	0.005 (0.007)	0.002 (0.003)	0.001 (0.001)	-0.010 (0.014)	-0.004 (0.006)	-0.002 (0.003)
0	0.8	0.67	∞	0.028	0.288	0.002 (0.003)	0.001 (0.001)	0.000 (0.001)	-0.007 (0.007)	-0.003 (0.003)	-0.001 (0.001)
0	0.8	0.84	∞	0.033	0.338	0.002 (0.004)	0.001 (0.002)	0.000 (0.001)	-0.007 (0.008)	-0.003 (0.003)	-0.001 (0.002)
0	0.9	1.04	∞	0.015	0.565	0.001 (0.002)	0.000 (0.001)	0.000 (0.000)	-0.006 (0.005)	-0.002 (0.002)	-0.001 (0.001)
0	0.9	1.34	∞	0.021	0.879	0.001 (0.003)	0.000 (0.001)	0.000 (0.001)	-0.006 (0.005)	-0.003 (0.002)	-0.001 (0.001)
0.05	0.6	0.84	1.15	0.061	0.129	0.054 (0.022)	0.051 (0.015)	0.051 (0.011)	0.040 (0.037)	0.047 (0.021)	0.048 (0.013)
0.05	0.7	0.84	1.53	0.040	0.240	0.053 (0.019)	0.051 (0.012)	0.051 (0.009)	0.044 (0.027)	0.048 (0.014)	0.049 (0.010)
0.05	0.8	0.84	1.73	0.018	0.352	0.051 (0.017)	0.050 (0.010)	0.050 (0.008)	0.044 (0.021)	0.048 (0.013)	0.049 (0.008)
0.05	0.9	1.04	2.01	0.003	0.577	0.050 (0.016)	0.050 (0.010)	0.050 (0.007)	0.045 (0.018)	0.048 (0.011)	0.049 (0.007)
0.1	0.6	0.84	0.67	0.040	0.150	0.099 (0.031)	0.099 (0.020)	0.100 (0.014)	0.096 (0.045)	0.098 (0.025)	0.100 (0.016)
0.1	0.7	0.84	1.15	0.023	0.257	0.099 (0.026)	0.100 (0.016)	0.100 (0.012)	0.095 (0.031)	0.099 (0.018)	0.099 (0.012)
0.1	0.8	0.84	1.38	0.002	0.368	0.099 (0.024)	0.100 (0.015)	0.100 (0.011)	0.096 (0.027)	0.099 (0.016)	0.099 (0.011)
Mis-specified model (logistic model)											
θ	ϕ	β_0^*	β_1^*	β_2^*	Estimate (SD)			CV estimate (SD)			
					$n_D = n_{\bar{D}}$			$n_D = n_{\bar{D}}$			
					100	250	100	250	100	250	
0	0.6	-2.284	0.47	0	0.009 (0.011)	0.004 (0.005)	0.002 (0.003)	-0.010 (0.028)	-0.005 (0.013)	-0.003 (0.007)	
0	0.7	-2.613	1.1	0	0.004 (0.006)	0.001 (0.002)	0.001 (0.001)	-0.008 (0.016)	-0.004 (0.006)	-0.002 (0.003)	
0	0.8	-3.358	1.94	0	0.001 (0.003)	0.001 (0.001)	0.000 (0.001)	-0.007 (0.009)	-0.003 (0.004)	-0.001 (0.002)	

Table 1. (Continued)

0	0.9	-5.501	3.90	0	0.001 (0.002)	0.000 (0.001)	0.000 (0.000)	-0.006 (0.005)	-0.002 (0.002)	-0.001 (0.001)
0.1	0.6	-2.680	0.5	1.05	0.111 (0.036)	0.108 (0.025)	0.105 (0.018)	0.099 (0.057)	0.103 (0.037)	0.100 (0.025)
0.1	0.7	-3.946	1.5	2.02	0.112 (0.035)	0.106 (0.023)	0.105 (0.016)	0.099 (0.053)	0.097 (0.033)	0.099 (0.022)
0.1	0.8	-19.534	10.5	10.9	0.108 (0.028)	0.106 (0.019)	0.104 (0.014)	0.098 (0.042)	0.101 (0.026)	0.100 (0.018)

Table 2. Type I error rate and power of one-sided test from the empirical bootstrap (EB), percentile bootstrap (PB), Wald using cross-validation (Wald.CV), and fuzzy p-value methods, under different underlying models and scenarios. ϕ indicates the performance of biomarker X alone and θ indicates the incremental value

