Statistica Sinica **33** (2023), 193-214
doi:https://doi.org/10.5705/ss.202021.0020

INTERVAL ESTIMATION FOR OPERATING CHARACTERISTIC OF CONTINUOUS BIOMARKERS WITH CONTROLLED SENSITIVITY OR SPECIFICITY

Yijian Huang[1], Isaac Parakati[2], Dattatraya H. Patil[1] and Martin G. Sanda[1]

[1]*Emory University and*
[2]*Ann & Robert H. Lurie Children's Hospital of Chicago*

Abstract: The receiver operating characteristic (ROC) curve provides a comprehensive performance assessment of a continuous biomarker over the full threshold spectrum. Nevertheless, a medical test often dictates operating at a certain high level of sensitivity or specificity. A diagnostic accuracy metric directly targeting clinical utility is specificity at the controlled sensitivity level, or vice versa. While the empirical point estimation is readily adopted in practice, the nonparametric interval estimation is difficult because the variance involves density functions, owing to the estimated threshold. In addition, even with a fixed threshold, many standard confidence intervals for the binomial proportion, including the Wald interval, can exhibit erratic behaviors. This study is motivated by the superior performance of the score interval for binomial proportion, and we propose a novel extension for the biomarker problem. We also develop an exact bootstrap procedure and establish the consistency of the bootstrap variance estimator. Both single-biomarker evaluation and two-biomarker comparison are investigated. Extensive simulation studies demonstrated competitive performance of our proposals. An application to aggressive prostate cancer diagnosis is also provided.

Key words and phrases: Diagnostic test, exact bootstrap, score confidence interval, sensitivity at controlled specificity, specificity at controlled sensitivity.

1. Introduction

Fueled by rapid recent advances in the scientific knowledge of molecular biology and high-throughput omics technologies, a large number of candidate biomarkers are being identified for disease diagnosis and prognosis, and the prediction of response to specific therapeutic interventions. Biomarker evaluation and comparison has become especially important for their validation and further clinical translation to ultimately improve and advance clinical practice (e.g., Tzoulaki, Siontis and Ioannidis (2011); Ioannidis and Panagiotou (2011)). Many

Corresponding author: Yijian Huang, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA. E-mail: yhuang5@emory.edu.

biomarkers are continuous, which means their dichotomization at a threshold is necessary for binary clinical testing. Sensitivity and specificity vary with the threshold, giving rise to the receiver operating characteristic (ROC) curve. While the ROC curve fully characterizes the performance of a biomarker over the complete threshold spectrum, only the point where the test is intended to operate is clinically relevant. For example, with aggressive prostate cancer diagnosis, a positive non-invasive test would be confirmed by biopsy. As a result, the cost of a false negative greatly outweighs that of a false positive. In this circumstance, the non-invasive test needs to attain a high sensitivity, say 95% (e.g., Catalona et al. (1998); Sanda et al. (2017)), to be clinically useful. Therefore, specificity at the controlled sensitivity level is a more sensible accuracy metric than, say, the area under the ROC curve, which is popular in practice. Of course, sensitivity at a controlled specificity level could be more relevant in a different clinical context. Nevertheless, the two correspond to the same statistical problem, upon transposing the roles of cases and controls. We focus on the former throughout this article.

While the empirical estimator of specificity at a controlled sensitivity level is straightforward to obtain, the nonparametric interval estimation is complicated by the fact that the variance involves density functions of the biomarker for case and control populations (cf. Linnet (1987); Pepe (2003)). Pepe (2003) suggested using kernel smoothing for the density estimation. However, this approach can be sensitive to the bandwidth choice, and choosing an appropriate bandwidth is often challenging with practical sample sizes. Furthermore, the approach is not invariant to a monotone transformation of the biomarker. As an alternative, Platt, Hanley and Yang (2000) and Zhou and Qin (2005) proposed adopting resampling bootstrap procedures. Nevertheless, the uncertainty from resampling affects reproducibility, despite that the error can be made small by increasing the resampling size.

If the threshold is fixed, the problem reduces to the interval estimation for binomial proportion. Nevertheless, even with this basic problem, many standard confidence intervals exhibit erratic behaviors (Agresti and Coull (1998); Brown, Cai and DasGupta (2001, 2002)). In particular, the simple and widely used Wald interval tends to have considerable under-coverage. Here, the score interval (Wilson (1927)) is recognized for its superior performance. Agresti and Coull (1998) suggested an adjusted Wald interval, mimicking the score interval to achieve better coverage performance. These results have influenced the interval estimation for our problem, that is, with an estimated threshold. Zhou and Qin (2005) incorporated the Agresti–Coull adjustment in their proposals, although the justification is not clear in this new context.

In this article, we propose a novel extension of the score interval for binomial proportion to specificity at a controlled sensitivity level. As another contribution, we develop an exact bootstrap procedure and establish the consistency of the bootstrap variance estimator. In Section 2, we evaluate a single biomarker. In Section 3, we compare two biomarkers, under both unpaired and paired designs (cf. Pepe (2003)). A bias analysis of the empirical specificity at controlled sensitivity is provided in Section 4, leading to an alternative point estimate and, subsequently, the associated confidence intervals. Simulations are reported in Section 5, and an illustration is given in Section 6 with prostate cancer detection. Final remarks are provided in Section 7. Technical details, including proofs, are relegated to the Appendix. An `R` package that implements the proposed methods is publicly available at the first author's website `http://web1.sph.emory.edu/users/yhuang5`.

2. Proposed Method for Single-Biomarker Evaluation

Consider a biomarker of interest M. Denote the case and control variables by $M_\bullet$ and $M_\circ$, respectively. Write their distribution functions as $F_\bullet(t) \equiv \Pr(M_\bullet \leq t)$ and $F_\circ(t) \equiv \Pr(M_\circ \leq t)$, respectively, and the quantile function of the former as $F_\bullet^{-1}(p) \equiv \inf\{t : F_\bullet(t) \geq p\}$. The case sample consists of $n_\bullet$ independent replicates of $M_\bullet$: $M_{\bullet i}$, $i = 1, \ldots, n_\bullet$, whereas the control sample comprises $n_\circ$ independent replicates of $M_\circ$: $M_{\circ i}$, $i = 1, \ldots, n_\circ$. Adopt the convention that reaching or exceeding a given threshold results in a positive diagnosis. With $\rho_0 \in (0, 1)$ as the controlled level of sensitivity, the largest threshold is $\tau_0 = F_\bullet^{-1}\{(1 - \rho_0)+\}$ such that the sensitivity defined as $\Pr(M_\bullet \geq \tau_0)$ is at least ρ_0. Accordingly, the specificity is $\phi_0 \equiv F_\circ(\tau_0-)$. Their natural plug-in estimators are given by

$$\widehat{\tau} = \widehat{F}_\bullet^{-1}\{(1 - \rho_0)+\}, \qquad \widehat{\phi} = \widehat{F}_\circ(\widehat{\tau}-),$$

where $\widehat{F}_\bullet$ and $\widehat{F}_\circ$ are the empirical versions of $F_\bullet$ and $F_\circ$, respectively. Under regularity conditions, $\widehat{\phi}$ is asymptotically normal with mean ϕ_0 and variance

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 \equiv \left\{\frac{F_\circ'(\tau_0)}{F_\bullet'(\tau_0)}\right\}^2 \frac{\rho_0(1 - \rho_0)}{n_\bullet} + \frac{\phi_0(1 - \phi_0)}{n_\circ}, \tag{2.1}$$

where $F_\bullet'$ and $F_\circ'$ are the derivatives of $F_\bullet$ and $F_\circ$, respectively; see Greenhouse and Mantel (1950), Hsieh and Turnbull (1996), and Pepe (2003), among others, and also Theorem 1, presented later. The variance has two components: σ_1^2, resulting from the threshold estimation, and σ_2^2, from the empirical specificity with given threshold τ_0. As discussed in Section 1, the involvement of density functions complicates the variance estimation.

2.1. Exact bootstrap

Bootstrapping is an effective approach to variance estimation. Platt, Hanley and Yang (2000) and Zhou and Qin (2005) suggested the routine resampling implementation. To improve reproducibility, we develop an exact bootstrap procedure and show its feasibility for this problem.

