LINEAR BIOMARKER COMBINATION FOR CONSTRAINED CLASSIFICATION

BY YIJIAN HUANG1 AND MARTIN G. SANDA2

1Department of Biostatistics and Bioinformatics, Emory University, yhuang5@emory.edu
2Department of Urology, Emory University, martinsanda@emory.edu

Multiple biomarkers are often combined to improve disease diagnosis. The uniformly optimal combination, i.e., with respect to all reasonable performance metrics, unfortunately requires excessive distributional modeling, to which the estimation can be sensitive. An alternative strategy is rather to pursue local optimality with respect to a specific performance metric. Nevertheless, existing methods may not target clinical utility of the intended medical test, which usually needs to operate above a certain sensitivity or specificity level, or do not have their statistical properties well studied and understood. In this article, we develop and investigate a linear combination method to maximize the clinical utility empirically for such a constrained classification. The combination coefficient is shown to have cube root asymptotics. The convergence rate and limiting distribution of the predictive performance are subsequently established, exhibiting robustness of the method in comparison with others. An algorithm with sound statistical justification is devised for efficient and high-quality computation. Simulations corroborate the theoretical results, and demonstrate good statistical and computational performance. Illustration with a clinical study on aggressive prostate cancer detection is provided.

1. Introduction. In this era of precision medicine, many biomarkers have been identified for disease diagnosis, as well as for disease prognosis and prediction of therapeutic response. Since a single biomarker often has only limited diagnostic accuracy, combining multiple biomarkers holds the promise for improved discrimination. This is a classical binary classification problem, with a vast relevant literature in statistics, machine learning, and econometrics (cf. Hastie, Tibshirani and Friedman, 2009). Nevertheless, challenges abound. Intuitively, the conditional probability of the disease status given the biomarkers, i.e., the posterior probability, is a uniformly optimal combination with respect to any reasonable performance metrics including classification error and expected cost. Since an optimal combination remains so under monotone transformation, interestingly such a combination may be determined even when the posterior probability cannot be identified from the observed data with, for example, case-control design as often adopted in biomarker studies. Specifically, as following from Bayes’ rule, the likelihood ratio is such a transformation and optimal; see McIntosh and Pepe (2002) and the references therein. These results lead to a common strategy via estimating the posterior probability or the biomarker distributions of cases and controls. Parametric modeling is often adopted, with popular methods including linear discriminant analysis, logistic regression, and probit regression. In particular, logistic regression is routinely employed to linearly combine biomarkers in practice. However, correct model specification can be difficult in, for example, cancer detection, where cancer biomarkers often have irregular distributions. Obviously, model misspecification could result in sub-optimality.


Keywords and phrases: Bahadur representation, Cube root asymptotics, Diagnostic test, Sensitivity, Specificity.
Semiparametric methods, including the monotonic density ratio model of Chen et al. (2016), have been pursued, but they still may not be sufficiently general. On the other hand, nonparametric density estimation with multiple biomarkers is notoriously difficult.

Of course, for classification purpose, a locally optimal combination, i.e., with respect to a given performance metric of interest, is sufficient. Such a combination does not need to be uniformly optimal and thus can be easier for identification and estimation (cf. Elliott and Lieli, 2013), as pursued by many methods. They include machine learning techniques such as random forest, support vector machine, and neural network. For instance, the support vector machine targets the hinge loss function. Another notable example involves the receiver operating characteristic curve (ROC) to use area under the curve (AUC) as the performance metric. Pepe, Cai and Longton (2006) proposed a linear combination to maximize the empirical AUC; see also Ma and Huang (2007) and Lin et al. (2011) which maximized smoothed versions of the empirical AUC, as well as related work by Vexler et al. (2006), Chen, Vexler and Markatou (2015), and Fong, Yin and Huang (2016). However, all these loss functions and performance metrics do not align well with clinical utility of diagnostic tests in practice. As an exception, Elliott and Lieli (2013) incorporated cost function into the classification optimization by extending the maximum score estimator of Manski (1975, 1985). Their empirical risk minimization accommodates differential costs associated with false positives and false negatives. Nevertheless, the cost quantification as required can be extremely difficult in practice, posing a potential limitation.

Considerable efforts have been made on locally optimal combination with respect to clinical utility-relevant performance metrics that are more practical. Note that a medical test typically needs to operate above a certain sensitivity or specificity level. Take aggressive prostate cancer diagnosis as an example. Since a positive non-invasive test would be confirmed with biopsy, a false negative has far more serious clinical consequences than a false positive. Thus, a high sensitivity, say, 95%, needs to be maintained (e.g., Catalona et al., 1998; Sanda et al., 2017). Partial AUC around the high sensitivity level then is a sensible metric, and perhaps more so is specificity at the controlled sensitivity level; indeed the latter may be viewed as a limiting special case. Of course, a different clinical context might rather call for a high specificity, which then needs to be controlled instead. Interestingly, specificity at controlled sensitivity or vice versa is in line with power at controlled type I error in standard statistical hypothesis testing; incidentally, this explains the applicability of the Neyman–Pearson lemma to optimal biomarker combination as discussed in McIntosh and Pepe (2002) and Eguchi and Copas (2002). A number of methods have been developed for optimal linear combination with respect to these metrics, including Pepe and Thompson (2000) and Yan et al. (2018) using partial AUC, and Meisner et al. (2021) targeting sensitivity at a controlled specificity level. Unfortunately, statistical properties of these procedures are not well understood.

Notably, most of these methods targeting a given performance metric are restricted to the class of linear combinations. As such, they may only approach the optimality in the linear class, but not necessarily so in general. Nevertheless, linear combination is of interest in its own right for a few reasons. First, it provides a first-order approximation to the general combination. Second, reliable estimation in a more general class may not be feasible in many applications where data are limited. Finally, nonlinear combinations can be accommodated in a linear combination method through the general linear basis expansion technique as adopted in, e.g., support vector machines (Hastie, Tibshirani and Friedman, 2009); see further discussion in Section 6.

In this article, we investigate and develop an optimal linear combination estimation method aiming at sensitivity- or specificity-constrained classification by maximizing the empirical utility. For simplicity in expressions, we opt to consider sensitivity at a controlled specificity level in the methodology development. The same method also applies to specificity
at a controlled sensitivity level as of interest in our motivating application, by transposing the roles of cases and controls. As natural as the empirical estimator appears, its asymptotic analysis and computation are challenging due to noncontinuity in both the objective function and the constraint of the optimization problem. Our approach is different from Meisner et al. (2021) which adopted kernel-smoothed estimates of sensitivity and specificity instead. Most importantly, the limiting distribution of our combination coefficients is established, to have non-standard cube root asymptotics. Subsequently, the convergence rate and limiting distribution of the predictive performance are obtained. Meanwhile, a novel computational algorithm with sound statistical justification is devised.

The rest of the paper is organized as follows. Section 2 will introduce the problem and the empirical utility maximization method, to set the stage. We present the asymptotic study in Section 3, and the computational algorithm in Section 4. Numerical studies are reported in Section 5. Final remarks are provided in Section 6. Technical details are deferred to the Appendices.

2. The problem and empirical utility maximization. Write a $k$-vector biomarker under consideration as $M_d$, for $k \geq 2$ and with $d = 1, 0$ denoting case and control, respectively. With coefficient $b$, a linear combination is given by $b^\top M_d$. Since sensitivity at controlled specificity is scale invariant to the combination, we limit to $\{b^\top M_d : \|b\|_1 = 1, b \in \mathbb{R}^k\}$ as the linear combination class under consideration, where $\|\cdot\|_1$ is the $\ell_1$ norm. Other norms for $b$ can serve the same purpose, but our choice has computational advantages as will be seen later.