Correctly specified model (logic model)							
θ	ϕ	c_1	$n_D = n_{\bar{D}}$	EB	PB	Wald.CV	Fuzzy p -value
Type I error rate							
0	0.6	0	100	0.015	0.132	0.006	0.056
			250	0.007	0.135	0.003	0.034
			500	0.023	0.344	0.008	0.055
		0.84	100	0.033	0.229	0.011	0.073
			250	0.015	0.256	0.002	0.042
			500	0.018	0.277	0.004	0.036
	0.7	0	100	0.001	0.009	0.001	0.038
			250	0.005	0.006	0.001	0.042
			500	0.003	0.011	0.001	0.047
		0.84	100	0.013	0.077	0.004	0.043
			250	0.012	0.087	0.000	0.041
			500	0.006	0.078	0.001	0.042
	0.8	0.67	100	0.001	0.007	0.001	0.031
			250	0.004	0.007	0.004	0.032
			500	0.002	0.011	0.000	0.036
		0.84	100	0.003	0.015	0.001	0.032
			250	0.005	0.019	0.001	0.034
			500	0.003	0.011	0.000	0.038
	0.9	1.04	100	0.000	0.002	0.002	0.026
			250	0.004	0.002	0.000	0.034
			500	0.000	0.001	0.000	0.037
		1.34	100	0.001	0.006	0.000	0.033
			250	0.002	0.002	0.000	0.040
			500	0.003	0.006	0.003	0.046
Power							
0.05	0.6	0.84	100	0.593	0.905	0.255	0.426
			250	0.945	0.996	0.710	0.690
			500	0.998	1.000	0.969	0.889
	0.7	0.84	100	0.801	0.942	0.499	0.462
			250	0.990	0.999	0.949	0.772
			500	1.000	1.000	0.997	0.945

Table 2. (Continued)

0.1	0.8	0.84	100	0.883	0.971	0.693	0.558
			250	0.998	1.000	0.998	0.868
			500	1.000	1.000	1.000	1.000
	0.9	1.04	100	0.944	0.980	0.870	0.822
			250	1.000	1.000	0.998	0.992
			500	1.000	1.000	1.000	1.000
	0.6	0.84	100	0.916	0.988	0.713	0.836
			250	1.000	1.000	0.991	0.984
			500	1.000	1.000	1.000	1.000
	0.7	0.84	100	0.991	0.997	0.908	0.890
			250	1.000	1.000	1.000	0.995
			500	1.000	1.000	1.000	1.000
	0.8	0.84	100	0.994	1.000	0.975	0.970
			250	1.000	1.000	1.000	1.000
			500	1.000	1.000	1.000	1.000
Mis-specified model (logistic model)							
θ	ϕ	$n_D = n_{\bar{D}}$	EB	PB	Wald.CV	Fuzzy p -value	
Type I error rate							
0	0.6	100	0.031	0.139	0.012	0.072	
		250	0.016	0.146	0.010	0.047	
		500	0.016	0.165	0.005	0.046	
	0.7	100	0.011	0.027	0.006	0.055	
		250	0.007	0.029	0.005	0.043	
		500	0.008	0.034	0.001	0.045	
	0.8	100	0.005	0.002	0.003	0.026	
		250	0.005	0.007	0.004	0.023	
		500	0.003	0.002	0.009	0.027	
	0.9	100	0.005	0.001	0.003	0.030	
		250	0.002	0.000	0.002	0.042	
		500	0.002	0.000	0.002	0.044	
Power							
0.1	0.6	100	0.873	0.999	0.558	0.805	
		250	0.996	1.000	0.884	0.967	
		500	1.000	1.000	0.995	0.996	
	0.7	100	0.928	0.999	0.631	0.862	
		250	0.995	1.000	0.927	0.981	
		500	1.000	1.000	0.997	1.000	
	0.8	100	0.977	1.000	0.798	0.932	
		250	0.999	1.000	0.942	0.989	
		500	1.000	1.000	1.000	1.000	

Moreover, to investigate the impact of increased sample size, we generate a larger dataset by randomly drawing 200 cases and 200 controls with replacement from the original data and apply our proposed methods. With the larger sample sizes, the one-sided CIs of incremental value based on EB and PB have lower limits 0.062 and 0.049, respectively, and the one-sided Wald.cv test has p -value $< 10^{-5}$, providing strong evidence that adding sLeX yields significantly better performance compared with using sLeA alone.

