<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2021-0020</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
</tr>
</tbody>
</table>
Interval estimation for operating characteristic of continuous biomarkers with controlled sensitivity or specificity

Yijian Huang¹, Isaac Parakati², Dattatraya H. Patil¹, and Martin G. Sanda¹

¹Emory University
and ²Ann & 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 pro-
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 et al., 2011; Ioannidis and Panagiotou, 2011). Many 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 et al. (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 et al., 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 empiri-
cal 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 $M_*$ and $M_o$, respectively. Write their distribution functions as $F_*(t) \equiv \Pr(M_* \leq t)$ and $F_o(t) \equiv \Pr(M_o \leq t)$, respectively, and the quantile function of the former as $F_*^{-1}(p) \equiv \inf\{t : F_*(t) \geq p\}$. The case sample consists of $n_*$ independent replicates of $M_*$: $M_{*i}$, $i = 1, \cdots, n_*$, whereas the control sample comprises $n_o$ independent replicates of $M_o$: $M_{oi}$, $i = 1, \cdots, n_o$. 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 $\tau_0 = F_*^{-1}\{(1 - \rho_0)^+\}$ such that the sensitivity defined as $\Pr(M_* \geq \tau_0)$ is at least $\rho_0$. Accordingly, the
2.1 Exact bootstrap

Specificity is $\phi_0 \equiv F_o(\tau_0 -)$. Their natural plug-in estimators are given by

$$
\hat{\tau} = \hat{F}_\bullet^{-1}\{(1 - \rho_0)^+\}, \quad \hat{\phi} = \hat{F}_o(\hat{\tau} -),
$$

where $\hat{F}_\bullet$ and $\hat{F}_o$ are the empirical versions of $F_\bullet$ and $F_o$, respectively. Under regularity conditions, $\hat{\phi}$ is asymptotically normal with mean $\phi_0$ and variance

$$
\sigma^2 = \sigma_1^2 + \sigma_2^2 \equiv \left\{\frac{F'_o(\tau_0)}{F'_o(\tau_0)}\right\}^2 \frac{\rho_0 (1 - \rho_0)}{n_\bullet} + \frac{\phi_0 (1 - \phi_0)}{n_o}, \quad (2.1)
$$

where $F'_\bullet$ and $F'_o$ are the derivatives of $F_\bullet$ and $F_o$, 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: $\sigma_1^2$, resulting from the threshold estimation, and $\sigma_2^2$, from the empirical specificity with given threshold $\tau_0$. 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 et al. (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
2.1 Exact bootstrap

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 $\hat{F}$ and its bootstrap counterpart by $F^\ast$. For any $p \in (0, 1)$, conditional on the observed data, $F^\ast^{-1}(p)$ has the same distribution as $\hat{F}^{-1}(B)$, where independent random variable $B$ follows 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 $\hat{F}^{-1}(p)$. That is, $\hat{F}^{-1}(p)$ 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 $\tau^\ast$ as the bootstrap counterpart of $\hat{\tau}$. Given that $\hat{\tau}$ is the $\{(1 - \rho_0)^+\}$-quantile, the bootstrap distribution of $\tau^\ast$ is the same as $\hat{F}^{-1}(1 - B_{\bullet})$, where $B_{\bullet} \sim \text{Beta}(n_{\bullet} - r + 1, r)$ and $r \equiv \lceil n_{\bullet}(1 - \rho_0)^+ \rceil$. That is, the bootstrap distribution assigns to order statistics $M_{\bullet[i]}$, $i = 1, \cdots, n_{\bullet}$, with the same probabilities as $1 - B_{\bullet}$ to the $n_{\bullet}$ intervals evenly split between zero and
2.1 Exact bootstrap

one. This resampling equivalence facilitates efficient computation with the exact bootstrap. Moreover, this novel perspective of $\tau^*$ 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 $\tau^*$. Let $\phi^*$ be the bootstrap counterpart of $\hat{\phi}$. Write $Pr^*$ as the bootstrap probability, that is, conditional on the observed data. The conditional bootstrap probability mass function is

$$Pr^*(\phi^* = \phi \mid \tau^*) = \binom{n_o}{n_o \phi} \hat{F}_o(\tau^*-\hat{\phi})^{n_o \phi} (1 - \hat{F}_o(\tau^*)^{-})^{n_o (1-\hat{\phi})}, \quad (2.2)$$

for $\phi \in \{0, 1/n_o, \cdots, (n_o-1)/n_o, 1\}$. Upon rescaling by $n_o$, this is a binomial distribution with size $n_o$ and success probability $\hat{F}_o(\tau^*)^-$. 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 $\rho_0$:

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

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

Write $E_*$ and $\text{Var}_*$ as the conditional expectation and variance, respectively, given the observed data. The bootstrap variance estimator of $\hat{\phi}$ is
2.1 Exact bootstrap

\( \text{Var}_* (\phi^*) : \)

\[
\hat{\sigma}^2 = \text{Var}_* \{ \hat{F}_o (\tau^* -) \} + E_* [n_0^{-1} \hat{F}_o (\tau^* -) \{ 1 - \hat{F}_o (\tau^* -) \}]
\]

\[
\equiv \hat{\sigma}_1^2 + \hat{\sigma}_2^2,
\]

which are estimators of \( \sigma_1^2 \) and \( \sigma_2^2 \), respectively, as components of \( \hat{\sigma}^2 \) given in (2.1).

The consistency and asymptotic normality of \( \hat{\phi} \) 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 \( \hat{\phi} \) 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 \( n_/ n_o \) converges to a nonzero finite constant as \( n_/ n_o \) approaches \( \infty \); (ii) \( \rho_0 \in (0, 1) \); and (iii) \( F_/ \) and \( F_o \) are differentiable at the threshold \( \tau_0 \), with \( F_/ '(\tau_0) > 0 \). Then, \( \hat{\phi} \) is strongly consistent for \( \phi_0 \).
and asymptotically normal with mean $\phi_0$ and variance $\sigma^2$, as given in (2.1). At the same time, $n^{1/2}_o(\phi^* - \hat{\phi})$ conditional on the data converges in distribution to the same limit as $n^{1/2}_o(\hat{\phi} - \phi_0)$. Furthermore, under the additional condition (iv) $F_\tau$ and $F_\sigma$ are continuously differentiable in a neighborhood around $\tau_0$, $n\cdot\hat{\sigma}_1^2$, $n\cdot\hat{\sigma}_2^2$, and subsequently $n_1\cdot\hat{\sigma}_1^2$ converge in probability to $n\cdot\sigma_1^2$, $n_1\cdot\sigma_2^2$, and $n_1\cdot\sigma_2^2$, respectively.

### 2.2 Confidence intervals

Using the bootstrap variance estimator $\hat{\sigma}^2$, a Wald $100(1 - \alpha)$% confidence interval is given by $\hat{\phi} \pm z_{\alpha/2}\hat{\sigma}$, 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 et al. (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 $\hat{\sigma}^2 \hat{\sigma}^{-2}$. The resulting score interval is given by

$$\left\{ \phi : \frac{(\hat{\phi} - \phi)^2}{n^{-1}\phi(1-\phi)\hat{\sigma}^2\hat{\sigma}^{-2}} < z_{\alpha/2}^2 \right\}, \quad (2.5)$$

which has an explicit expression, with the two bounds as solutions to a quadratic equation. Because the denominator approaches zero as $\phi$ 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 $\rho_0$, 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 $\hat{\tau}_X$ and $\hat{\tau}_Y$, we obtain the corresponding estimated specificities $\hat{\phi}_X$ and $\hat{\phi}_Y$ and, subsequently, their estimated difference $\hat{\delta} = \hat{\phi}_X - \hat{\phi}_Y$.

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 $\hat{\phi}_X$ and $\hat{\phi}_Y$ are independent of each other. Then, the bootstrap variance estimator and the distribution of the difference in specificity $\hat{\delta}$ can be easily obtained. Subsequently, the Wald and percentile confidence intervals for $\delta_0$ 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 $\delta$. We suggest instead combining $\delta$ and the other biomarker’s estimated specificity, that is, $\delta + \hat{\phi}_Y$ and $\hat{\phi}_X - \delta$, as
3.2 Paired comparison

estimated specificities under the null for biomarkers $X$ and $Y$, respectively.