We focus first on the threshold estimation with the cases. Maritz and Jarrett (1978) and Efron (1979) derived the exact bootstrap distribution for a sample order statistic. From their result, we obtain a resampling scheme that is equivalent to the bootstrap resampling with respect to a sample quantile. Denote the ceiling function by $\lceil \cdot \rceil$.

Lemma 1. *Consider an independent and identically distributed sample of a random variable with size n. Denote the empirical cumulative distribution function by $\widehat{F}$ and its bootstrap counterpart by F^*. For any $p \in (0, 1)$, conditional on the observed data, $F^{*-1}(p)$ has the same distribution as $\widehat{F}^{-1}(B)$, where independent random variable B follows* $\mathrm{Beta}(\lceil np \rceil, n - \lceil np \rceil + 1)$.

Remark 1. This lemma was deduced from Maritz and Jarrett (1978) and Efron (1979). However, it might become more intuitive in light of a representation of the distribution of $\widehat{F}^{-1}(p)$. That is, $\widehat{F}^{-1}(p)$ as an order statistic can be shown to have the same distribution as $F^{-1}(B)$, where F is the underlying cumulative distribution function under consideration.

This result does not impose any restriction on the underlying distribution, which can be continuous, discrete, or a mixture of the two. Write τ^* as the bootstrap counterpart of $\widehat{\tau}$. Given that $\widehat{\tau}$ is the $\{(1 - \rho_0)+\}$-quantile, the bootstrap distribution of τ^* is the same as $\widehat{F}_{\bullet}^{-1}(1 - B_{\bullet})$, where $B_{\bullet} \sim \mathrm{Beta}(n_{\bullet} - r + 1, r)$ and $r \equiv \lceil n_{\bullet}(1 - \rho_0)+ \rceil$. That is, the bootstrap distribution assigns to order statistics $M_{\bullet[i]}$, $i = 1, \ldots, n_{\bullet}$, with the same probabilities as $1 - B_{\bullet}$ to the $n_{\bullet}$ intervals evenly split between zero and one. This resampling equivalence facilitates efficient computation with the exact bootstrap. Moreover, this novel perspective of τ^* can be exploited in a large-sample study, as shown later.

Now, we turn to the specificity estimation with the controls, at a given threshold τ^*. Let ϕ^* be the bootstrap counterpart of $\widehat{\phi}$. Write $\Pr^*$ as the bootstrap probability, that is, conditional on the observed data. The conditional bootstrap probability mass function is

$$\Pr^*(\phi^* = \phi \mid \tau^*) = \binom{n_\circ}{n_\circ\phi} \widehat{F}_\circ(\tau^*-)^{n_\circ\phi}\{1 - \widehat{F}_\circ(\tau^*-)\}^{n_\circ(1-\phi)}, \tag{2.2}$$

for $\phi \in \{0, 1/n_\circ, \ldots, (n_\circ - 1)/n_\circ, 1\}$. Upon rescaling by $n_\circ$, this is a binomial distribution with size $n_\circ$ and success probability $\widehat{F}_\circ(\tau^*-)$.

Because the case and control samples are independent of each other, combining the preceding results gives the bootstrap distribution of the specificity at controlled sensitivity ρ_0:

$$\Pr^*(\phi^* = \phi) = \sum_{i=1}^{n_\bullet} \Pr^*(\tau^* = M_{\bullet[i]})\Pr^*(\phi^* = \phi \mid \tau^* = M_{\bullet[i]}), \tag{2.3}$$

for $\phi \in \{0, 1/n_\circ, \ldots, (n_\circ - 1)/n_\circ, 1\}$. This exact bootstrap distribution is feasible to compute, although care is needed to avoid numerical underflow and overflow.

Write E_* and Var_* as the conditional expectation and variance, respectively, given the observed data. The bootstrap variance estimator of $\widehat{\phi}$ is $\mathrm{Var}_*(\phi^*)$:

$$\begin{aligned}\widehat{\sigma}^2 &= \mathrm{Var}_*\{\widehat{F}_\circ(\tau^*-)\} + E_*[n_\circ^{-1}\widehat{F}_\circ(\tau^*-)\{1 - \widehat{F}_\circ(\tau^*-)\}] \\ &\equiv \widehat{\sigma}_1^2 + \widehat{\sigma}_2^2, \end{aligned} \tag{2.4}$$

which are estimators of σ_1^2 and σ_2^2, respectively, as components of σ^2 given in (2.1).

The consistency and asymptotic normality of $\widehat{\phi}$ have long been known; see, for example, Greenhouse and Mantel (1950). However, theoretical justification may not have been provided even for the consistency of the bootstrap distribution, and much less for that of the bootstrap variance estimator; in general, the former does not necessarily imply the latter (e.g., Ghosh et al. (1984); Shao (1990)). The following result focuses on the bootstrap distribution and variance. Nevertheless, the asymptotic properties of $\widehat{\phi}$ are also stated, mostly for completeness, with weaker assumptions imposed. We also provide a proof, from which the consistency of the bootstrap distribution immediately follows.

Theorem 1. *Suppose that the following conditions hold: (i) the size ratio of the cases and controls $n_\bullet/n_\circ$ converges to a nonzero finite constant as $n_\bullet + n_\circ$ approaches ∞; (ii) $\rho_0 \in (0, 1)$; and (iii) $F_\bullet$ and $F_\circ$ are differentiable at the threshold τ_0, with $F'_\bullet(\tau_0) > 0$. Then, $\widehat{\phi}$ is strongly consistent for ϕ_0, and asymptotically normal with mean ϕ_0 and variance σ^2, as given in (2.1). At the same time, $n_\circ^{1/2}(\phi^* - \widehat{\phi})$ conditional on the data converges in distribution to the same limit as $n_\circ^{1/2}(\widehat{\phi} - \phi_0)$. Furthermore, under the additional condition (iv) $F_\bullet$ and $F_\circ$ are continuously differentiable in a neighborhood around τ_0, $n_\bullet\widehat{\sigma}_1^2$, $n_\circ\widehat{\sigma}_2^2$, and subsequently $n_\circ\widehat{\sigma}^2$ converge in probability to $n_\bullet\sigma_1^2$, $n_\circ\sigma_2^2$, and $n_\circ\sigma^2$, respectively.*

2.2. Confidence intervals

Using the bootstrap variance estimator $\widehat{\sigma}^2$, a Wald $100(1-\alpha)\%$ confidence interval is given by $\widehat{\phi} \pm z_{\alpha/2}\widehat{\sigma}$, where $z_{\alpha/2}$ is the $(\alpha/2)$-quantile of the standard normal distribution. The interval is truncated with $[0,1]$ to respect the parameter range. Another common and simple interval is the percentile $100(1-\alpha)\%$ confidence interval, which is the interval between the $\alpha/2$- and $(1-\alpha/2)$-quantiles of the bootstrap distribution.

We propose a novel confidence interval. For binomial proportion, Agresti and Coull (1998) and Brown, Cai and DasGupta (2001, 2002), among others, showed that the score interval (Wilson (1927)) has good coverage accuracy, even for a very small sample size. It outperforms many other competitors, including the Wald interval and the "exact" interval of Clopper and Pearson (1934). Like the Wald interval, the score interval is inverted from a hypothesis test. However, the score interval adopts the null variance. Unfortunately, the estimated specificity at controlled sensitivity is the proportion of an overdispersed binomial, owing to the estimated threshold. Therefore, its variance is not fully determined by the null specificity. We overcome this issue by estimating the overdispersion factor using $\widehat{\sigma}^2\widehat{\sigma}_2^{-2}$. The resulting score interval is given by

$$\left\{\phi:\ \frac{(\widehat{\phi}-\phi)^2}{n_\circ^{-1}\phi(1-\phi)\widehat{\sigma}^2\widehat{\sigma}_2^{-2}} < z_{\alpha/2}^2\right\}, \tag{2.5}$$

which has an explicit expression, with the two bounds as solutions to a quadratic equation. Because the denominator approaches zero as ϕ goes to zero or one, this interval is guaranteed to be contained in the parameter range $[0,1]$.

All three confidence intervals are invariant to a monotone transformation of the biomarker. Owing to the exact bootstrap, they are also perfectly reproducible.