2.1. Optimal linear combination. Adopt the convention that a positive diagnosis is associated with a larger combination value. With $\rho \in (0, 1)$ as the given control level of specificity, the optimal combination coefficient $\beta$ to maximize the sensitivity is the solution $b$ of the following optimization problem:

$$
\max_{b: \|b\|_1 = 1} \Pr(b^\top M_1 > t) \quad \text{subject to } \Pr(b^\top M_0 \leq t) \geq \rho,
$$

where $t$ corresponds to a test threshold.

Write $F_d(t; b) = \Pr(b^\top M_d \leq t)$ for the cumulative distribution function, $d = 1, 0$, and $F_0^{-1}(\rho; b) \equiv \inf\{t : F_0(t; b) \geq \rho\}$ for the control quantile function. With a given $b$, the smallest threshold $t$ with the specificity being at least $\rho$ and the resulting constrained sensitivity are given by

$$
\tau(b) = F_0^{-1}(\rho; b), \quad \phi(b) = 1 - F_1\{\tau(b); b\},
$$

respectively. Then, we have

$$
\beta = \arg \max_{b: \|b\|_1 = 1} \phi(b),
$$

as an identity for the optimal combination coefficient.

2.2. Empirical utility maximization. Consider a case-control study, with $n_d$ independent replicates of $M_d$: $M_{d, [i]}, i = 1, \ldots, n_d$, where $d = 1, 0$ represents the case and control samples, respectively. As a natural estimator of $\beta$, $\hat{\beta}$ is the solution $b$ of the following problem to optimize the empirical utility:

$$
\max_{b: \|b\|_1 = 1} \hat{E}(b^\top M_1 > t) \quad \text{subject to } \hat{E}(b^\top M_0 \leq t) \geq \rho,
$$
where \( I(\cdot) \) is the indicator function and \( \hat{E} \) denotes the empirical average, over the case or control sample as appropriate; for example, \( \hat{E} I(b^T M_1 > t) = n_1^{-1} \sum_{i=1}^{n_1} I(b^T M_{1,[i]} > t) \). Let \( \hat{F}_d(t; b) \) denote the empirical counterpart of \( F_d(t; b) \). The empirical estimators for the threshold and associated constrained sensitivity are then given by

\[
\hat{\tau}(b) = \hat{F}_0^{-1}(\rho; b), \quad \hat{\phi}(b) = 1 - \hat{F}_1\{\tau(b); b\},
\]

respectively. Subsequently,

\[
\hat{\beta} = \arg \max_{b: ||b||_1 = 1} \hat{\phi}(b),
\]

which will facilitate later analysis of the estimator.

### 2.3. A related problem.
Consider the circumstance that a biomarker is known \textit{a priori} to have a non-zero coefficient with a specific sign in the optimal linear combination. Without loss of generality, suppose that the first biomarker is such a so-called anchor with a positive coefficient, which can always be achieved by biomarker reordering and sign-flipping. Write \( M_d \equiv (M_{d,1}, M_{d,-1})^T \), \( \beta \equiv (\beta_1, \beta_{-1})^T \), and \( \hat{\beta} \equiv (\hat{\beta}_1, \hat{\beta}_{-1})^T \). Then, instead of \( \{b^T M_d : ||b||_1 = 1, b \in \mathbb{R}^k\} \), one may work with a restricted class \( \{M_{d,1} + h^T M_{d,-1} : h \in \mathbb{R}^{k-1}\} \).

With a slight abuse of notation, in this context, write \( F_d(t; h) \equiv F_d(t; (1, h^T)^T) \) and similarly for \( F_d(t; h), \tau(h), \hat{\tau}(h), \phi(h), \) and \( \hat{\phi}(h) \). With \( \beta_1 > 0 \) and scale invariance of \( \phi(b) \) to \( b \), we have the optimal combination coefficient in the new class,

\[
\eta = \beta_{-1}/\beta_1 = \arg \max_{h} \phi(h).
\]

Its empirical estimator is given by

\[
\hat{\eta} = \arg \max_{h} \hat{\phi}(h),
\]

or, equivalently, the solution \( h \) of

\[
\max_{h, t} \hat{E} I(M_{1,1} + h^T M_{1,-1} > t) \quad \text{subject to} \quad \hat{E} I(M_{0,1} + h^T M_{0,-1} \leq t) \geq \rho.
\]

The optimization problem (6) is clearly a simplification, and such a restricted class has been commonly adopted for linear combinations in the literature, e.g., Pepe, Cai and Longton (2006). However, the requirement of prior knowledge imposes a restriction, and we do not take it as an equivalent problem in general, especially for computational purposes. Nevertheless, provided that the consistency of \( \hat{\beta} \) is established for \( \beta \), this restricted problem may be taken as a surrogate for the purposes of weak convergence. Specifically, under the set-up of \( \beta_1 > 0, \hat{\beta}_1 > 0 \) and subsequently \( \hat{\eta} = \hat{\beta}_{-1}/\hat{\beta}_1 \) hold with probability tending to 1. In that asymptotic sense, these two estimated combinations differ by a scale and thus are equivalent in terms of performance. Therefore, for weak convergence, we can instead work with \( \hat{\eta} \) in the restricted problem, which is advantageous since the distribution of \( \hat{\beta} \) is concentrated on a \((k-1)\)-dimensional subspace of \( \mathbb{R}^k \) by definition.

### 3. Asymptotic theory.
An asymptotic study for this non-standard estimation problem, via optimization (3), is conducted under the circumstance that the number of biomarkers \( k \) is held fixed, and the total sample size \( n = n_1 + n_0 \) approaches \( \infty \). Recall that the case and control samples are independent of each other, and each of them consists of independent and identically distributed observations. Consistency shall be tackled first. We then establish a uniform Bahadur representation for the estimated thresholds to obtain an approximation for the empirical constrained sensitivities, of nearly optimal combinations. On the basis of this approximation, weak convergence results are subsequently pursued.
3.1. **Strong consistency.** Mild regularity conditions are imposed.

**CONDITION 1** (Case and control sizes). The size ratio \( n_1/n_0 \) converges to a constant \( \gamma \in (0, \infty) \) as \( n \to \infty \).

**CONDITION 2** (Threshold uniqueness). For any \( \varepsilon > 0 \), the threshold \( \tau(b) \) satisfies

\[
\sup_{b: \|b\|_1 = 1} F_0(\tau(b) - \varepsilon; b) < \rho < \inf_{b: \|b\|_1 = 1} F_0(\tau(b) + \varepsilon; b).
\]

**CONDITION 3** (Sensitivity continuity). For each \( b \) such that \( \|b\|_1 = 1 \), \( F_1(t; b) \) is continuous at \( t = \tau(b) \).

Condition 1 implies that \( n_1 \) and \( n_0 \) in \( O(\cdot) \) and \( o(\cdot) \) notation expressions can be equivalently replaced with \( n \), as we shall do consistently. Conditions 2 and 3 concern the biomarker distributions at the thresholds for the combinations, but not elsewhere. Nevertheless, the requirement for each combination might not be necessary especially for \( b \) such that \( \phi(b) \) is far away from the optimum. Simplicity is part of the consideration in adopting the current form.

The conditions so far guarantee the existence of an optimal combination. However, the optimal combination is not necessarily unique as in the case, for example, that \( M_1 \) and \( M_0 \) lie in a linear subspace of dimension less than \( k \) with probability 1. Then, \( \beta \) is not well defined and much less the notion of consistency for its estimation. For that reason, we further require the following uniqueness condition.