To this end, we propose the following score confidence interval for $\delta_0$:

$$\left[ \delta : (\hat{\delta} - \delta)^2 < \frac{z_{\alpha/2}^2}{n_{\circ X} \hat{\sigma}_X^2}(\delta + \hat{\phi}_Y)(1 - \delta - \hat{\phi}_Y) ight. 
\left. + \frac{\hat{\sigma}_Y^2}{n_{\circ Y} \hat{\sigma}_Y^2}(\hat{\phi}_X - \delta)(1 - \hat{\phi}_X + \delta) \right]. \quad (3.1)$$

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 $\delta + \hat{\phi}_Y$ and $\hat{\phi}_X - \delta$ 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 $(X_\circ, Y_\circ)^\top$ for a case and $(X_\circ, Y_\circ)^\top$ for a control. The case sample consists of $n_\circ$ independent replicates, $(X_{\circ i}, Y_{\circ i})^\top$, for $i = 1, \cdots, n_\circ$, and the control sample comprises $n_\circ$ independent replicates, $(X_{\circ i}, Y_{\circ i})^\top$, for $i = 1, \cdots, n_\circ$. 
3.2 Paired comparison

To derive the exact bootstrap distribution, we start with the cases. For

\( s, t = 0, 1 \), introduce

\[
\hat{m}_{st}(x, y) = \# \{ I(X_{oi} \leq x) = s, I(Y_{oi} \leq y) = t : i = 1, \cdots, n \}.
\]

Write \( \hat{m}_s(x, y) = (\hat{m}_{11}(x, y), \hat{m}_{10}(x, y), \hat{m}_{01}(x, y), \hat{m}_{00}(x, y))^\top \) and \( m^*_s(x, y) = (m^*_{11}(x, y), m^*_{10}(x, y), m^*_{01}(x, y), m^*_{00}(x, y))^\top \) as its bootstrap counterpart.

Because \( (\tau^*_X, \tau^*_Y)^\top \) may take a value only in \( \Omega = \{ X_i : i = 1, \cdots, n \} \times \{ Y_j : j = 1, \cdots, n \} \), we have

\[
\Pr^*(\tau^*_X \leq X_{oi}, \tau^*_Y \leq Y_{oj}) = \Pr^*\{ m^*_{11}(X_{oi}, Y_{oj}) + m^*_{10}(X_{oi}, Y_{oj}) \geq r, m^*_{11}(X_{oi}, Y_{oj}) + m^*_{01}(X_{oi}, Y_{oj}) \geq r \}; \quad (3.2)
\]

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

\[
m^*_s(X_{oi}, Y_{oj}) \mid \text{observed data} \sim \text{Multinomial}\{ n_s, \hat{m}_s(X_{oi}, Y_{oj})/n_s \}.
\]

Now, with the controls, we derive the bootstrap distribution of \( \delta^* \), the bootstrap counterparts of \( \hat{\delta} \), conditional on the thresholds \( \tau^*_X \) and \( \tau^*_Y \). In parallel to their case counterparts, introduce

\[
\hat{m}_{ost}(x, y) = \# \{ I(O_{oi} \leq x) = s, I(O_{oi} \leq y) = t : i = 1, \cdots, n \},
\]

for \( s, t = 0, 1 \), and subsequently \( \hat{m}_o(x, y) \) and \( m^*_o(x, y) \). It is clear that

\[
m^*_o(\tau^*_X, \tau^*_Y) \mid \text{observed data, } \tau^*_X, \tau^*_Y \sim \text{Multinomial}\{ n_o, \hat{m}_o(\tau^*_X, \tau^*_Y)/n_o \},
\]
from which the bootstrap distribution of $\delta^*$ given $\tau^*_X$ and $\tau^*_Y$ can be obtained. For that purpose, $\delta^* = \{m^*_{\circ10}(\tau^*_X, \tau^*_Y) - m^*_{\circ01}(\tau^*_X, \tau^*_Y)\}/n_\circ$ as $\hat{\delta} = \{\hat{m}_{\circ10}(\hat{\tau}_X, \hat{\tau}_Y) - \hat{m}_{\circ01}(\hat{\tau}_X, \hat{\tau}_Y)\}/n_\circ$.

Combining the preceding results gives the bootstrap distribution of $\delta^*$.