3. Two-Biomarker Comparison

Another common task in biomarker research is to compare two biomarkers, say X and Y. Denote the quantities in Section 2 associated with each biomarker by adding a subscript "X" or "Y." At the common controlled sensitivity ρ_0, the specificity difference $\delta_0 \equiv \phi_{0X} - \phi_{0Y}$ between the two provides a meaningful measure of their clinical utility difference. Using the estimated thresholds $\widehat{\tau}_X$ and $\widehat{\tau}_Y$, we obtain the corresponding estimated specificities $\widehat{\phi}_X$ and $\widehat{\phi}_Y$ and, subsequently, their estimated difference $\widehat{\delta} = \widehat{\phi}_X - \widehat{\phi}_Y$. This point estimation procedure remains the same for the two biomarkers measured in two independent studies or in the same one, that is, under an unpaired or a paired design, respectively (cf. Pepe (2003)). However, the inference is different, and also more complicated.

3.1. Unpaired comparison

With two biomarkers measured in independent studies, the bootstrap distributions of $\widehat{\phi}_X$ and $\widehat{\phi}_Y$ are independent of each other. Then, the bootstrap variance estimator and the distribution of the difference in specificity $\widehat{\delta}$ can be easily obtained. Subsequently, the Wald and percentile confidence intervals for δ_0 are constructed in the same fashion as for the single-biomarker evaluation in Section 2.2.

Nevertheless, it is unclear how to construct a score interval. To follow the approach for the single-biomarker evaluation, the two null specificities at the controlled sensitivity would be needed. However, they are not determined, except for their difference δ. We suggest instead combining δ and the other biomarker's estimated specificity, that is, $\delta + \widehat{\phi}_Y$ and $\widehat{\phi}_X - \delta$, as estimated specificities under the null for biomarkers X and Y, respectively. To this end, we propose the following score confidence interval for δ_0:

$$\Bigg[\delta : (\widehat{\delta} - \delta)^2 < z^2_{\alpha/2}\Bigg\{\frac{\widehat{\sigma}^2_X}{n_{\circ X}\widehat{\sigma}^2_{2X}}(\delta + \widehat{\phi}_Y)(1 - \delta - \widehat{\phi}_Y) + \frac{\widehat{\sigma}^2_Y}{n_{\circ Y}\widehat{\sigma}^2_{2Y}}(\widehat{\phi}_X - \delta)(1 - \widehat{\phi}_X + \delta)\Bigg\}\Bigg]. \tag{3.1}$$

In contrast to the more standard form as in (2.5), the variance component above is moved to the other side of the inequality. Because $\delta + \widehat{\phi}_Y$ and $\widehat{\phi}_X - \delta$ are not guaranteed to be bounded between zero and one, the variance component may not necessarily be positive. The current form is more sensible, because a negative variance component is against, rather than for, the null. Just like (2.5) for the single-biomarker evaluation, the confidence interval (3.1) has an explicit expression, with the two bounds being the solutions to a quadratic equation; the existence of the solutions can be easily shown. The interval is truncated by $[-1, 1]$ to respect the parameter range.

3.2. Paired comparison

Denote the pair of biomarkers by $(X_\bullet, Y_\bullet)^\top$ for a case and $(X_\circ, Y_\circ)^\top$ for a control. The case sample consists of $n_\bullet$ independent replicates, $(X_{\bullet i}, Y_{\bullet i})^\top$, for $i = 1, \ldots, n_\bullet$, and the control sample comprises $n_\circ$ independent replicates, $(X_{\circ i}, Y_{\circ i})^\top$, for $i = 1, \ldots, n_\circ$.

To derive the exact bootstrap distribution, we start with the cases. For $s,t=0,1$, introduce

$$\widehat{m}_{\bullet st}(x,y)=\#\{I(X_{\bullet i}\le x)=s, I(Y_{\bullet i}\le y)=t: i=1,\ldots,n_\bullet\}.$$

Write $\widehat{\mathbf{m}}_\bullet(x,y)=\{\widehat{m}_{\bullet 11}(x,y),\widehat{m}_{\bullet 10}(x,y),\widehat{m}_{\bullet 01}(x,y),\widehat{m}_{\bullet 00}(x,y)\}^\top$ and $\mathbf{m}^*_\bullet(x,y)=\{m^*_{\bullet 11}(x,y),m^*_{\bullet 10}(x,y),m^*_{\bullet 01}(x,y),m^*_{\bullet 00}(x,y)\}^\top$ as its bootstrap counterpart. Because $(\tau^*_X,\tau^*_Y)^\top$ may take a value only in $\Omega=\{X_{\bullet i}:i=1,\ldots,n_\bullet\}\times\{Y_{\bullet j}:j=1,\ldots,n_\bullet\}$, we have

$$\begin{aligned}\Pr{}^*(\tau^*_X\le X_{\bullet i},\tau^*_Y\le Y_{\bullet j})=\Pr{}^*\{&m^*_{\bullet 11}(X_{\bullet i},Y_{\bullet j})+m^*_{\bullet 10}(X_{\bullet i},Y_{\bullet j})\ge r,\\ &m^*_{\bullet 11}(X_{\bullet i},Y_{\bullet j})+m^*_{\bullet 01}(X_{\bullet i},Y_{\bullet j})\ge r\};\end{aligned}\tag{3.2}$$

recall $r\equiv\lceil n_\bullet(1-\rho_0)+\rceil$. The right-hand side above can be calculated from the fact that

$$\mathbf{m}^*_\bullet(X_{\bullet i},Y_{\bullet j})\mid\text{observed data}\sim\text{Multinomial}\left\{n_\bullet,\frac{\widehat{\mathbf{m}}_\bullet(X_{\bullet i},Y_{\bullet j})}{n_\bullet}\right\}.$$

Now, with the controls, we derive the bootstrap distribution of δ^*, the bootstrap counterparts of $\widehat{\delta}$, conditional on the thresholds τ^*_X and τ^*_Y. In parallel to their case counterparts, introduce

$$\widehat{m}_{\circ st}(x,y)=\#\{I(X_{\circ i}\le x)=s,\ I(Y_{\circ i}\le y)=t: i=1,\ldots,n_\circ\},$$

for $s,t=0,1$, and subsequently $\widehat{\mathbf{m}}_\circ(x,y)$ and $\mathbf{m}^*_\circ(x,y)$. It is clear that

$$\mathbf{m}^*_\circ(\tau^*_X,\tau^*_Y)\mid\text{observed data},\tau^*_X,\tau^*_Y\sim\text{Multinomial}\left\{n_\circ,\frac{\widehat{\mathbf{m}}_\circ(\tau^*_X,\tau^*_Y)}{n_\circ}\right\},$$

from which the bootstrap distribution of δ^* given τ^*_X and τ^*_Y can be obtained. For that purpose, $\delta^*=\{m^*_{\circ 10}(\tau^*_X,\tau^*_Y)-m^*_{\circ 01}(\tau^*_X,\tau^*_Y)\}/n_\circ$ as $\widehat{\delta}=\{\widehat{m}_{\circ 10}(\widehat{\tau}_X,\widehat{\tau}_Y)-\widehat{m}_{\circ 01}(\widehat{\tau}_X,\widehat{\tau}_Y)\}/n_\circ$.

Combining the preceding results gives the bootstrap distribution of δ^*. The bootstrap variance estimator of $\widehat{\delta}$ is given by

$$\begin{aligned}\widehat{\sigma}^2_\delta&=\text{Var}_*\{E_*(\delta^*\mid\tau^*_X,\tau^*_Y)\}+E_*\{\text{Var}_*(\delta^*\mid\tau^*_X,\tau^*_Y)\}\\&=\sum_{(x,y)^\top\in\Omega}\Big[\{n_\circ^{-1}\widehat{m}_{\circ 10}(x,y)-n_\circ^{-1}\widehat{m}_{\circ 01}(x,y)-E_*\delta^*\}^2\\&\quad+n_\circ^{-2}\widehat{m}_{\circ 10}(x,y)\{1-n_\circ^{-1}\widehat{m}_{\circ 10}(x,y)\}+n_\circ^{-2}\widehat{m}_{\circ 01}(x,y)\{1-n_\circ^{-1}\widehat{m}_{\circ 01}(x,y)\}\\&\quad+2n_\circ^{-3}\widehat{m}_{\circ 10}(x,y)\widehat{m}_{\circ 01}(x,y)\Big]\Pr{}^*(\tau^*_X=x,\tau^*_Y=y),\end{aligned}\tag{3.3}$$

where $E_*\delta^* = \sum_{(x,y)^\top \in \Omega} n_\circ^{-1}\{\widehat{m}_{\circ 10}(x,y) - \widehat{m}_{\circ 01}(x,y)\}\mathrm{Pr}^*(\tau_X^* = x, \tau_Y^* = y)$. This computation is obviously more intensive than that for the single-biomarker evaluation, but is still feasible.