**CONDITION 4** (Optimal combination uniqueness). The maximizer \( \beta \) of \( \phi(b) \) over \( \{b: \|b\|_1 = 1, b \in \mathbb{R}^k\} \) is unique.

**THEOREM 3.1.** Under Conditions 1–3, a maximizer \( \beta \) of \( \phi(b) \) over \( \{b: \|b\|_1 = 1, b \in \mathbb{R}^k\} \) exists. Consider an estimator \( \hat{\beta} \) such that \( \|\hat{\beta}\|_1 = 1 \), satisfying

\[
\hat{\phi}(\hat{\beta}) - \max_{b: \|b\|_1 = 1} \hat{\phi}(b) = o(1),
\]

almost surely. Then, both \( \phi(\hat{\beta}) \) and \( \hat{\phi}(\hat{\beta}) \) converge to \( \phi(\beta) \) almost surely. If, in addition, Condition 4 holds, then \( \hat{\beta} \) converges to \( \beta \) almost surely.

This result concerns a near maximizer \( \tilde{\beta} \), of which \( \hat{\beta} \) is a special case. While \( \phi(\beta) \) represents the ideal performance, the predictive performance \( \hat{\phi}(\hat{\beta}) \) reflects how the learned combination \( \beta \) performs on future data once implemented. Their difference \( \phi(\beta) - \hat{\phi}(\hat{\beta}) \), referred to as the performance deficiency, converges to 0 almost surely. Meanwhile, the empirical estimate \( \hat{\phi}(\hat{\beta}) \) is consistent for both \( \phi(\beta) \) and \( \phi(\beta) \). Appealingly, this consistency may hold even when \( \beta \) is not unique and thus \( \tilde{\beta} \) might not converge. Indeed, of most interest is the performance rather than the combination coefficient, which is a distinctive feature of this problem. Nevertheless, the uniqueness of \( \beta \) and consistency of \( \hat{\beta} \) would facilitate a rigorous study on weak convergence of the predictive performance. In some circumstances, non-uniqueness of \( \beta \) may be resolved as with the example that \( M_1 \) and \( M_0 \) lie in a linear subspace of dimension less than \( k \) by means of reducing the biomarker dimension.

3.2. **Approximating \( \hat{\phi}(b) \) via uniform Bahadur representation of \( \tau(h) \).** As discussed in Section 2.3, we may now switch to the simplified problem dealing with \( \eta \) and \( \hat{\eta} \) instead of \( \beta \) and \( \hat{\beta} \). Recall \( \beta_1 > 0 \). Then, by Theorem 3.1 under Conditions 1–4, \( \hat{\eta} \) converges to \( \eta \) almost surely. Write \( m^+ \equiv mI(m \geq 0) \) and \( m^- \equiv mI(m \leq 0) \) and, in the case of a vector, the operators \( + \) and \( - \) apply componentwise. Additional conditions are adopted.
CONDITION 5 (Biomarker integrability). Biomarker $M_{d,-1}$ is integrable for $d = 1, 0$.

CONDITION 6 (Smoothness of control quantile). With the control anchor biomarker $M_{0,1}$, (i) the conditional distribution function $\Pr(M_{0,1} + h^T M_{0,-1} \leq t \mid M_{0,-1})$ has a bounded density for $(h^T, t)^\top$ in a neighborhood of $\{\eta^\top, \tau(\eta)\}^\top$; and (ii) possibly upon a location shift for $M_{0,-1}$, the density of the marginal distribution function $\Pr(M_{0,1} + h^T M_{0,-1} + h^T M_{0,-1} \leq t)$ is bounded away from 0 for $(h^T, t)^\top$ in a neighborhood of $\{\eta^\top, -\eta^\top, \tau(\eta)\}^\top$.

CONDITION 7 (Smoothness of case and control distributions). For $d = 1, 0$, the conditional distribution function $\Pr(M_{d,1} + h^T M_{d,-1} \leq t \mid M_{d,-1})$ has a bounded second derivative with respect to $t$, for $(h^T, t)^\top$ in a neighborhood of $\{\eta^\top, \tau(\eta)\}^\top$.

Conditions 6 and 7 require the marginal distribution of the anchor biomarker $M_{d,1}$, $d = 1, 0$, to have certain smoothness, but not so for other biomarkers. However, these non-anchor biomarkers need to have finite expectations by Condition 5.

REMARK 1. The biomarker combination in part (ii) of Condition 6 is more general than the linear combination under consideration, since the coefficients for positive or negative $M_{0,-1}$, $h_+$ or $-h_-$, respectively, may or may not be the same. However, in certain situations, the two are actually equivalent. Notable in the condition is the possible location shift for $M_{0,-1}$, to which both the linear combination coefficient and the constrained sensitivity are invariant. Nevertheless, the split point of a biomarker for different coefficients then may correspond to any finite value before the shift. Therefore, when each component of $M_{0,-1}$ has a finite upper or lower bound for the support of its distribution, only one of the corresponding components of $h_+$ and $-h_-$ is relevant upon shifting the finite bound to 0, to give rise to a linear combination. The more general combination is only essential when a component of $M_{0,-1}$ has a distribution with support on the whole real line as with, e.g., a normal distribution.

REMARK 2. When the distribution of $M_{d,-1}$ has a bounded support, $|\langle h - \eta \rangle^T M_{d,-1}|$ can be made arbitrarily small when $h$ is sufficiently close to $\eta$. In that case, Condition 6 is satisfied when the conditional density of the optimal combination $M_{0,1} + \eta^T M_{0,-1}$ given $M_{0,-1}$ is bounded from above and away from 0 in a neighborhood of $\tau(\eta)$. Similarly, Condition 7 is met when the conditional distribution of the optimal combination $M_{d,1} + \eta^T M_{d,-1}$ given $M_{d,-1}$ has a bounded second derivative in a neighborhood of $\tau(\eta)$.

REMARK 3. We have made these conditions slightly less general in the interest of clarity in exposition. The first biomarker, i.e., the anchor, serves a role different from the others in Conditions 5–7. Actually, that role may be taken by any biomarker with a non-zero coefficient in the optimal combination. Also, the three such biomarkers, one for Condition 6 and two for Condition 7 corresponding to $d = 1, 0$ separately, do not need to be the same. These conditions serve to ensure the marginal distribution of a linear combination to have certain smoothness around the threshold and around the optimal combination; see Appendix B. Only existence, but not identification, of these aforementioned biomarkers is required.

We first provide a uniform convergence rate of $\hat{\tau}(h)$, extending the classical result on the empirical quantile (cf. Serfling, 1980, lemma 2.5.4.B). Write $\| \cdot \|_\infty$ as the maximum norm.
**Lemma 3.2.** Under Conditions 1–6, there exists a constant \( \varepsilon > 0 \) such that
\[
\sup_{\|h - \eta\|_\infty \leq \varepsilon} |\hat{\tau}(h) - \tau(h)| = O\left\{ n^{-1/2} (\log n)^{1/2} \right\},
\]
almost surely.

Meanwhile, the local behavior result of the empirical distribution function (cf. Serfling, 1980, lemma 2.5.4.E) can also be generalized.