Table 3. Coverage of 95% two-sided confidence interval (CI) and corresponding length in the parenthesis using Wald with cross-validation method, under different underlying models and scenarios. ϕ indicates the performance of biomarker X alone and θ indicates the incremental value

Coverage (length) of two-sided 95% CI					
Correctly specified model (logic model)					
θ	ϕ	c_1	100	$n_D = n_{\bar{D}}$ 250	500
0	0.6	0	98.3% (0.117)	99.0% (0.053)	97.8% (0.051)
		0.84	98.5% (0.123)	98.8% (0.060)	98.2% (0.034)
	0.7	0	98.0% (0.056)	96.8% (0.019)	95.9% (0.008)
		0.84	97.3% (0.071)	97.9% (0.030)	97.8% (0.015)
	0.8	0.67	95.5% (0.036)	95.1% (0.014)	95.2% (0.007)
		0.84	96.0% (0.040)	95.2% (0.016)	95.6% (0.008)
	0.9	1.04	92.2% (0.023)	93.1% (0.009)	93.9% (0.005)
		1.34	95.1% (0.026)	94.2% (0.011)	93.5% (0.006)
0.05	0.6	0.84	94.3% (0.154)	96.0% (0.089)	96.5% (0.056)
	0.7	0.84	93.8% (0.108)	96.4% (0.060)	95.5% (0.039)
0.1	0.8	0.84	92.2% (0.083)	94.7% (0.049)	94.9% (0.033)
	0.9	1.04	91.7% (0.069)	93.4% (0.042)	94.0% (0.029)
	0.6	0.84	95.4% (0.175)	95.7% (0.100)	96.9% (0.065)
	0.7	0.84	95.3% (0.128)	95.6% (0.074)	95.7% (0.049)
	0.8	0.84	94.5% (0.105)	94.5% (0.063)	95.3% (0.043)
Mis-specified model (logistic model)					
θ	ϕ		100	$n_D = n_{\bar{D}}$ 250	500
0	0.6		98.3% (0.121)	99.0% (0.062)	98.4% (0.036)
	0.7		98.1% (0.074)	97.5% (0.032)	98.3% (0.017)
	0.8		96.2% (0.044)	96.0% (0.019)	96.0% (0.010)
	0.9		93.5% (0.025)	94.6% (0.011)	93.8% (0.005)
0.1	0.6		93.4% (0.212)	93.3% (0.137)	95.0% (0.097)
0.1	0.7		91.9% (0.189)	92.9% (0.120)	94.9% (0.084)
0.1	0.8		93.0% (0.156)	94.0% (0.099)	95.1% (0.069)

The fuzzy p -value approach rejects the null hypothesis with probability 1. The Wald.CV two-sided CI for incremental value of sLeX equals [0.066, 0.164].

5. CONCLUDING REMARKS

In this paper, we considered an inference problem about the incremental value of a new biomarker when combined with an established biomarker using an OR rule, motivated by the example in early detection of pancreatic cancer, where the standard biomarker CA19-9 is only elevated in a subclass of cancer cases. Thus, identifying a new biomarker that is present in the other subclasses to combine with CA19-9 is of primary interest. We considered a nonparametric estimator of incremental value of the new biomarker, based on an estimator of the OR rule that maximizes the weighted average of sensitivity and specificity. We proposed different procedures based on bootstrap, cross-validation, and a novel fuzzy p -value approach, to test and make inference about a new biomarker's incremental value. Through extensive numerical studies,

we found that the hypothesis test based on one-sided empirical bootstrap CI has satisfactory performance in terms of well-controlled Type I error rate and decent power for declaring the usefulness of the new marker, while the popular percentile bootstrap CI should be avoided due to its inflated Type I error rate. When it is of interest to provide uncertainty about the estimated incremental value, we found that two-sided Wald-type CI based on cross-validated estimates of incremental value performs very well, with coverage close to the nominal level. The novel fuzzy p -value method we proposed for testing the incremental value also has satisfactory performance. Moreover, the fuzzy p -value method can be particularly appealing as a testing procedure given its theoretical foundation and its potential to be extended to other biomarker testing problems when non-regularity is an issue. Importantly, our findings are based not only on settings where the true risk model conditional on biomarkers follows a logic model with an OR combination. They are also based on settings where the logic risk model does not hold, but the OR rule is used as a practical way to combine biomarkers for simplicity and interpretability. Our findings provide valuable guidance on selecting appropriate methods for testing and making inference about the incremental value of a new biomarker. Such threshold-based decision rules are of interest not only in the specific classification problems we considered here, but also in other problem settings such as the identification of optimal dynamic treatment regimens (Laber and Zhao, 2015; Wang and Rudin, 2015; Zhang and others, 2015), where the decision is to predict optimal treatment allocation instead of disease/non-disease status.