The bootstrap variance estimator of $\hat{\delta}$ is given by

$$
\hat{\sigma}^2 = \text{Var}_*\{E_*(\delta^* | \tau^*_X, \tau^*_Y)\} + E_*\{\text{Var}_*(\delta^* | \tau^*_X, \tau^*_Y)\}
= \sum_{(x,y) \in \Omega} \left[\{n^{-1}_\circ \hat{m}_{\circ10}(x,y) - m^{-1}_\circ \hat{m}_{\circ01}(x,y) - E_*\delta^*\}^2
+ n^{-2}_\circ \hat{m}_{\circ10}(x,y)\{1 - m^{-1}_\circ \hat{m}_{\circ10}(x,y)\}
+ n^{-2}_\circ \hat{m}_{\circ01}(x,y)\{1 - m^{-1}_\circ \hat{m}_{\circ01}(x,y)\}
+ 2n^{-3}_\circ \hat{m}_{\circ10}(x,y)\hat{m}_{\circ01}(x,y)\Pr^*(\tau^*_X = x, \tau^*_Y = y)\right],
$$

where $E_*\delta^* = \sum_{(x,y) \in \Omega} n^{-1}_\circ \{\hat{m}_{\circ10}(x,y) - \hat{m}_{\circ01}(x,y)\}\Pr^*(\tau^*_X = x, \tau^*_Y = y)$.

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 $I(X_\bullet \leq \tau_{0X})$ and $I(Y_\bullet \leq \tau_{0Y})$ and between $I(X_\circ \leq \tau_{0X})$ and $I(Y_\circ \leq \tau_{0Y})$, is less than one. Under conditions (i), (ii), and (iii) in Theorem 1, as for each biomarker, conditional on the data, $n^{1/2}_\circ(\hat{\delta} - \delta)$ converges in distribution to the same limit as $n^{1/2}_\circ(\hat{\delta} - \delta_0)$, which is normal with mean zero.
If condition (iv) in Theorem 1 holds additionally for each biomarker, \( n_0 \hat{\sigma}^2_\delta \) converges in probability to the asymptotic variance of \( n_0^{1/2}(\hat{\delta} - \delta_0) \).

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

Using the bootstrap variance estimator and bootstrap distribution of \( \hat{\delta} \), the Wald and percentile confidence intervals can be constructed for \( \delta_0 \). 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 \( \hat{\sigma}^2_\delta \) and \( \hat{\sigma}^2_X + \hat{\sigma}^2_Y \):

\[
\left[ \delta : (\hat{\delta} - \delta)^2 < \frac{z_{\alpha/2}^2 \hat{\sigma}^2_\delta}{n_0 (\hat{\sigma}^2_X + \hat{\sigma}^2_Y)} \left\{ \frac{\hat{\sigma}^2_X}{\hat{\sigma}^2_{2X}} (\delta + \hat{\phi}_Y)(1 - \delta - \hat{\phi}_Y) + \frac{\hat{\sigma}^2_Y}{\hat{\sigma}^2_{2Y}} (\hat{\phi}_X - \delta)(1 - \hat{\phi}_X + \delta) \right\} \right].
\]

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 \( \hat{\phi} \) would have an oscil-
lating component with variable $n_\bullet$. A more formal analysis, given in the Appendix, shows that

$$E(\hat{\phi}) - \phi_0 = \frac{F'_0(\tau_0)}{F'_0(\tau_0)} \left( \frac{[n_\bullet(1 - \rho_0)] - 1 + \rho_0}{n_\bullet + 1} \right)$$

$$+ \left\{ \frac{F''_0(\tau_0)}{F'_0(\tau_0)^2} - \frac{F'_0(\tau_0)F''_0(\tau_0)}{2F'_0(\tau_0)^3} \right\} \frac{\rho_0(1 - \rho_0)}{2n_\bullet} + o(n_\bullet^{-1}), \quad (4.1)$$