**Lemma 3.3.** Suppose that Conditions 1–5 and 7 hold and take \( c_0 > 0 \) as any given constant. For \( d = 1, 0 \) separately, there exists a constant \( \varepsilon > 0 \) such that
\[
\sup_{\|h - \eta\|_\infty \leq \varepsilon} \left| \hat{F}_d(\tau(h) + u; h) - \hat{F}_d(\tau(h); h) - F_d(\tau(h) + u; h) + F_d(\tau(h); h) \right| = O\left\{ n^{-3/4} (\log n)^{3/4} \right\},
\]
almost surely.

**Remark 4.** The proofs of the uniform results in Lemmas 3.2 and 3.3, given in Appendix B, exploit a monotonicity property of \( \hat{\tau}(h) \) and \( \hat{F}_d(t; h) \). That is, in the special case that the non-anchor biomarkers, \( M_{d,-1} \), \( d = 1, 0 \), are nonnegative, \( \hat{\tau}(h) \) is non-decreasing in each component of \( h \) whereas \( \hat{F}_d(t; h) \) is non-decreasing in \( t \) and non-increasing in each component of \( h \). Furthermore, the general problem can be so reformulated. This can be easily achieved, by means of biomarker shifting and sign-flipping, if the case and control distributions for each component of \( M_{d,-1} \) are bounded in the union of their supports from either above or below. More generally, we now develop a novel biomarker splitting technique to extend the problem to a linear combination with \( 2k - 1 \) biomarkers, \( (M_{d,1}, M_{d,-1}^T, M_{d,-1}^-) \). Now \( \eta \) in the original problem translates to \( (\eta^+, -\eta^-)^T \) in the extended one, with all non-anchor biomarkers being nonnegative. Focus on Lemma 3.2 with Condition 6. Note that an extended linear combination can be equivalently written as:
\[
M_{0,1} + h^+_1 M_{0,-1}^+ + h^-_1 M_{0,-1}^- = M_{0,1} + \{ h_+ \odot I(M_{0,-1} \geq 0) - h_- \odot I(M_{0,-1} < 0) \}^T M_{0,-1},
\]
where the indicator functions apply to \( M_{0,-1} \) componentwise, and \( \odot \) is the Hadamard product. Furthermore, conditioning on \( M_{0,-1}^+ \) and \( M_{0,-1}^- \) is the same as that on \( M_{0,-1} \). Therefore, part (i) of Condition 6 on the conditional distribution automatically holds for the extended problem if it does for the original one. Meanwhile, part (ii) of Condition 6 on the marginal distributions accommodates this extension as discussed in Remark 1. Therefore, the extended problem, with the monotonicity property, can be utilized to establish Lemma 3.2, upon noting that the result for the extended problem is more general. The same approach applies to Lemma 3.3 with Condition 7.

Write \( f_d(t; h) \) as the probability density function of \( F_d(t; h) \), \( d = 1, 0 \), if it exists. Let \( \lambda \equiv f_1(\tau(\eta) \eta) f_0(\tau(\eta) \eta)^{-1} \). The preceding lemmas give rise to a uniform Bahadur representation of the empirical threshold \( \hat{\tau}(h) \) and subsequently to an approximation of the empirical constrained sensitivity \( \bar{\phi}(h) \) by
\[
\bar{\phi}(h) = 1 - \hat{F}_1(\tau(h); h) + \lambda \{ \hat{F}_0(\tau(h); h) - \rho \},
\]
which is more amenable to analysis.
THEOREM 3.4. Suppose that Conditions 1–7 hold. There exists a constant \( \varepsilon > 0 \) such that
\[
(7) \quad \sup_{\|h - \eta\|_\infty \leq \varepsilon} |\hat{\mathbf{T}}(h) - \mathbf{T}(h) + f_0(\mathbf{T}(h); h)^{-1}[\hat{F}_0(\mathbf{T}(h); h) - \rho]| = O\{n^{-3/4}(\log n)^{3/4}\},
\]
almost surely. Furthermore, for any \( h_n \to \eta \),
\[
(8) \quad \hat{\phi}(h_n) - \phi(h_n) = \|h_n - \eta\|_\infty O\{n^{-1/2}(\log n)^{1/2}\} + O\{n^{-3/4}(\log n)^{3/4}\},
\]
almost surely.

The order of the remainder in (7), \( O\{n^{-3/4}(\log n)^{3/4}\} \), may be improved in the light of the sharper bound in the standard Bahadur representation (Bahadur, 1966; Kiefer, 1967). Nevertheless, for our purposes, the remainder needs only to be \( o_p(n^{-2/3}) \). In fact, a weaker version of (8),
\[
(9) \quad \hat{\phi}(h_n) - \phi(h_n) = \|h_n - \eta\|_\infty o_p(n^{-1/3}) + o_p(n^{-2/3}),
\]
is adequate for the ensuing results.

3.3. Weak convergence: cube root asymptotics. In their seminal work, Kim and Pollard (1990) established cube root asymptotics with their main theorem dealing with an unconstrained maximization of a one-sample empirical process. Our objective function \( \hat{\phi}(h) \), however, is a rather complicated functional of empirical processes involving two independent samples. We shall extend their result to our problem, by exploiting the approximation (9) through \( \hat{\phi}(h) \) as a linear combination of empirical processes from independent samples.

THEOREM 3.5. Suppose that Conditions 1–7 hold. Then, the process \( n_1^{2/3}\{\hat{\phi}(\eta) + n_1^{-1/3}a - \hat{\phi}(\eta)\} \) converges weakly to a Gaussian process \( Z(a) \) with continuous sample paths, mean \( a^\top H a / 2 \), and covariance kernel \( V \), where \( H \) and \( V \) are given by (26) and (27), respectively, in Appendix B. Consider \( \eta \) as an estimator of \( \eta \) satisfying
\[
\hat{\phi}(\eta) \geq \max_h \hat{\phi}(h) + o_p(n^{-2/3}).
\]
If \( H \) is negative definite and \( Z \) has nondegenerate increments, i.e., \( V(a, a) \neq 0 \) for \( a \neq 0 \), \( n_1^{1/3}(\eta - \eta) \) converges in distribution to \( U = \arg \max_a Z(a) \). Meanwhile, \( n_1^{2/3}\{\phi(\eta) - \phi(\eta)\} \) converges in distribution to \( U^\top H U / 2 \).

Of course, \( \eta \) is a special case of \( \eta \). The estimated combination coefficient \( \hat{\eta} \) has a \( n^{-1/3} \) convergence rate, leading to a \( n^{-2/3} \) convergence rate of the predictive performance \( \phi(\eta) \) to the ideal one \( \phi(\eta) \). This exhibits a distinctive and more robust convergence profile in comparison with other linear combination methods, with respect to the constrained sensitivity under consideration. Note that a number of methods originally developed for cohort studies, such as support vector classifier and the maximum score estimator (Manski, 1975, 1985), can be applied to case-control studies as well, although their loss functions and thus interpretation may change accordingly. Among existing methods, the maximum score estimator has the same \( n^{-1/3} \) convergence rate as \( \eta \) but may not converge to the same limit. Thus, its predictive performance does not approach the ideal one \( \phi(\eta) \) in general. Meanwhile, a large number of methods, including linear discriminant analysis, logistic regression, support vector classifier, and AUC maximization (Pepe, Cai and Longton, 2006), have the faster parametric \( n^{-1/2} \) convergence rate of their combination coefficients. Accordingly, in the case that their limits coincide with \( \eta \), their associated predictive performances have a faster \( n^{-1} \) convergence rate; this occurs, for example, for logistic regression when the model is correctly specified. Under other circumstances, however, their predictive performances may not approach the ideal one \( \phi(\eta) \), and also typically converge at a slower \( n^{-1/2} \) rate.
4. Computational algorithm. The optimization problem (3) is computationally challenging. Obviously brute-force grid search can be used, but the computational burden is prohibitive except with very few biomarkers. We have also approached the problem via modern mixed integer linear programming. Algorithmic advances and hardware improvements over the past three decades have dramatically sped up mixed integer optimization. The last few years has seen applications to several statistical problems, related or unrelated to ours, that were once regarded as intractable; see Florios and Skouras (2008) and Bertsimas, King and Mazumder (2016) among others. Most appealingly, this approach permits exact optimization or an approximate one with a definite error bound. Nevertheless, our experience showed that the computation is still too intensive, at least with personal computers, for typical datasets in practice. In the following, we suggest a different approach, not a purely computational one but rather a combined statistical and computational solution to balance efficiency and high quality. We shall devise an asymptotically equivalent optimization problem and then develop a novel computational algorithm, partly inspired by Ou, Zeng and Cai (2016) on a different problem.