Theorem 2. *Suppose that at least one of the two correlations, between $I(X_\bullet \le \tau_{0X})$ and $I(Y_\bullet \le \tau_{0Y})$ and between $I(X_\circ \le \tau_{0X})$ and $I(Y_\circ \le \tau_{0Y})$, is less than one. Under conditions (i), (ii), and (iii) in Theorem 1, as for each biomarker, conditional on the data, $n_\circ^{1/2}(\delta^* - \widehat{\delta})$ converges in distribution to the same limit as $n_\circ^{1/2}(\widehat{\delta} - \delta_0)$, which is normal with mean zero. If condition (iv) in Theorem* 1 *holds additionally for each biomarker, $n_\circ \widehat{\sigma}_\delta^2$ converges in probability to the asymptotic variance of $n_\circ^{1/2}(\widehat{\delta} - \delta_0)$.*

Asymptotic degeneracy of the joint distribution of $\widehat{\phi}_X$ and $\widehat{\phi}_Y$ is avoided with the above correlation condition.

Using the bootstrap variance estimator and bootstrap distribution of $\widehat{\delta}$, the Wald and percentile confidence intervals can be constructed for δ_0. For the score confidence interval, we build upon that for the unpaired comparison, as given in (3.1), by further accounting for the difference between $\widehat{\sigma}_\delta^2$ and $\widehat{\sigma}_X^2 + \widehat{\sigma}_Y^2$:

$$\Bigg[\delta : (\widehat{\delta} - \delta)^2 < \frac{z_{\alpha/2}^2 \widehat{\sigma}_\delta^2}{n_\circ(\widehat{\sigma}_X^2 + \widehat{\sigma}_Y^2)} \Bigg\{ \frac{\widehat{\sigma}_X^2}{\widehat{\sigma}_{2X}^2}(\delta + \widehat{\phi}_Y)(1 - \delta - \widehat{\phi}_Y) + \frac{\widehat{\sigma}_Y^2}{\widehat{\sigma}_{2Y}^2}(\widehat{\phi}_X - \delta)(1 - \widehat{\phi}_X + \delta) \Bigg\} \Bigg]. \tag{3.4}$$

Like the previous ones, this score confidence interval has an explicit expression, with the two bounds as solutions to a quadratic equation. We truncate the interval by $[-1, 1]$ to respect the parameter range.

4. Alternative Point Estimator for Improved Performance

Here, we focus on the single-biomarker evaluation, as in Section 2. The estimated threshold is the $\lceil n_\bullet(1 - \rho_0) + \rceil$th order statistic of the cases. Because of the discrete nature, intuitively, the bias of $\widehat{\phi}$ would have an oscillating component with variable $n_\bullet$. A more formal analysis, given in the Appendix, shows that

$$E(\widehat{\phi}) - \phi_0 = \frac{F_\circ'(\tau_0)}{F_\bullet'(\tau_0)} \left(\frac{\lceil n_\bullet(1 - \rho_0) + \rceil}{n_\bullet + 1} - 1 + \rho_0 \right) + \left\{ \frac{F_\circ''(\tau_0)}{F_\bullet'(\tau_0)^2} - \frac{F_\circ'(\tau_0)F_\bullet''(\tau_0)}{F_\bullet'(\tau_0)^3} \right\} \frac{\rho_0(1 - \rho_0)}{2n_\bullet} + o(n_\bullet^{-1}), \tag{4.1}$$

provided $F_\bullet$ is continuous and strictly increasing, and the second derivatives, $F''_\bullet$ and $F''_\circ$, exist and are continuous in a neighborhood of τ_0. This bias result is sharper than that of Lloyd and Yong (1999, Thm. 2). While both bias terms are of order $n_\bullet^{-1}$, the first one is the oscillating component. For example, with fixed $\rho_0 = 0.95$, the first term vanishes whenever $0.05(n_\bullet + 1)$ is an integer, which occurs at increments of 20.

This bias analysis suggests an alternative point estimator free of this oscillating bias component. Write $\lfloor\cdot\rfloor$ as the floor function. Consider two threshold estimates as the $\lfloor(n_\bullet+1)(1-\rho_0)\rfloor$th and the $\lceil(n_\bullet+1)(1-\rho_0)\rceil$th order statistics of the case biomarkers. If the two order statistics are the same, they lead to the estimated specificity $\widehat{\phi}$, because $\lfloor(n_\bullet+1)(1-\rho_0)\rfloor \le \lceil n_\bullet(1-\rho_0)+\rceil \le \lceil(n_\bullet+1)(1-\rho_0)\rceil$. Otherwise, obtain the weighted average of the empirical specificities at these two thresholds as $\widetilde{\phi}$, with weights $\lceil(n_\bullet + 1)(1 - \rho_0)\rceil - (n_\bullet + 1)(1 - \rho_0)$ and $(n_\bullet + 1)(1 - \rho_0) - \lfloor(n_\bullet + 1)(1 - \rho_0)\rfloor$, respectively. It is straightforward to show

$$E(\widetilde{\phi}) - \phi_0 = \left\{ \frac{F''_\circ(\tau_0)}{F'_\bullet(\tau_0)^2} - \frac{F'_\circ(\tau_0)F''_\bullet(\tau_0)}{F'_\bullet(\tau_0)^3} \right\} \frac{\rho_0(1-\rho_0)}{2n_\bullet} + o(n_\bullet^{-1}). \tag{4.2}$$

The two estimators, $\widehat{\phi}$ and $\widetilde{\phi}$, are asymptotically equivalent to each other to the first order, and coincide when $(n_\bullet + 1)(1 - \rho_0)$ is an integer. This construction is reminiscent of the usual definition of the sample median, which is the average of the two middle-order statistics in the case of an even sample size. Indeed, it can be applied to the threshold estimation instead. However, we do not do so because the resulting specificity estimator would no longer be invariant to a monotone transformation of the biomarker.

By replacing $\widehat{\phi}$ with $\widetilde{\phi}$, we obtain alternatives to the Wald and score confidence intervals in Section 2.2. The variance components are kept the same, although the exact bootstrap for $\widetilde{\phi}$ may be developed. This same approach also leads to new Wald and score intervals for the two-biomarker comparison discussed in Section 3.

Note that $\widetilde{\phi}$ still shares the same non-oscillating bias component of order $n_\bullet^{-1}$ with $\widehat{\phi}$. It is possible to further develop an estimator that is unbiased to order $n_\bullet^{-1}$ using, for example, a delete-d jackknife. However, this bias reduction may not reduce mean squared error. In fact, preliminary numerical studies did not show a performance improvement in the resulting confidence intervals. Therefore, we did not pursue this further.

5. Simulations

Extensive simulations were conducted to evaluate the proposed methods under practical sample sizes. Throughout, the controlled sensitivity level was set to 95%, and the nominal level of the confidence intervals was set to 95%. Under each setup, 1,000 replications were simulated. The setups with equal case and control sizes are reported here. The others are included in the Supplementary Material; the results were largely similar.

5.1. Single-biomarker evaluation

For comparison, we included several existing confidence intervals. In one, we employed kernel smoothing for the density estimation to construct a Wald interval. To preserve the parameter range, this interval for logit-transformed specificity was first formulated and then back-transformed, as described in Pepe (2003). We adopted the univariate adaptive kernel density estimation of Silverman (1986), as implemented in function `ajk()` of `R` package `Quantreg` with default tuning parameters. Unlike other confidence intervals being studied, this kernel-smoothing approach is not invariant to a monotone transformation of the biomarker. Therefore, we also applied the same method to the data after an exponential transformation of the biomarker, because biomarkers are often nonnegative and have skewed distributions in practice. Among confidence intervals based on the resampling bootstrap, Zhou and Qin (2005) reported that their "BTII" interval performed best. This Zhou–Qin interval was implemented with two resampling sizes, 200 and 500. A size of 200 is usually considered sufficient for Wald confidence intervals (cf. Efron and Tibshirani (1994, Sec. 6.4)), although a size of 500 was adopted in the simulations of Zhou and Qin (2005). In addition, we constructed this interval based on our exact bootstrap, equivalent to the case of an infinite resampling size.