provided $F_\bullet$ is continuous and strictly increasing, and the second derivatives, $F''_0$ and $F''_\omega$, exist and are continuous in a neighborhood of $\tau_0$. This bias result is sharper than that of Lloyd and Yong (1999, theorem 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 $[\cdot]$ as the floor function. Consider two threshold estimates as the $[(n_\bullet + 1)(1 - \rho_0)]$th and the $[(n_\bullet + 1)(1 - \rho_0)]$th order statistics of the case biomarkers. If the two order statistics are the same, they lead to the estimated specificity $\hat{\phi}$, because $[(n_\bullet + 1)(1 - \rho_0)] \leq [n_\bullet(1 - \rho_0)] \leq [(n_\bullet + 1)(1 - \rho_0)]$. Otherwise, obtain the weighted average of the empirical specificities at these two thresholds as $\tilde{\phi}$, with weights $[(n_\bullet + 1)(1 - \rho_0)] - (n_\bullet + 1)(1 - \rho_0)$ and $(n_\bullet + 1)(1 - \rho_0) - [(n_\bullet + 1)(1 - \rho_0)]$, respectively.
respectively. It is straightforward to show
\[ E(\tilde{\phi}) - \phi_0 = \left\{ \frac{F''(\tau_0)}{F'_*(\tau_0)^2} - \frac{F'_*(\tau_0)F''(\tau_0)}{F'_*(\tau_0)^3} \right\} \rho_0 (1 - \rho_0) \frac{1}{2n} + o(n^{-1}). \] (4.2)

The two estimators, \( \hat{\phi} \) and \( \tilde{\phi} \), are asymptotically equivalent to each other to the first order, and coincide when \((n + 1)(1 - \rho_0)\) 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 \( \hat{\phi} \) with \( \tilde{\phi} \), 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 \( \tilde{\phi} \) 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 \( \tilde{\phi} \) still shares the same non-oscillating bias component of order \( n^{-1} \) with \( \hat{\phi} \). It is possible to further develop an estimator that is unbiased to order \( n^{-1} \) 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, 1000 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 dis-
5.1 Single-biomarker evaluation

tributions 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, section 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 $\phi_0$ 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 biomarker is measured. With both of the scales considered, these intervals tended to be much wider.
5.1 Single-biomarker evaluation

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 \( \phi_0 \), 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 \( \phi_0 \). Between them, the score intervals tended to have a shorter length in the middle value range of \( \phi_0 \), 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 \( n^* = n_0 \) from 20 to 200. The three best performers, the two score intervals using \( \hat{\phi} \) and \( \tilde{\phi} \) 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 the adopted point estimate \( \hat{\phi} \), as discussed in Section 4, and considerable under-coverage could arise at certain sample sizes. Not surprisingly, this occurred to the score interval using \( \hat{\phi} \) as well. In contrast, the coverage probability of the score interval using \( \tilde{\phi} \) was much more stable, with little oscillation.
5.2 Two-biomarker comparison

Many of these confidence intervals are closely related to those for binomial proportion (Agresti and Coull, 1998; Brown et al., 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 $n_x = n_{oX} = n_y = n_{oY}$ 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 \( \geq 7 \), 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 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 et al. (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 et al. (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 $nF^*(x) \sim \text{Binomial}\{n, \hat{F}(x)\}$ given the data,

\[
\Pr^*\{F^*(x) \geq p\} = \Pr^*\{nF^*(x) \geq \lceil np \rceil\} = \sum_{k=\lceil np \rceil}^{n} \binom{n}{k} \hat{F}(x)^k(1 - \hat{F}(x))^{n-k} = \lceil np \rceil \left( \frac{n}{\lceil np \rceil} \right) \int_{0}^{\hat{F}(x)} y^{\lceil np \rceil - 1}(1 - y)^{n-\lceil np \rceil} dy = \Pr^*\{B \leq \hat{F}(x)\},
\]