Two features of the optimization problem (3) contribute to its poor computational properties. One is the nonlinear equality constraint, on \(|b|_1\), and the other is the indicator function in the objective function and the inequality constraint. For the former, because of scale invariance of \(\hat{\phi}(b)\) to \(b\),

\[
\hat{\beta} = \arg \max_{b, \|b\|_1 \leq 1} \hat{\phi}(b) + w\|b\|_1,
\]

with any constant \(w > 0\). This results in an equivalent optimization problem with inequality constraints only,

\[
\min_{b, t} \hat{E}I(b^\top M_1 \leq t) - w\|b\|_1
\]

subject to \(\rho - \hat{E}I(b^\top M_0 \leq t) \leq 0\)

\[
\|b\|_1 - 1 \leq 0;
\]

such a minimization formulation is more standard in the optimization literature. For the indicator function, approximating \(I(x \leq 0)\) by \(\sigma^{-1}\{x^- - (x + \sigma)^-\}\), for a small \(\sigma > 0\), gives rise to:

\[
\min_{b, t} \sigma^{-1}\hat{E}(b^\top M_1 - t)^- - \{\sigma^{-1}\hat{E}(b^\top M_1 - t + \sigma)^- + w\|b\|_1\}
\]

subject to \(\{\rho + \sigma^{-1}\hat{E}(b^\top M_0 - t + \sigma)^-\} - \sigma^{-1}\hat{E}(b^\top M_0 - t)^- \leq 0\)

\[
\|b\|_1 - 1 \leq 0.
\]

However, the solution \(b\) of this problem, say, \(\hat{\xi}_\sigma\), may no longer have unity \(\ell_1\) norm. A rescaling leads to \(\hat{\beta}_\sigma = \hat{\xi}_\sigma / \|\hat{\xi}_\sigma\|_1\).

This estimation procedure is better elucidated through an analysis. Add subscript \(\sigma\) to \(\hat{F}_d\), \(\hat{\tau}\), and \(\hat{\phi}\) to denote their counterparts after the indicator function approximation. From (11), we can see

\[
\hat{\xi}_\sigma = \arg \max_{b, \|b\|_1 \leq 1} \hat{\phi}_\sigma(b) + w\|b\|_1.
\]

With \(\sigma > 0\), \(\hat{\phi}_\sigma(b)\) is no longer scale invariant to \(b\) as \(\hat{\phi}_\sigma(sb) = \hat{\phi}_{s^{-1}\sigma}(b)\) for \(s > 0\), which explains that \(\|\hat{\xi}_\sigma\|_1\) may not be 1. Nevertheless, since \(\hat{\phi}_\sigma(b) \in [0, 1]\), we have \(1 + w\|\hat{\xi}_\sigma\|_1 \geq w\) or \(\|\hat{\xi}_\sigma\|_1 \geq 1 - w^{-1}\). Suppose \(w > 1\) and we then obtain

\[
\hat{\xi}_\sigma = \arg \max_{b, \|b\|_1 = s} \hat{\phi}_\sigma(b),
\]
for some data-dependent $s \in [1 - w^{-1}, 1]$. Subsequently,
\begin{equation}
\hat{\beta}_\sigma = \arg \max_{b} \left( \phi_{\sigma^{-1}} (b) \right).
\end{equation}

Meanwhile, in the circumstance of the first biomarker being the anchor as described in Section 2.3, write $\hat{\beta}_\sigma = (\hat{\beta}_{1,\sigma}, \hat{\beta}_{-1,\sigma})^\top$ and $\hat{\eta}_\sigma = \hat{\beta}_{1,\sigma} / |\hat{\beta}_{1,\sigma}|$. When $\hat{\beta}_{1,\sigma} > 0$, we have a similar identity,
\begin{equation}
\hat{\eta}_\sigma = \arg \max_{h} \left( \phi_{(s^{1/\sigma})^{-1}} (h) \right).
\end{equation}

On the other hand, from $\mathbb{I}(x \leq -\sigma) \leq \sigma^{-1} \{ \sigma - (x + \sigma) \} \leq \mathbb{I}(x \leq 0)$, we have $\hat{\beta}_d(t - \sigma; b) \leq \hat{\beta}_d(t; b) \leq \hat{\beta}_d(t; b), \hat{\tau}(b) \leq \hat{\tau}_\sigma(b) \leq \hat{\tau}(b) + \sigma$, and subsequently
\begin{equation}
1 - \hat{\beta}_1 \{ \hat{\tau}(b) + \sigma; b \} \leq \hat{\phi}_\sigma(b) \leq 1 - \hat{\beta}_1 \{ \hat{\tau}(b) - \sigma; b \}.
\end{equation}

Identities (12) and (13) together with (14) suggest that, with $\sigma$ sufficiently small, impact of the approximation is asymptotically negligible such that the consistency and weak convergence results in Theorems 3.1 and 3.5 hold.

\textbf{Corollary 4.1.} Set the finite constant $w > 1$ in optimization problem (11). If $\sigma = o(1)$, then $\hat{\beta}_\sigma$ is a special case of $\hat{\beta}$ defined in Theorem 3.1. Furthermore, when the first biomarker is the anchor and $\sigma = o(n^{-2/3})$, $\hat{\beta}_\sigma$ and $(1, \hat{\eta}_\sigma^\top)^\top$ differ by a scale with probability tending to 1 and $\hat{\eta}_\sigma$ is a special case of $\eta$ defined in Theorem 3.5.

Now, focus on the computation with problem (11). The objective function and the left-hand side of the first constraint are sums of convex and concave components, whereas the left-hand side of the second constraint is convex. We can then adopt the concave-convex procedure (Yuille and Rangarajan, 2003) as extended by Lipp and Boyd (2016) to accommodate constraints, which is the core of Algorithm 1. At each step, the two concave components are replaced by their tangent planes at the current variable value resulting in a convex optimization problem. With our application, the convex optimization is actually a linear program as given by (15), upon adopting slack variables, thanks partly to the adopted $\ell_1$ norm on $b$. The current variable value is then updated with the optimizer, which always satisfies the original constraints and improves the original objective function. Such steps can be repeated until the original objective function could not be further improved, and the algorithmic convergence in the objective function of (11) is guaranteed.

Nevertheless, one issue with the implementation is how to choose $\sigma$. Corollary 4.1 suggests a small value, which, however, could lead to a local optimizer near the initial value. Indeed, multiple local optimizers may exist and the concave-convex procedure does not guarantee to reach the global one (cf. Lipp and Boyd, 2016). In Algorithm 1, rather than a single small value, a sequence of decreasing $\sigma$ values are taken. Thanks to the specific approximation adopted for the indicator function, the transition from one $\sigma$ value to the next is seamless since the variable value remains feasible, i.e., satisfying the constraints, as $\sigma$ decreases. With a larger $\sigma$, conceivably the objective function and the constraint are more smooth to have fewer local optima. As the $\sigma$ value decreases gradually, the optimizer might thus more likely approach the global one.