Both the case and the control biomarkers followed normal distributions with unit variance but different means, to achieve a specified specificity at the controlled 95% sensitivity. In the simulations, the original Zhou–Qin interval showed considerable variability from the resampling. Nevertheless, this variability was not reflected in the coverage probability and averaged length. Thus, only the exact bootstrap version is included in the reporting; the resampling variability is shown in Section 6. Figure 1 shows the coverage probabilities and averaged lengths of these confidence intervals over a grid of ϕ_0 between 0.1 and 0.9, with mesh size 0.005, in the setting of the case and control sample sizes $(n_\bullet, n_\circ)$ being (50,50). The kernel smoothing method could be sensitive to the scale on which a

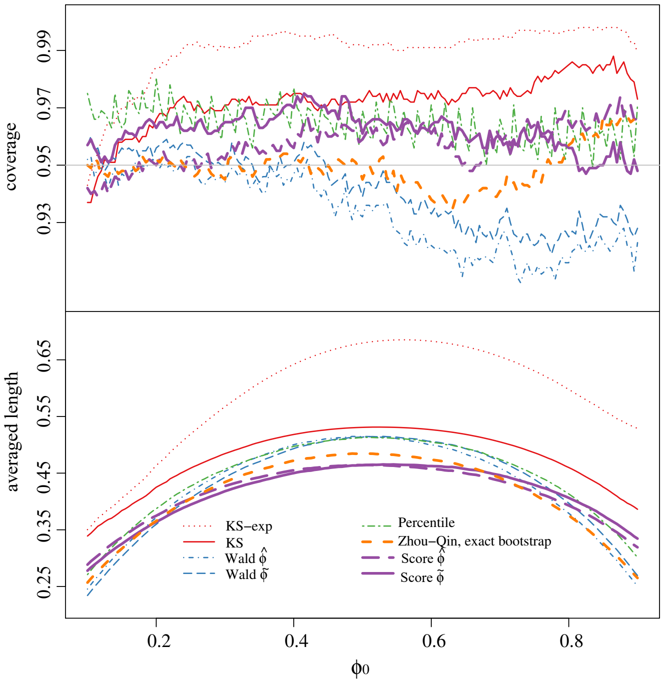

Figure 1. Simulation summaries of 95% confidence intervals for specificity at controlled 95% sensitivity in the single-biomarker evaluation, under fixed sizes $n_\bullet = n_\circ = 50$ and variable ϕ_0. KS is the kernel smoothing-based Wald confidence intervals, as in Pepe (2003), whereas KS-exp corresponds to the KS applied to exponentially transformed biomarker data.

biomarker is measured. With both of the scales considered, these intervals tended to be much wider and more conservative than the others. The exact bootstrap-based Wald and percentile intervals had similar averaged length, shorter than the kernel smoothing ones, but longer than the Zhou–Qin and score intervals. However, the Wald intervals had considerable under-coverage, except at small ϕ_0, whereas the percentile interval had coverage always above the nominal level. Overall, the Zhou–Qin and score intervals performed best, reaching the nominal coverage level with the shortest averaged length over most of ϕ_0. Between them, the score intervals tended to have a shorter length in the middle value range of ϕ_0, whereas the Zhou–Qin interval was tighter at the extremes.

These confidence intervals were also evaluated with fixed $\phi_0 = 0.2$, 0.4, 0.6, and 0.8, and variable sample size $n_\bullet = n_\circ$ from 20 to 200. The three best performers, the two score intervals using $\widehat{\phi}$ and $\widetilde{\phi}$ and the Zhou–Qin interval, are reported in Figure 2. Except for the case of $\phi_0 = 0.8$, the Zhou–Qin interval had an oscillating pattern in the coverage probability, apparently due to the bias of

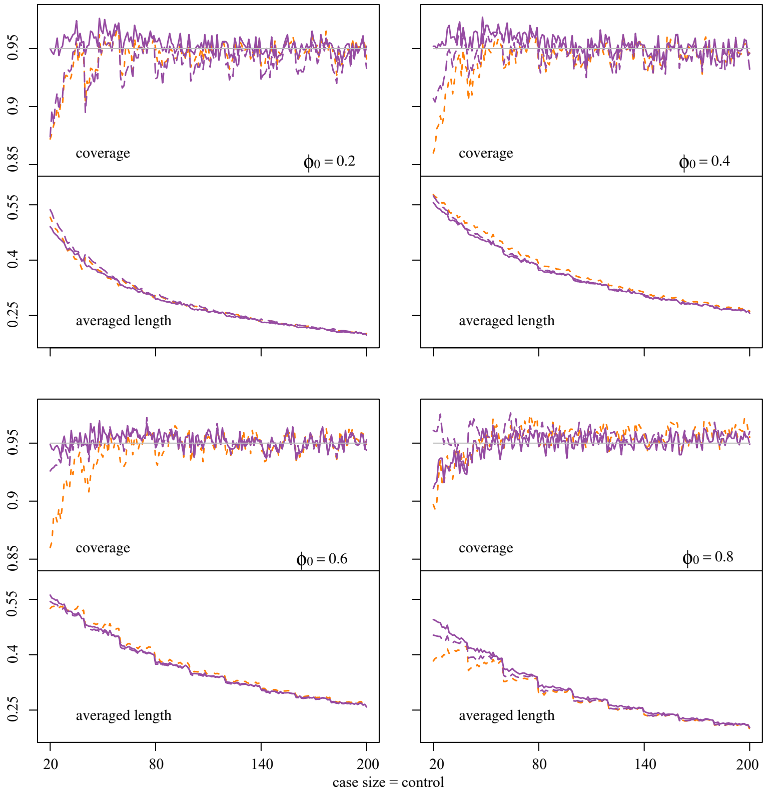

Figure 2. Simulation summaries of 95% confidence intervals for specificity at controlled 95% sensitivity in the single-biomarker evaluation, under fixed ϕ_0 and variable sizes $n_\bullet = n_\circ$. Score intervals, based on $\widehat{\phi}$ and $\widetilde{\phi}$, and the exact bootstrap-based Zhou–Qin are included, with the same labeling as in Figure 1.

the adopted point estimate $\widehat{\phi}$, as discussed in Section 4, and considerable undercoverage could arise at certain sample sizes. Not surprisingly, this occurred to the score interval using $\widehat{\phi}$ as well. In contrast, the coverage probability of the score interval using $\widetilde{\phi}$ was much more stable, with little oscillation.

Many of these confidence intervals are closely related to those for binomial proportion (Agresti and Coull (1998); Brown, Cai and DasGupta (2001)). Nevertheless, the behavior patterns appeared to be different, at least when the case and control sizes were comparable. Oscillation in coverage could arise, but mainly because of the way that the threshold is estimated.

5.2. Two-biomarker comparison

We report unpaired comparison studies with $n_{\bullet X} = n_{\circ X} = n_{\bullet Y} = n_{\circ Y}$ and paired comparison studies with $n_\bullet = n_\circ$; the results with other sample size setups were similar. Case and control sizes of 50 and 200 were considered. With the unpaired comparison, each biomarker was simulated in the same fashion as in Section 5.1. Under the paired comparison, $(X_\circ, Y_\circ)^\top$ followed the standard bivariate normal distribution with a 0.5 correlation coefficient, and $(X_\bullet, Y_\bullet)^\top$ had a location shift from that distribution to attain the specified specificity at controlled 95% sensitivity for each biomarker.

Table 1 shows the coverage probabilities and averaged lengths of the exact bootstrap-based Wald, percentile, and score confidence intervals for difference in specificity at controlled 95% sensitivity. All confidence intervals were reasonably close to the nominal level, but they all tended to be conservative when the sample size was smaller. The two score intervals, without and with the oscillating bias-correction for the point estimate, were similar, both being considerably shorter than the other three, for both unpaired and paired comparisons alike.

The kernel smoothing-based Wald interval can be extended to the two-biomarker comparison. However, it was not included in our study in light of its less competitive performance in the single-biomarker evaluation. On the other hand, the Zhou–Qin interval may be extended as well. We studied its exact bootstrap version, although the results are not included in the table. The Zhou–Qin interval also tended to be conservative. It was slightly shorter than the Wald and percentile intervals, but much wider than the score ones.