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), $F^*(x) \geq p$ if and only if $F^*\left\{-1\left( (1 - \rho_0)^+ \right) \right\} \leq x$, and $B \leq \hat{F}(x)$ if and only if $\hat{F}^{-1}(B) \leq x$. 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 $\phi = \hat{F}_o[\hat{F}_o^{-1}\{(1 - \rho_0)^+\}-]$, $\phi$ is the plug-in estimator in the map $\{F_\bullet, F_o\} \mapsto \phi_0$ decomposed
as
\[
\begin{align*}
F_\ast \in D[a,b] &\mapsto \tau_0 \in R, \\
F_0 \in D[a,b] &\mapsto \phi_0 \in R. \\
\end{align*}
\]
With \( F_\ast \) being differentiable with positive derivative at \( \tau_0 \), \( F_\ast \mapsto \tau_0 \) is Hadamard-differentiable at \( F_\ast \) tangentially to the set of functions \( h \in D[a,b] \) that are continuous at \( \tau_0 \), with derivative \( -h(\tau_0)/F'_\ast(\tau_0) \) (e.g., van der Vaart, 1998, lemma 21.3). Meanwhile, given \( F_0 \) being differentiable at \( \tau_0 \), it can be shown that \( (\tau_0, F_0) \mapsto \phi_0 \) is Hadamard-differentiable at \( (\tau_0, F_0) \) tangentially to the set \( \{k : k \in R\} \times \{l : l \in D[a,b], \text{continuous at } \tau_0\}, \) with derivative \( F'_0(\tau_0)k + l(\tau_0) \). By the chain rule, \( \{F_\ast, F_0\} \mapsto \phi_0 \) is then Hadamard-differentiable at \( \{F_\ast, F_0\} \) tangentially to the set \( \{h : h \in D[a,b], \text{continuous at } \tau_0\} \times \{l : l \in D[a,b], \text{continuous at } \tau_0\}, \) with derivative \( -h(\tau_0)F'_0(\tau_0)/F'_\ast(\tau_0) + l(\tau_0) \). As the Hadamard-differentiability implies continuity, \( \hat{\phi} \) is strongly consistent for \( \phi_0 \) following the strong consistency of \( \{\hat{F}_\ast, \hat{F}_0\} \) for \( \{F_\ast, F_0\} \) by the Glivenko–Cantelli theorem. With the weak convergence of \( n^{1/2}_\ast (\hat{F}_\ast - F_\ast) \) and \( n^{1/2}_0 (\hat{F}_0 - F_0) \), \( \hat{\phi} \) is \( AN(\phi_0, \sigma^2) \) following the functional delta method and the asymptotic normality of \( \{\hat{F}_\ast(\tau_0), \hat{F}_0(\tau_0)\} \) by the central limit theorem. Furthermore, by the results on empirical bootstrap and delta method for bootstrap (van der Vaart, 1998, theorems 23.7 and 23.9), the conditional distribution of \( n^{1/2}_0 (\phi^* - \hat{\phi}) \) given the observed data converges in distribution...
to the same limit as $n_o^{1/2}(\tilde{\phi} - \phi_0)$.

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

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

for any constant $b$.

The asymptotic normality result on $\tilde{\phi}$ can be extended to the weak convergence of $\tilde{\phi}(\rho)$ in a neighborhood of $\rho$ around $\rho_0$ 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 \( n_0^{1/2} \{ \hat{\phi}(\rho) - \phi(\rho) \} \) converges weakly to a Gaussian process in a neighborhood of \( \rho \) around \( \rho_0 \), for any \( d_{n_0} \downarrow 0 \), one can show

\[
\sup_{|d| \leq d_{n_0}} |\hat{\phi}(\rho_0 + d) - \hat{\phi}(\rho_0) - \phi(\rho_0 + d) + \phi(\rho_0)| = o_p(n_0^{-1/2});
\]

see, for example, Huang (2017, appendix). Meanwhile, the differentiability of \( \phi(\rho) \) at \( \rho_0 \) implies

\[
\sup_{|d| \leq d_{n_0}} d^{-1}|\phi(\rho_0 + d) - \phi(\rho_0) - \phi'(\rho_0)d| = o(1),
\]

where \( \phi'(\rho_0) = -F'(\tau_0)/F'(\tau_0) \).

Let \( c_{n_0} = \{ \log n_0/(n_0 + 2) \}^{1/2} \). Following (A.1), \( \Pr\{|B_\bullet - E(B_\bullet)| > c_{n_0}\} \leq 2n_0^{-2} \). Given \( E(B_\bullet) - \rho_0 = O(n_0^{-1}) \), \( |B_\bullet - \rho_0| = |B_\bullet - E(B_\bullet)| + O(n_0^{-1}) \).