We used grid search to identify the threshold parameters. Though not a focus of this work, a computationally efficient implementation could find $\hat{c}$ in $O(n \log n)$ time by first sorting X , and subsequently using cumulative sums to iteratively compute the c -specific weighted combination of sensitivity and specificity. An analogous $O(n^2 \log n)$ time procedure could be used to derive $\hat{c}_1$ and $\hat{c}_2$ by fixing a candidate c_2 , again sorting X , and iteratively computing the c_2 -specific sum at different values of c_1 . Repeating this over all n candidate values for $\hat{c}_2$ will find the maximizer.

Furthermore, it is worth mentioning that while our current work focuses on the estimation and inference of the incremental value of a new single biomarker, the framework could be in general extended to make inference about the incremental value raised by multiple new biomarkers. Let X be an established biomarker for predicting the disease outcome D and let $Y_1, \dots, Y_K$ be the K new biomarkers where $K \geq 1$ is an integer. When combining the new biomarkers $Y_1, \dots, Y_K$ with the established biomarker X , we suppose a case is declared if either one of $X, Y_1, \dots, Y_K$ is elevated. The sensitivity and specificity using the OR rule combining $Y_1, \dots, Y_K$ with X are defined as $Sen(c_1, c_2, \dots, c_{K+1}) = P(X > c_1 \text{ OR } Y_1 > c_2 \text{ OR } \dots \text{ OR } Y_K > c_{K+1} \mid D = 1)$ and $Spe(c_1, c_2, \dots, c_{K+1}) = P(X \leq c_1 \text{ AND } Y_1 \leq c_2 \text{ AND } \dots \text{ AND } Y_K \leq c_{K+1} \mid D = 0)$, respectively, for some thresholds $c_1, \dots, c_{K+1}$. Notice that the problem considered in this article is the case with $K = 1$. Our current framework under $K = 1$ can be naturally applied to the scenarios with multiple new biomarkers, i.e., $K \geq 2$. More generally the method could extend to include both OR/AND combinations. When the number of markers to be combined is increased, one might consider other algorithms such as LOGIC regression or tree-based algorithms, which are more computationally efficient than the grid search method in finding the maximizers $\hat{c}_1, \dots, \hat{c}_{K+1}$ of the weighted average of empirical estimates of sensitivity and specificity.

6. SOFTWARE

The R code is available at <https://github.com/WangLu88/ORrules.git>.

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

This work was supported by the U.S. National Institutes of Health grant R01 GM106177-01.

REFERENCES

- BCAKER, S. G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* **56**, 1082–1087.
- BICKEL, P. J., KLAASSEN, C. A., BICKEL, P. J., RITOV, Y., KLAASSEN, J., WELLNER, J. A. and RITOV, Y. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*, Volume 2. New York: Springer New York.
- EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. Boca Raton: CRC press.
- ETZIONI, R., KOOPERBERG, C., PEPE, M., SMITH, R. and GANN, P. H. (2003). Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics* **4**, 523–538.
- FENG, Z. (2010). Classification versus association models: should the same methods apply? *Scandinavian Journal of Clinical and Laboratory Investigation* **70**, 53–58.
- GANN, P. H., MA, J., CATALONA, W. J. and STAMPFER, M. J. (2002). Strategies combining total and percent free prostate specific antigen for detecting prostate cancer: a prospective evaluation. *The Journal of Urology* **167**, 2427–2434.
- GEYER, C. J. and MEEDEN, G. D. (2005). Fuzzy and randomized confidence intervals and p-values. *Statistical Science* **20**, 358–366.
- HAN, J., PEI, J. and KAMBER, M. (2011). *Data Mining: Concepts and Techniques*. Burlington: Elsevier.
- LABER, E. B. and ZHAO, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* **102**, 501–514.
- MCINTOSH, M. W. and PEPE, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58**, 657–664.
- RUCZINSKI, I., KOOPERBERG, C. and LEBLANC, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics* **12**, 475–511.
- TANG, H., SINGH, S., PARTYKA, K., KLETTER, D., HSUEH, P., YADAV, J., ENSINK, E., BERN, M., HOSTETTER, G., HARTMAN, D. and others. (2015). Glycan motif profiling reveals plasma sialyl-lewis X elevations in pancreatic cancers that are negative for sialyl-lewis A. *Molecular & Cellular Proteomics* **14**, 1323–1333.
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics*, Volume 3. Cambridge: Cambridge University Press.
- WANG, F. and RUDIN, C. (2015). Falling rule lists. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, PMLR **38**, 1013–1022.
- YOUDEM, W. J. (1950). Index for rating diagnostic tests. *Cancer* **3**, 32–35.
- ZHANG, Y., LABER, E. B., TSIATIS, A. and DAVIDIAN, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71**, 895–904.

[Received May 01, 2018; revised October 08, 2018; accepted for publication October 17, 2018]