6. Illustration with Aggressive Prostate Cancer Detection

This development was motivated by prostate cancer research to evaluate biomarkers for the detection of aggressive prostate cancer, that is, a Gleason score ≥ 7, among men undergoing their first-time biopsy. Two biomarkers of interest are serum prostate health index (phi) and urine PCA3. A total of 512 participants enrolled from four urology groups affiliated with three academic medical centers, consisting of 155 cases and 357 controls, per pathology testing on prostate biopsies (Sanda et al. (2017)). They provided post-urinary specimens after digital rectal examination and serum specimens, both before biopsy, and had their phi and PCA3 assayed. The metric of specificity at 95% sensitivity was adopted to evaluate the biomarker performance.

Figure 3 shows the analysis results. To reach 95% sensitivity, the estimated phi and PCA3 thresholds were 22.4 and 7.6, respectively. Their corresponding empirical specificities were 24.6% and 17.4%, which became 24.1% and 16.1% upon the oscillating bias correction. For each biomarker, various 95% confidence intervals for the specificity were constructed. The original Zhou–Qin interval, via

Table 1. Simulation summary statistics of exact bootstrap-based 95% confidence intervals for difference in specificity at controlled 95% sensitivity in the two-biomarker comparison.

				unpaired biomarkers					paired biomarkers				
				Wald		Pct	Score		Wald		Pct	Score	
ϕ_{0X}	δ_0	size		$\widehat{\phi}$	$\widetilde{\phi}$		$\widehat{\phi}$	$\widetilde{\phi}$	$\widehat{\phi}$	$\widetilde{\phi}$		$\widehat{\phi}$	$\widetilde{\phi}$
0.2	0.0	50	C	966	968	987	976	979	979	978	991	982	986
			L	579	579	576	467	445	516	516	515	430	409
		200	C	950	951	975	956	955	959	956	979	963	958
			L	293	293	292	275	268	261	261	261	248	242
0.4	0.2	50	C	956	963	985	970	970	964	974	986	970	965
			L	671	671	663	540	525	600	600	595	500	486
		200	C	939	942	966	942	944	948	949	972	955	953
			L	339	339	337	318	313	303	303	302	288	284
	0.0	50	C	964	958	981	968	962	968	974	993	971	978
			L	749	749	744	594	584	658	658	657	545	535
		200	C	938	933	963	941	935	946	940	972	951	945
			L	380	380	378	355	352	337	337	336	319	317
0.6	0.4	50	C	955	956	984	946	947	962	968	989	955	958
			L	671	674	668	556	551	607	608	605	517	512
		200	C	955	953	971	959	942	964	955	980	961	957
			L	339	339	337	321	319	306	306	304	293	291
	0.2	50	C	957	959	981	957	955	964	970	991	963	969
			L	752	753	751	601	600	663	663	664	552	552
		200	C	944	941	968	949	943	961	959	979	966	957
			L	381	381	379	356	356	339	339	337	321	321
	0.0	50	C	962	960	980	959	956	970	964	993	966	962
			L	760	760	765	599	609	669	669	675	551	560
		200	C	956	949	970	954	949	964	959	983	966	959
			L	380	380	379	355	358	339	339	338	321	324
0.8	0.6	50	C	960	959	980	927	928	967	969	979	927	932
			L	582	586	592	519	522	537	539	544	486	488
		200	C	953	950	969	946	944	960	956	971	949	944
			L	290	290	288	281	281	264	264	263	257	258
	0.4	50	C	957	962	973	936	939	968	969	978	936	942
			L	679	679	685	562	569	608	608	612	519	526
		200	C	956	949	966	947	948	961	957	979	960	954
			L	337	337	335	319	321	305	305	304	291	293
	0.2	50	C	957	963	979	944	958	965	966	985	953	951
			L	690	690	701	552	569	614	614	622	510	526
		200	C	961	960	967	956	947	961	962	979	963	967
			L	337	337	336	316	321	304	304	303	288	293
	0.0	50	C	970	961	989	957	946	971	971	990	970	969
			L	619	619	639	492	519	551	551	569	454	480
		200	C	965	956	975	962	952	974	967	983	967	963
			L	289	289	289	272	279	258	258	259	245	252

size $= n_{\bullet X} = n_{\circ X} = n_{\bullet Y} = n_{\circ Y}$ for unpaired, and size $= n_{\bullet} = n_{\circ}$ for paired comparison.

C: coverage probability (×1000); L: average interval length (×1000). Pct: percentile interval.

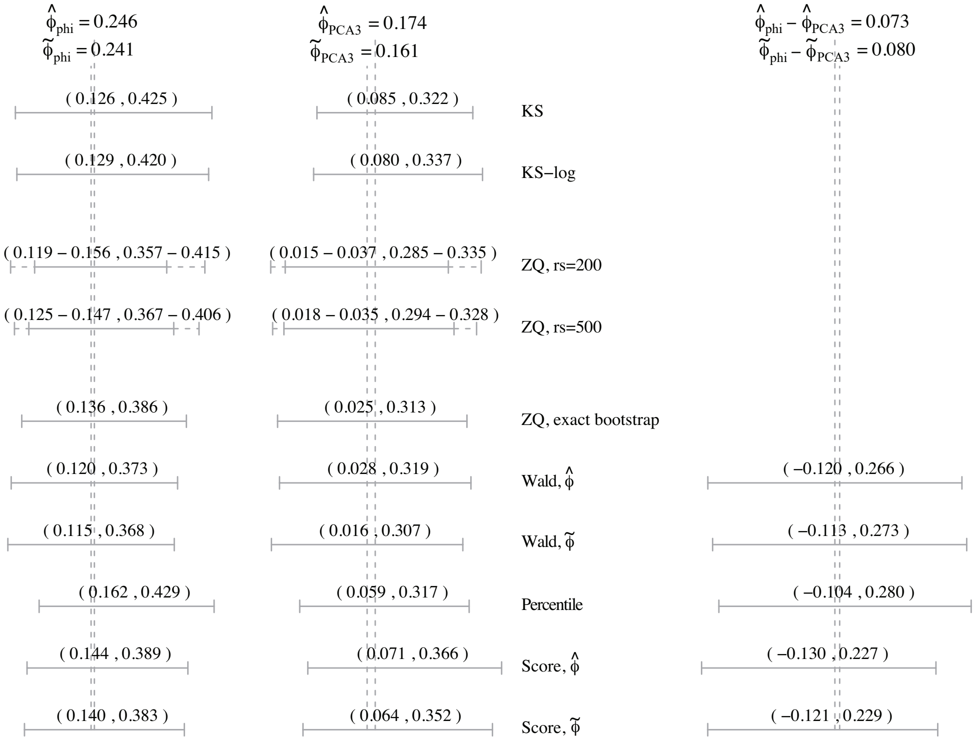

Figure 3. Analysis results of the prostate cancer study: point estimates and 95% confidence intervals of specificity at 95% sensitivity for phi and PCA3 as diagnostic biomarkers for aggressive prostate cancer. KS and KS-log are the kernel smoothing-based Wald confidence intervals, as applied to untransformed and logarithm-transformed biomarker data, respectively. For the resampling Zhou–Qin (ZQ) confidence interval with a given resampling size (rs), ranges of the left and right bounds over 100 runs are provided.

the resampling bootstrap, exhibited considerable variability under resampling sizes of both 200 and 500. Other intervals also showed appreciable differences. A paired comparison was made between phi and PCA3, and 95% confidence intervals were constructed for their difference in specificity at controlled 95% sensitivity. The score intervals were tighter than the Wald and percentile ones, which is consistent with the simulation results.

We also estimated the areas under the ROC curves. They were 0.792 (95% confidence interval: 0.750 – 0.835) and 0.696 (95% confidence interval: 0.647 – 0.744) for phi and PCA3, respectively. The difference was statistically significant, with a p-value of 0.003. Although the area under the ROC curve is a commonly used accuracy metric, its interpretation is different. As noted in Section 1, specificity at controlled 95% sensitivity is clinically more sensible in this application.

7. Discussion

We have investigated interval estimation for specificity at a controlled sensitivity level. Exact bootstrap is advocated over kernel smoothing and resampling bootstrap for the inference. Furthermore, we have proposed novel score confidence intervals, which showed competitive or superior performance in comparison with existing ones in the single-biomarker evaluation and the two-biomarker comparison.