Note that \( E_\bullet \) takes expectation over \( B_\bullet \). Then,

\[
E_\bullet \{ \hat{\phi}(B_\bullet) \} = E_\bullet \hat{\phi}(B_\bullet) I\{|B_\bullet - E(B_\bullet)| \leq c_{n_0}\} + O(n_0^{-2})
\]

\[
= \hat{\phi}(\rho_0) + \phi'(\rho_0)E[(B_\bullet - \rho_0)I\{|B_\bullet - E(B_\bullet)| \leq c_{n_0}\}]
\]

\[
+ o_p\{n_0^{-1/2} + E|B_\bullet - \rho_0|\}
\]

\[
= \hat{\phi} + o_p(n_0^{-1/2}),
\]

since \( E|B_\bullet - E(B_\bullet)| = O(n_0^{-1/2}) \) following \( \text{Var}(B_\bullet) = O(n_0^{-1}) \) by Jensen’s inequality. Similarly, \( E_\bullet \{ \hat{\phi}(B_\bullet)^2 \} = \hat{\phi}^2 + o_p(n_0^{-1/2}) \). Therefore, \( n_0 \hat{\sigma}_2^2 = \)
$E_*(\hat{\phi}(B_\bullet)\{1 - \hat{\phi}(B_\bullet)\})$ converges to $n_\circ \sigma_2^2$ in probability. By similar arguments,

$$\hat{\sigma}_1^2 = E_*[\{\hat{\phi}(B_\bullet) - \hat{\phi} + o_p(n_\circ^{-1/2})\}^2]$$

$$= E[\{\phi'(\rho_0)(B_\bullet - \rho_0)\}^2I\{|B_\bullet - E(B_\bullet)| \leq c_{n_\circ}\}] + o(n_\circ^{-1})$$

$$= \sigma_1^2 + o(n_\circ^{-1}).$$

Then, $n_\circ \hat{\sigma}_1^2$ converges to $n_\circ \sigma_1^2$ in probability. Thus, $n_\circ \hat{\sigma}^2$ 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 $\hat{\phi}$ in Section 4**

Following Remark 1, $\hat{\tau}$ has the same distribution as $F_\circ^{-1}(1 - B_\bullet)$. Thus,

$$E(\hat{\phi}) = E\{E(\hat{\phi} \mid \hat{\tau})\} = E\{F_\circ(\hat{\tau})\} = E\{\phi(B_\bullet)\}.$$
\[ c_n = \{ \log n / (n + 2) \}^{1/2}, \]

\[
E\{ \phi(B_\bullet) \} = E\{ \phi(B_\bullet) I(|B_\bullet - E(B_\bullet)| \leq c_n) \} + O(n^{-2})
\]

\[
= \phi\{E(B_\bullet)\} + \phi'[E(B_\bullet)] E[\{B_\bullet - E(B_\bullet)\} I(|B_\bullet - E(B_\bullet)| \leq c_n)]
\]

\[
+ E[\phi''(E(B_\bullet)) / 2 + o(1)]\{B_\bullet - E(B_\bullet)\}^2 I(|B_\bullet - E(B_\bullet)| \leq c_n)\}
\]

\[
+ O(n^{-2})
\]

\[
= \phi\{E(B_\bullet)\} + \{ \phi''(\rho_0) / 2 + o(1) \} E[\{B_\bullet - E(B_\bullet)\}^2] + O(n^{-2})
\]

\[
= \phi_0 + \phi'(\rho_0) \left( 1 - \frac{r}{n_\bullet + 1} - \rho_0 \right) + \phi''(\rho_0) \rho_0 (1 - \rho_0) \frac{1}{2n_\bullet} + o(n^{-1}),
\]

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 \( \rho_0 \) under the assumptions given.

References


Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal
REFERENCES


Department of Biostatistics and Bioinformatics, Emory University

E-mail: yhuang5@emory.edu

Ann & Robert H. Lurie Children’s Hospital of Chicago

E-mail: iparakati@luriechildrens.org

Department of Urology, Emory University

E-mail: dattatraya.patil@emory.edu, martinsanda@emory.edu
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.