With shrinking $\sigma$ in Algorithm 1, the choice of constant $w > 1$ becomes less essential since $w$ is absorbed into $\sigma$ in identities (12) and (13). A convenient value, say, 2, suffices. Finally, an initial feasible value is required for $(b^\top, t)^\top$ with the starting $\sigma$ value. For that purpose, a working method such as logistic regression or a coarse grid search may be adopted.

As a note, this developed algorithm is not limited to the proposed method but rather is generally applicable for other optimization problems of a similar form. The maximum score estimator of Manski (1975, 1985) is one example.
Finally, the kernel smoothing-based estimator of Meisner et al. (2021) as implemented in the 
package does not appear to be readily available, we adapted our algorithm in Section 4 for its computation.

As software for the maximum score estimator of Manski (1975, 1985) did not appear to be readily available, we adopted our algorithm in Section 4 for its computation. Finally, the kernel smoothing-based estimator of Meisner et al. (2021) as implemented in the R package maxTPr with default parameters was assessed as well.

For all the following numerical studies, specificity at controlled 95% sensitivity was adopted as the performance metric, and so targeted by both the proposed empirical utility maximization and the kernel smoothing-based estimator of Meisner et al. (2021).

Algorithm 1: Pseudo code for optimization problem (11) with shrinking $\sigma$ value

**Input:** starting value of $\sigma$
- initial feasible value $(b_0^T, t_0)$ of $(b^T, t)^T$
- shrinking factor of $\sigma$

**Output:** optimizer for $(b^T, t)^T$

$i \leftarrow 0$.

repeat /* Concave-convex procedure with fixed $\sigma$ */

1. **Convexify.** Form

$$T_1(b, t; b_i, t_i) \equiv \frac{1}{\sigma}E[(b_i^T M_1 - t_i + \sigma)^- + w\|b_i\|_1]$$
$$+ \frac{1}{\sigma}E[(b - b_i)^T M_1 - t + t_i] I(b_i^T M_1 \leq t_i - \sigma)]$$
$$+ w \text{sign}(b_i)^T (b - b_i),$$

$$T_0(b, t; b_i, t_i) \equiv \frac{1}{\sigma}E[(b_i^T M_0 - t_i)^-]$$
$$+ \frac{1}{\sigma}E[(b - b_i)^T M_0 - t + t_i] I(b_i^T M_0 \leq t_i)].$$

2. **Solve.** Set the value of $(b_{i+1}, t_{i+1})^T$ to the solution of

$$\min_{b, t} \frac{1}{\sigma}E[b^T M_1 - t]^- - T_1(b, t; b_i, t_i)$$
subject to $\{\rho + \frac{1}{\sigma}E[(b - b_i)^T M_0 - t + \sigma)^-] - T_0(b, t; b_i, t_i) \leq 0$,

3. **Update iteration.** $i \leftarrow i + 1$.

until sufficiently small improvement in objective function of (11) with current $\sigma$.
Shrink $\sigma$ value by the given factor.

until sufficiently small improvement in objective function of (10).

5. **Numerical studies.** The proposed empirical utility maximization with Algorithm 1 has been implemented using a linear programming solver from R package Rmosek. In the studies reported below, $w$ was set to 2. The logistic regression coefficient served as the initial value; the results using a coarse grid search were similar. With the initial coefficient, the maximum gap among the cases and that among the controls, between adjacent order statistics of the combinations, were computed. The larger of the two was taken as the starting $\sigma$ value. A feasible initial value for the threshold was then calculated. Each subsequent $\sigma$ value was shrunk by a factor of 0.8.

Seven existing linear combination methods were included in our numerical studies for comparison. Linear discriminant analysis and logistic regression are both standard methods; the latter in particular is routinely adopted in biomarker research. As a semiparametric method, the monotonic density ratio model of Chen et al. (2016) is more general. For the support vector classifier, function svm in R package e1071 was used. To target the AUC, we adopted the smoothed empirical AUC maximizer of Lin et al. (2011) as implemented in R package aucm. As software for the maximum score estimator of Manski (1975, 1985) did not appear to be readily available, we adapted our algorithm in Section 4 for its computation. Finally, the kernel smoothing-based estimator of Meisner et al. (2021) as implemented in R package maxTPr with default parameters was assessed as well.

For all the following numerical studies, specificity at controlled 95% sensitivity was adopted as the performance metric, and so targeted by both the proposed empirical utility maximization and the kernel smoothing-based estimator of Meisner et al. (2021).
The computer code for the proposed empirical utility maximization and the maximum score estimator of Manski (1975, 1985) is available on the first author’s website (http://web1.sph.emory.edu/users/yhuang5).

5.1. Simulations. Across all set-ups, the control biomarkers were independent and identically distributed with the standard normal distribution, whereas the case biomarkers had different distributions from one set-up to another. Either three or six biomarkers were considered for combination. For the former, all the biomarkers were informative, i.e., having different case and control distributions. Then, in the latter case, three independent and non-informative biomarkers were additionally included; that is, each followed the same standard normal distribution for the cases as for the controls. Four different scenarios were constructed for the three informative biomarkers:

Scenario A. The case biomarkers were independent and normally distributed each with mean 0.9 and variance 1.

Scenario B. The case biomarkers were independent and normally distributed with the same mean of 0.8 but different variances, 0.5, 1, and 2.

Scenario C. The case distribution was jointly normal with the same mean of 1, variances of 0.5, 1, and 2, and pairwise correlation coefficient of 0.5.

Scenario D. The case biomarkers followed a mixture of independent normal distributions. With probability 2/3, they had means of 1.7, 1.7, and 0, and variances of 0.5, 2, and 1, respectively. Then, with probability 1/3, their means became 0, 0, and 1.7, and the variances had the same value of 1.

These scenarios were motivated from cancer detection applications. In Scenarios A, B, and C, all the three case biomarkers were elevated in comparison to their control counterparts. However, they might be independent or correlated, with the same or different variability. Scenarios D mimicked cancer heterogeneity as common in cancer biology, involving two subtypes. The first two biomarkers were only elevated in one subtype, whereas the last was so in the other. The assumptions for linear discriminant analysis, logistic regression, and the monotonic density ratio model hold under Scenario A, but not under Scenarios B, C, and D. At controlled 95% sensitivity, the optimal specificities with linear combinations were 0.466, 0.452, 0.444, and 0.442 under Scenarios A, B, C, and D, respectively. The case and control sizes, \( n_1 \) and \( n_0 \), were set equal with values from 100 to 500.