We have limited our scope to confidence intervals on the basis of the empirical specificity at controlled sensitivity or its variants, for their robustness. However, there is an extensive body of literature on kernel-based estimators of the ROC curve, including Zou, Hall and Shapiro (1997) and Lloyd (1998). Lloyd and Yong (1999) showed that such estimators have a smaller mean squared error asymptotically than that of the empirical estimator, given a proper choice of the smoothing bandwidth. Hall, Hyndman and Fan (2004) investigated the confidence intervals based on the kernel-smoothed ROC curve and kernel-based variance estimation. Theoretically, this approach could lead to an accuracy gain. Nevertheless, an appropriate selection of a multitude of smoothing parameters is required, and further development would be needed for wide practical adoption.

Supplementary Material

Additional simulation results, with unequal case and control sizes, are available in the Supplementary Material. They are for both the single-biomarker evaluation and the two-biomarker comparison.

Acknowledgments

The authors thank the reviewers for their helpful comments and suggestions. Support by grants CA230268, CA113913, and AI050409 from the US National Institutes of Health is gratefully acknowledged.

Appendix: Proofs and Other Technical Details

Proof of Lemma 1

Since $nF^*(x) \sim \text{Binomial}\{n, \widehat{F}(x)\}$ given the data,

$$\begin{aligned}\Pr{}^*\{F^*(x) \geq p\} &= \Pr{}^*\{nF^*(x) \geq \lceil np \rceil\} \\ &= \sum_{k=\lceil np \rceil}^{n} \binom{n}{k} \widehat{F}(x)^k \{1 - \widehat{F}(x)\}^{n-k}\end{aligned}$$

$$= \lceil np \rceil \binom{n}{\lceil np \rceil} \int_0^{\widehat{F}(x)} y^{\lceil np \rceil - 1}(1-y)^{n-\lceil np \rceil}\, dy$$
$$= \Pr{}^*\{B \le \widehat{F}(x)\},$$

where the third equality follows by induction from n downward as the value of $\lceil np \rceil$. By a basic result of quantile function (e.g., Serfling (1980, Lemma 1.1.4)), $F^*(x) \ge p$ if and only if $F^{*-1}(p) \le x$, and $B \le \widehat{F}(x)$ if and only if $\widehat{F}^{-1}(B) \le x$. The assertion then follows.

Proof of Theorem 1

Write $D[a,b]$ as the space of cadlag functions in $[a,b]$ with some a and b such that $a < \tau_0 < b$. Endow such a space with the supremum norm and their product with the max supremum norm. In light of $\widehat{\phi} = \widehat{F}_\circ[\widehat{F}_\bullet^{-1}\{(1-\rho_0)+\}-]$, $\widehat{\phi}$ is the plug-in estimator in the map $\{F_\bullet, F_\circ\} \mapsto \phi_0$ decomposed as

$$\left.\begin{array}{l} F_\bullet \in D[a,b] \mapsto \tau_0 \in R \\ \\ F_\circ \in D[a,b] \end{array}\right\} \mapsto \phi_0 \in R.$$

With $F_\bullet$ being differentiable with positive derivative at τ_0, $F_\bullet \mapsto \tau_0$ is Hadamard-differentiable at $F_\bullet$ tangentially to the set of functions $h \in D[a,b]$ that are continuous at τ_0, with derivative $-h(\tau_0)/F_\bullet'(\tau_0)$ (e.g., van der Vaart (1998, Lemma 21.3)). Meanwhile, given $F_\circ$ being differentiable at τ_0, it can be shown that $(\tau_0, F_\circ) \mapsto \phi_0$ is Hadamard-differentiable at $(\tau_0, F_\circ)$ tangentially to the set $\{k : \in R\} \times \{l : \in D[a,b]$, continuous at $\tau_0\}$, with derivative $F_\circ'(\tau_0)k + l(\tau_0)$. By the chain rule, $\{F_\bullet, F_\circ\} \mapsto \phi_0$ is then Hadamard-differentiable at $\{F_\bullet, F_\circ\}$ tangentially to the set $\{h : \in D[a,b]$, continuous at $\tau_0\} \times \{l : \in D[a,b]$, continuous at $\tau_0\}$, with derivative $-h(\tau_0)F_\circ'(\tau_0)/F_\bullet'(\tau_0) + l(\tau_0)$. As the Hadamard-differentiability implies continuity, $\widehat{\phi}$ is strongly consistent for ϕ_0 following the strong consistency of $\{\widehat{F}_\bullet, \widehat{F}_\circ\}$ for $\{F_\bullet, F_\circ\}$ by the Glivenko–Cantelli theorem. With the weak convergence of $n_\bullet^{1/2}(\widehat{F}_\bullet - F_\bullet)$ and $n_\circ^{1/2}(\widehat{F}_\circ - F_\circ)$, $\widehat{\phi}$ is $AN(\phi_0, \sigma^2)$ following the functional delta method and the asymptotic normality of $\{\widehat{F}_\bullet(\tau_0), \widehat{F}_\circ(\tau_0)\}$ by the central limit theorem. Furthermore, by the results on empirical bootstrap and delta method for bootstrap (van der Vaart (1998, Thm. 23.7 and 23.9)), the conditional distribution of $n_\circ^{1/2}(\phi^* - \widehat{\phi})$ given the observed data converges in distribution to the same limit as $n_\circ^{1/2}(\widehat{\phi} - \phi_0)$.

To investigate the bootstrap variance estimator, $\widehat{F}_\circ(\tau^*-)$ is an important building block as shown in (2.4). Write $\phi(\rho) = F_\circ[F_\bullet^{-1}\{(1-\rho)+\}-]$ and $\widehat{\phi}(\rho) = \widehat{F}_\circ[\widehat{F}_\bullet^{-1}\{(1-\rho)+\}-]$; of course, $\phi_0 \equiv \phi(\rho_0)$ and $\widehat{\phi} \equiv \widehat{\phi}(\rho_0)$. By Lemma 1,

$\widehat{F}_\circ(\tau^*-)$ is equivalent to $\widehat{\phi}(B_\bullet)$ in conditional distribution. As a fact, $B_\bullet$ has the same distribution as $C_1/(C_1+C_2)$ with $C_1 \sim \chi^2(2n_\bullet - 2r + 2)$, $C_2 \sim \chi^2(2r)$, and $C_1 \perp\!\!\!\perp C_2$. Therefore, $B_\bullet$ converges almost surely to ρ_0 by strong law of large numbers and continuous mapping theorem, and $n_\bullet^{1/2}(B_\bullet - \rho_0)$ converges in distribution to $N\{0, \rho_0(1-\rho_0)\}$ by the central limit theorem and delta method. Furthermore, we give a bound on the tail probability of $B_\bullet$ on the basis of the sub-Gaussianity of the Beta distribution (Marchal and Arbel (2017)). Since both $B_\bullet$ and $1 - B_\bullet$ are $\{4(n_\bullet + 2)\}^{-1}$ sub-Gaussian,

$$\Pr\{|B_\bullet - E(B_\bullet)| > b\} \le 2\exp\{-2(n_\bullet + 2)b^2\} \tag{A.1}$$

for any constant b.

The asymptotic normality result on $\widehat{\phi}$ can be extended to the weak convergence of $\widehat{\phi}(\rho)$ in a neighborhood of ρ around ρ_0 under the additional condition (iv). The arguments are essentially the same, upon appropriate modifications of domain and range spaces of the functions involved and with extended result on Hadamard differentiability of the quantile function (van der Vaart (1998, Lemma 21.4)). Given that $n_\circ^{1/2}\{\widehat{\phi}(\rho) - \phi(\rho)\}$ converges weakly to a Gaussian process in a neighborhood of ρ around ρ_0, for any $d_{n_\circ} \downarrow 0$, one can show

$$\sup_{|d| \le d_{n_\circ}} |\widehat{\phi}(\rho_0 + d) - \widehat{\phi}(\rho_0) - \phi(\rho_0 + d) + \phi(\rho_0)| = o_p(n_\circ^{-1/2});$$

see, for example, Huang (2017, Appendix). Meanwhile, the differentiability of $\phi(\rho)$ at ρ_0 implies

$$\sup_{|d| \le d_{n_\circ}} d^{-1}|\phi(\rho_0 + d) - \phi(\rho_0) - \phi'(\rho_0)d| = o(1),$$

where $\phi'(\rho_0) = -F'_\circ(\tau_0)/F'_\bullet(\tau_0)$.