<table>
<thead>
<tr>
<th>$\phi_{0X}$</th>
<th>$\delta_0$</th>
<th>n</th>
<th>unpaired biomarkers</th>
<th>paired biomarkers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Wald Pct Score Wald Pct Score</td>
<td>Wald Pct Score</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0</td>
<td>50</td>
<td>C 966 968 987 976 979</td>
<td>979 978 991 982 986</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 579 579 576 467 445</td>
<td>516 516 515 430 409</td>
</tr>
<tr>
<td>200</td>
<td>C 950 951 975 956 955</td>
<td>959 956 979 963 958</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 293 293 292 275 268</td>
<td>261 261 261 248 242</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
<td>50</td>
<td>C 956 963 985 970 970</td>
<td>964 974 986 970 965</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 671 671 663 540 525</td>
<td>600 600 595 500 486</td>
</tr>
<tr>
<td>200</td>
<td>C 939 942 966 942 944</td>
<td>948 949 972 955 953</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 339 339 337 318 313</td>
<td>303 303 302 288 284</td>
</tr>
<tr>
<td>0.6</td>
<td>0.4</td>
<td>50</td>
<td>C 955 956 984 946 947</td>
<td>962 968 989 955 958</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 671 674 668 556 551</td>
<td>607 608 605 517 512</td>
</tr>
<tr>
<td>200</td>
<td>C 955 953 971 959 942</td>
<td>964 955 980 961 957</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 339 339 337 321 319</td>
<td>306 306 304 293 291</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0</td>
<td>50</td>
<td>C 957 959 981 957 955</td>
<td>964 970 991 963 969</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 752 753 751 601 600</td>
<td>663 663 664 552 552</td>
</tr>
<tr>
<td>200</td>
<td>C 944 941 968 949 943</td>
<td>961 959 979 966 957</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 381 381 379 356 356</td>
<td>339 339 337 321 321</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>50</td>
<td>C 962 960 980 959 956</td>
<td>970 964 993 966 962</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 760 760 765 599 609</td>
<td>669 669 675 551 560</td>
</tr>
<tr>
<td>200</td>
<td>C 956 949 970 954 949</td>
<td>964 959 983 966 959</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 380 380 379 355 358</td>
<td>339 339 338 321 324</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>50</td>
<td>C 960 959 980 927 928</td>
<td>967 969 979 927 932</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 582 586 592 519 522</td>
<td>537 539 544 486 488</td>
</tr>
<tr>
<td>200</td>
<td>C 953 950 969 946 944</td>
<td>960 956 971 949 944</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 290 290 288 281 281</td>
<td>264 264 263 257 258</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
<td>50</td>
<td>C 957 962 973 936 939</td>
<td>968 969 978 936 942</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 679 679 685 562 569</td>
<td>608 608 612 519 526</td>
</tr>
<tr>
<td>200</td>
<td>C 956 949 966 947 948</td>
<td>961 957 979 960 954</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 337 337 335 319 321</td>
<td>305 305 304 291 293</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>50</td>
<td>C 957 963 979 944 958</td>
<td>965 966 985 953 951</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 690 690 701 552 569</td>
<td>614 614 622 510 526</td>
</tr>
<tr>
<td>200</td>
<td>C 961 960 967 956 947</td>
<td>961 962 979 963 967</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 337 337 336 316 321</td>
<td>304 304 303 288 293</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>50</td>
<td>C 970 961 989 957 946</td>
<td>971 971 990 970 969</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 619 619 639 492 519</td>
<td>551 551 569 454 480</td>
</tr>
<tr>
<td>200</td>
<td>C 965 956 975 962 952</td>
<td>974 967 983 967 963</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L 289 289 289 272 279</td>
<td>258 258 259 245 252</td>
</tr>
</tbody>
</table>

size = $n_{\cdot X} = n_{\cdot Y} = n_{\cdot Y}$ for unpaired, and size = $n_{\cdot} = n_{\cdot}$ for paired comparison.
C: coverage probability ($\times 1000$); L: average interval length ($\times 1000$). Pct: percentile interval.
Figure 1: Simulation summaries of 95% confidence intervals for specificity at controlled 95% sensitivity in the single-biomarker evaluation, under fixed sizes \( n_* = n_o = 50 \) and variable \( \phi_0 \). 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.
Figure 2: Simulation summaries of 95% confidence intervals for specificity at controlled 95% sensitivity in the single-biomarker evaluation, under fixed $\phi_0$ and variable sizes $n_\bullet = n_\circ$. Score intervals, based on $\hat{\phi}$ and $\tilde{\phi}$, and the exact bootstrap-based Zhou–Qin are included, with the same labeling as in Figure 1.
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.