Results were obtained from 1000 simulations for each set-up. Table 1 shows summary statistics of performance deficiencies for the proposed empirical utility maximization along with the seven existing methods. The averaged performance deficiencies for a subset of these methods are also displayed in Figure 1. For all the methods under each scenario, performance deficiency decreased with larger sample size as expected, and also from 6 to 3 biomarkers after the elimination of non-informative ones. In Scenario A, all the estimated combination coefficients converge to the optimal one and so do their predictive performances. Not surprisingly, linear discriminant analysis and logistic regression performed the best. They were followed by monotonic density ratio model, support vector classifier, and empirical AUC maximizer. The maximum score estimator and the kernel smoothing-based estimation also had better performance than the empirical utility maximization, although the differences were fairly small. However, in Scenario B and more so in Scenario C, the empirical utility maximization had the best performance whereas linear discriminant analysis and logistic regression performed rather poorly and so did monotonic density ratio model, support vector classifier, and empirical AUC maximizer. Under Scenario D, the relative performance varied with sample size and the number of biomarkers; the empirical utility maximization became the best as the sample size increased. Overall, the proposed method showed a different performance profile, even in comparison with the kernel smoothing-based method. Across all
**Table 1**
Simulation results on performance deficiencies of specificity at 95% sensitivity

<table>
<thead>
<tr>
<th>n₁ = n₀</th>
<th>3 biomarkers</th>
<th>6 biomarkers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M D M D M D M D</td>
<td>M D M D M D M D</td>
</tr>
<tr>
<td>100</td>
<td>8 8 4 4 3 3 2 2 2 2</td>
<td>20 13 10 6 7 4 5 3 4 3</td>
</tr>
<tr>
<td>200</td>
<td>8 8 4 4 3 3 2 2 2 2</td>
<td>21 14 10 7 7 5 5 3 4 3</td>
</tr>
<tr>
<td>300</td>
<td>10 10 5 5 4 4 3 3 2 3</td>
<td>23 15 12 8 8 5 6 4 5 4</td>
</tr>
<tr>
<td>400</td>
<td>9 10 5 5 3 3 2 2 2 2</td>
<td>24 15 12 8 8 5 6 4 5 3</td>
</tr>
<tr>
<td>500</td>
<td>9 15 4 4 3 3 2 2 2 2</td>
<td>24 23 11 7 7 5 6 4 5 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>37 25 31 17 30 14 30 12 29 11</td>
</tr>
<tr>
<td>LR</td>
<td>37 26 31 17 30 14 29 12 28 11</td>
</tr>
<tr>
<td>MDR</td>
<td>41 29 34 21 33 17 31 15 30 13</td>
</tr>
<tr>
<td>SVC</td>
<td>38 28 31 19 30 16 29 13 28 12</td>
</tr>
<tr>
<td>AUC</td>
<td>38 30 32 18 31 15 30 13 29 11</td>
</tr>
<tr>
<td>MSE</td>
<td>43 41 35 31 30 25 27 22 26 20</td>
</tr>
<tr>
<td>KS</td>
<td>33 33 21 18 20 22 18 21 16 13</td>
</tr>
<tr>
<td>EUM</td>
<td>30 30 19 18 15 15 12 12 11 10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>94 36 92 26 91 20 90 19 91 17</td>
</tr>
<tr>
<td>LR</td>
<td>98 39 97 28 96 22 95 21 96 18</td>
</tr>
<tr>
<td>MDR</td>
<td>106 42 107 32 108 28 108 27 110 25</td>
</tr>
<tr>
<td>SVC</td>
<td>106 42 105 31 104 25 103 23 104 21</td>
</tr>
<tr>
<td>AUC</td>
<td>95 40 92 28 92 22 91 20 91 18</td>
</tr>
<tr>
<td>MSE</td>
<td>106 63 104 53 104 49 108 47 107 45</td>
</tr>
<tr>
<td>KS</td>
<td>61 57 44 45 39 39 30 35 26 33</td>
</tr>
<tr>
<td>EUM</td>
<td>37 39 22 24 18 19 15 16 13 13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>34 27 28 18 27 15 26 13 25 11</td>
</tr>
<tr>
<td>LR</td>
<td>33 25 26 17 25 14 25 12 24 11</td>
</tr>
<tr>
<td>MDR</td>
<td>37 33 31 23 30 21 29 18 28 16</td>
</tr>
<tr>
<td>SVC</td>
<td>40 33 30 22 28 17 28 14 26 13</td>
</tr>
<tr>
<td>AUC</td>
<td>33 26 25 17 24 14 23 12 22 10</td>
</tr>
<tr>
<td>MSE</td>
<td>54 52 41 39 37 35 35 32 33 29</td>
</tr>
<tr>
<td>KS</td>
<td>36 42 21 20 19 25 16 20 14 22</td>
</tr>
<tr>
<td>EUM</td>
<td>33 31 22 21 17 16 14 13 12 12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario D</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>8 8 4 4 3 3 2 2 2 2</td>
</tr>
<tr>
<td>LR</td>
<td>8 8 4 4 3 3 2 2 2 2</td>
</tr>
<tr>
<td>MDR</td>
<td>10 10 5 5 4 4 3 3 2 3</td>
</tr>
<tr>
<td>SVC</td>
<td>9 10 5 5 3 3 2 2 2 2</td>
</tr>
<tr>
<td>AUC</td>
<td>9 15 4 4 3 3 2 2 2 2</td>
</tr>
<tr>
<td>MSE</td>
<td>22 21 14 13 10 10</td>
</tr>
<tr>
<td>KS</td>
<td>20 26 13 14 9 17</td>
</tr>
<tr>
<td>EUM</td>
<td>30 30 19 18 15 15</td>
</tr>
</tbody>
</table>

M: empirical mean (×1000); D: empirical standard deviation (×1000).

LDA: linear discriminant analysis; LR: logistic regression; MDR: monotonic density ratio model; SVC: support vector classifier; AUC: smoothed empirical AUC maximizer; MSE: maximum score estimator; KS: kernel smoothing-based method of Meisner et al. (2021); EUM: proposed empirical utility maximization.

set-ups, on a logarithmic scale for both variables, the averaged performance deficiency and the sample size showed roughly a linear relationship with a slope of $-2/3$ for the empirical utility maximization. This is consistent with the asymptotic result of $n^{-2/3}$ convergence rate.

These simulations were performed on a 2020 MacBook Pro laptop with 2.3 GHz Intel Core i7. For the empirical utility maximization, the average CPU time for a single dataset ranged from 0.82 seconds in a case of $n_1 = n_0 = 100$ and 3 biomarkers to 4.43 seconds with $n_1 = n_0 = 500$ and 6 biomarkers.
FIG 1. Simulation results on linear combination via the proposed empirical utility maximization (●), in comparison with logistic regression (○), smoothed empirical AUC maximizer (∆), maximum score estimator (+), and kernel smoothing-based method of Meisner et al. (2021) (×). The least-squares fitting lines of −2/3 slope are shown for the empirical utility maximization.

5.2. Application to prostate cancer detection. The proposed methodology was motivated by prostate cancer research, to improve the detection of aggressive cancer, i.e., Gleason score ≥ 7, using non-invasive biomarkers among men undergoing their first-time biopsy. Among a limited number of commercially available test assays, prostate health index (phi) is an FDA-approved blood test analysis by combining three forms of prostate-specific antigen (PSA) from serum, total PSA (tPSA), free PSA (fPSA), and isoform [-2]proPSA (p2PSA):

\[
\text{phi} = \frac{\text{p2PSA}}{\text{fPSA}} \times \sqrt{\text{tPSA}},
\]

which is a proprietary calculation developed by Beckman Coulter Inc. The test has been evaluated and adopted to distinguish aggressive cancer from indolent or no cancer (Catalona et al., 2011). However, it was unclear whether the combination of the three PSA forms could be improved to achieve better specificity at a controlled high sensitivity level, in particular, 95%. To address it, we analyzed 156 cases and 358 controls, i.e., with and without aggressive prostate cancer, respectively, per pathology testing on prostate biopsies, enrolled in academic urology groups (Sanda et al., 2017). Serum specimens of these participants, obtained prior to biopsy, were assayed for phi.