Let $c_{n_\bullet} = \{\log n_\bullet/(n_\bullet + 2)\}^{1/2}$. Following (A.1), $\Pr\{|B_\bullet - E(B_\bullet)| > c_{n_\bullet}\} \le 2n_\bullet^{-2}$. Given $E(B_\bullet) - \rho_0 = O(n_\bullet^{-1})$, $|B_\bullet - \rho_0| = |B_\bullet - E(B_\bullet)| + O(n_\bullet^{-1})$. Note that E_* takes expectation over $B_\bullet$. Then,

$$\begin{aligned}
E_*\{\widehat{\phi}(B_\bullet)\} &= E_*[\widehat{\phi}(B_\bullet)I\{|B_\bullet - E(B_\bullet)| \le c_{n_\bullet}\}] + O(n_\bullet^{-2}) \\
&= \widehat{\phi}(\rho_0) + \phi'(\rho_0)E[(B_\bullet - \rho_0)I\{|B_\bullet - E(B_\bullet)| \le c_{n_\bullet}\}] \\
&\quad + o_p\{n_\bullet^{-1/2} + E|B_\bullet - \rho_0|\} \\
&= \widehat{\phi} + o_p(n_\bullet^{-1/2}),
\end{aligned}$$

since $E|B_\bullet - E(B_\bullet)| = O(n_\bullet^{-1/2})$ following $\mathrm{Var}(B_\bullet) = O(n_\bullet^{-1})$ by Jensen's inequality. Similarly, $E_*\{\widehat{\phi}(B_\bullet)^2\} = \widehat{\phi}^2 + o_p(n_\bullet^{-1/2})$. Therefore, $n_\circ\widehat{\sigma}_2^2 = E_*[\widehat{\phi}(B_\bullet)\{1 - \widehat{\phi}(B_\bullet)\}]$ converges to $n_\circ\sigma_2^2$ in probability. By similar arguments,

$$\begin{aligned}\widehat{\sigma}_1^2 &= E_*[\{\widehat{\phi}(B_\bullet) - \widehat{\phi} + o_p(n_\bullet^{-1/2})\}^2]\\ &= E[\{\phi'(\rho_0)(B_\bullet - \rho_0)\}^2 I\{|B_\bullet - E(B_\bullet)| \le c_{n_\bullet}\}] + o(n_\bullet^{-1})\\ &= \sigma_1^2 + o(n_\bullet^{-1}).\end{aligned}$$

Then, $n_\bullet\widehat{\sigma}_1^2$ converges to $n_\bullet\sigma_1^2$ in probability. Thus, $n_\circ\widehat{\sigma}^2$ converges to $n_\circ\sigma^2$ in probability as well.

Proof of Theorem 2

The arguments for the proof of Theorem 1 with a single biomarker extend in a straightforward fashion to the two-biomarker problem, for the estimation with correlated specificities at a common controlled sensitivity level. Subsequently, the claims on the difference in specificity follow.

Bias analysis of $\widehat{\phi}$ in Section 4

Following Remark 1, $\widehat{\tau}$ has the same distribution as $F_\bullet^{-1}(1 - B_\bullet)$. Thus, $E(\widehat{\phi}) = E\{E(\widehat{\phi} \mid \widehat{\tau})\} = E\{F_\circ(\widehat{\tau})\} = E\{\phi(B_\bullet)\}$. In light of (A.1) and with $c_{n_\bullet} = \{\log n_\bullet/(n_\bullet + 2)\}^{1/2}$,

$$\begin{aligned}E\{\phi(B_\bullet)\} &= E\{\phi(B_\bullet)I(|B_\bullet - E(B_\bullet)| \le c_{n_\bullet})\} + O(n_\bullet^{-2})\\ &= \phi\{E(B_\bullet)\} + \phi'\{E(B_\bullet)\}E[\{B_\bullet - E(B_\bullet)\}I(|B_\bullet - E(B_\bullet)| \le c_{n_\bullet})]\\ &\quad + E\left\{\left[\frac{\phi''\{E(B_\bullet)\}}{2} + o(1)\right]\{B_\bullet - E(B_\bullet)\}^2 I(|B_\bullet - E(B_\bullet)| \le c_{n_\bullet})\right\}\\ &\quad + O(n_\bullet^{-2})\\ &= \phi\{E(B_\bullet)\} + \left\{\frac{\phi''(\rho_0)}{2} + o(1)\right\}E[\{B_\bullet - E(B_\bullet)\}^2] + O(n_\bullet^{-2})\\ &= \phi_0 + \phi'(\rho_0)\left(1 - \frac{r}{n_\bullet + 1} - \rho_0\right) + \phi''(\rho_0)\frac{\rho_0(1 - \rho_0)}{2n_\bullet} + o(n_\bullet^{-1}),\end{aligned}$$

which gives equation (4.1). The second equation above is an application of Taylor expansion, with the existence and continuity of $\phi''(\cdot)$ in a neighborhood of ρ_0 under the assumptions given.

References

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* **52**, 119–126.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation of a binomial proportion. *Statistical Science* **16**, 101–133.

Brown, L. D., Cai, T. T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics* **30**, 160–201.

Catalona, W. J., Partin, A. W., Slawin, K. M., Brawer, M. K., Flanigan, R. C., Patel, A. et al. (1998). Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: A prospective multicenter clinical trial. *JAMA* **279**, 1542–1547.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics* **7**, 1–26.

Efron B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap.* Chapman & Hall/CRC, New York.

Ghosh, M., Parr, W. C., Singh, K. and Babu, G. J. (1984). A note on bootstrapping the sample median. *The Annals of Statistics* **12**, 1130–1135.

Greenhouse, S. W. and Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics* **6**, 399–412.

Hall, P., Hyndman, R. J. and Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika* **91**, 743–750.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics* **24**, 25–40.

Huang, Y. (2017). Restoration of monotonicity respecting in dynamic regression. *Journal of the American Statistical Association* **112**, 613–622.

Ioannidis, J. P. and Panagiotou, O. A. (2011). Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA* **305**, 2200–2210.

Linnet, K. (1987). Comparison of quantitative diagnostic tests: Type I error, power and sample size. *Statistics in Medicine* **6**, 147–158.

Lloyd, C. J. (1998). The use of smoothed ROC curves to summarise and compare diagnostic systems. *Journal of the American Statistical Association* **93**, 1356–1364.

Lloyd, C. J. and Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statistics & Probability Letters* **44**, 221–228.

Marchal, O. and Arbel, J. (2017). On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability* **22**, 1–14.

Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association* **73**, 194–196.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, Oxford.

Platt, R. W., Hanley, J. A. and Yang, H. (2000). Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test. *Statistics in Medicine* **19**, 313–322.

Sanda, M. G., Feng, Z., Howard, D. H., Tomlins, S. A., Sokoll, L. J., Chan, D. W. et al. (2017). Association between combined TMPRSS2: ERG and PCA3 RNA urinary testing and detection of aggressive prostate cancer. *JAMA Oncology* **3**, 1085–1093.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

Shao, J. (1990). Bootstrap estimation of the asymptotic variances of statistical functionals. *Annals of the Institute of Statistical Mathematics* **42**, 737–752.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Tzoulaki, I., Siontis, K. C. and Ioannidis, J. P. (2011). Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: Meta-epidemiology study. *BMJ* **343**, d6829.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, New York.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212.

Zhou, X.-H. and Qin, G. (2005). Improved confidence intervals for the sensitivity at a fixed level of specificity of a continuous-scale diagnostic test. *Statistics in Medicine* **24**, 465–477.

Zou, K. H., Hall, W. J. and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* **16**, 2143–2156.

Yijian Huang

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA.

E-mail: yhuang5@emory.edu

Isaac Parakati

Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL 60611, USA.

E-mail: iparakati@luriechildrens.org

Dattatraya H. Patil

Department of Urology, Emory University, Atlanta, GA 30322, USA.

E-mail: dattatraya.patil@emory.edu

Martin G. Sanda

Department of Urology, Emory University, Atlanta, GA 30322, USA.

E-mail: martinsanda@emory.edu

(Received January 2021; accepted May 2021)