Note that phi is a linear combination of logarithmic transformed tPSA, fPSA, and p2PSA. We applied our proposed method and the existing ones considered in the earlier simulations to this data set, as reported in Table 2. The combination coefficient of the empirical utility maximization appeared to deviate considerably from those of the existing methods, and


### Table 2

Analysis results of the prostate cancer study

<table>
<thead>
<tr>
<th></th>
<th>log(PSA)</th>
<th>log(PSA)</th>
<th>log(p2PSA)</th>
<th>empirical</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>phi</td>
<td>0.200</td>
<td>-0.400</td>
<td>0.400</td>
<td>0.246</td>
<td>–</td>
</tr>
<tr>
<td>LDA</td>
<td>0.275</td>
<td>-0.441</td>
<td>0.283</td>
<td>0.277</td>
<td>0.296</td>
</tr>
<tr>
<td>LR</td>
<td>0.266</td>
<td>-0.434</td>
<td>0.301</td>
<td>0.293</td>
<td>0.285</td>
</tr>
<tr>
<td>MDR</td>
<td>0.257</td>
<td>-0.425</td>
<td>0.319</td>
<td>0.268</td>
<td>0.285</td>
</tr>
<tr>
<td>SVC</td>
<td>0.270</td>
<td>-0.432</td>
<td>0.299</td>
<td>0.293</td>
<td>0.290</td>
</tr>
<tr>
<td>AUC</td>
<td>0.257</td>
<td>-0.439</td>
<td>0.304</td>
<td>0.251</td>
<td>0.277</td>
</tr>
<tr>
<td>MSE</td>
<td>0.272</td>
<td>-0.424</td>
<td>0.305</td>
<td>0.288</td>
<td>0.276</td>
</tr>
<tr>
<td>KS</td>
<td>0.350</td>
<td>-0.429</td>
<td>0.221</td>
<td>0.299</td>
<td>0.282</td>
</tr>
<tr>
<td>EUM</td>
<td>0.427</td>
<td>-0.445</td>
<td>0.128</td>
<td>0.369</td>
<td>0.297</td>
</tr>
</tbody>
</table>

LDA: linear discriminant analysis; LR: logistic regression; MDR: monotonic density ratio model; SVC: support vector classifier; AUC: smoothed empirical AUC maximizer; MSE: maximum score estimator; KS: kernel smoothing-based method of Meisner et al. (2021); EUM: proposed empirical utility maximization.

All combination coefficients are scaled to have unity $\ell_1$ norm.

Performance: specificity at 95% sensitivity; CV: median of three-fold cross-validation estimates from 100 random splits.

---

even more so from that of phi. Furthermore, the empirical utility maximization had a substantially better empirical estimate of specificity at 95% sensitivity. However, except for phi, all these methods had their combinations trained in the same data set and thus their empirical performance estimates might not be taken as unbiased for predictive performances. In particular, the empirical performance estimate of the empirical utility maximization method tends to over-estimate, and possibly so does the kernel smoothing-based method. However, for other combination methods which target different metrics, their empirical performances do not necessarily over-estimate and could actually under-estimate. As an attempt to address such biases, three-fold cross-validation was performed. This specific fold choice was driven by the fact that the validation subset could not be made too small due to the need of threshold estimation. With the cross-validation estimates, not surprisingly, the edge of the combination from the empirical utility maximization shrunk considerably. Yet the improvement over phi still appeared clinically meaningful, although it became only marginal over some of the other combinations. Nevertheless, the cross-validation results correspond to learning from a considerably smaller sample size. Thus, the difference in the cross-validation between the proposed method and those targeting different metrics might be conservative, since the former would approach the ideal performance as sample size increases. An independent validation study should provide a more definitive assessment.

#### 6. Discussion

We have developed a linear biomarker combination method that empirically maximizes clinical utility of the intended medical test. The estimated combination coefficient and predictive performance have been rigorously investigated with their limiting distributions established. The proposed empirical utility maximization is shown to be more robust in comparison with several common linear combination methods with respect to our performance metric of interest. Nevertheless, several topics warrant further investigation.

First of all, predictive performance estimation with the training data is of great value, for identification and selection of promising combinations to be validated in future studies.
Unfortunately, as discussed in Section 5.2, the apparent empirical estimate tends to over-estimate whereas standard cross-validation might be conservative. The asymptotic theory may need further development to guide this pursuit of more reliable estimation.

Second, optimal biomarker combination without restriction to a specific class is the ultimate goal. Indeed, linear combination may not be effective with, for example, heterogeneous diseases, where various biomarkers are discriminative for certain subtypes but not others. As indicated in Section 1, the proposed linear combination method may accommodate nonlinear combinations of biomarkers via the linear basis expansion technique. However, the number of basis functions under consideration can be large or even infinite. Selection and regularization methods are thus needed; see Hastie, Tibshirani and Friedman (2009, section 5). They indicated in Section 1, the proposed linear combination method may accommodate nonlinear estimate whereas standard cross-validation might be conservative. The asymptotic theory may need further development to guide this pursuit of more reliable estimation.

Finally, a generalization to high-dimensional biomarkers is also being explored. High throughput technology is becoming increasingly available. Combination of such biomarkers also calls for selection and regularization methods. Further efficiency improvement in computation might be critical as well.

APPENDIX A: PROOFS OF RESULTS IN SECTION 3.1

Proof of Theorem 3.1. We first show the existence of a maximizer \( \beta \) of \( \phi(\mathbf{b}) \). Consider a fixed \( \mathbf{b} \) and an arbitrary \( \mathbf{b}_* \) such that \( \|\mathbf{b}\|_1 = \|\mathbf{b}_*\|_1 = 1 \). For any \( \varepsilon > 0 \), Condition 2 implies

\[
\Pr\{\mathbf{b}^\top \mathbf{M}_0 \leq \tau(\mathbf{b}) + \varepsilon/2\} > \rho \geq \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 < \tau(\mathbf{b}_*)\}.
\]

Thus, there exists a constant \( c_1 > 0 \), independent of \( \mathbf{b}_* \), such that \( \Pr(\|\mathbf{M}_0\|_\infty > c_1) \) is sufficiently small to satisfy

\[
(16) \quad \Pr\{\mathbf{b}^\top \mathbf{M}_0 \leq \tau(\mathbf{b}) + \varepsilon/2\} \geq \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 < \tau(\mathbf{b}_*)\} + \Pr(\|\mathbf{M}_0\|_\infty > c_1).
\]

When \( \|\mathbf{M}_0\|_\infty \leq c_1 \), we can have \( \|\mathbf{b} - \mathbf{b}_*\|^\top \mathbf{M}_0 \leq \varepsilon/2 \) so long as \( \mathbf{b}_* \) is sufficiently close to \( \mathbf{b} \). With such a \( \mathbf{b}_* \),

\[
(17) \quad \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 < \tau(\mathbf{b}_*)\} = \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 < \tau(\mathbf{b}_*) + (\mathbf{b} - \mathbf{b}_*)^\top \mathbf{M}_0\}
\]

\[
\geq \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 < \tau(\mathbf{b}_*) + (\mathbf{b} - \mathbf{b}_*)^\top \mathbf{M}_0, \|\mathbf{M}_0\|_\infty \leq c_1\}
\]

\[
\geq \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 < \tau(\mathbf{b}_*) - \varepsilon/2, \|\mathbf{M}_0\|_\infty \leq c_1\}
\]

\[
\geq \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 < \tau(\mathbf{b}_*) - \varepsilon/2\} - \Pr(\|\mathbf{M}_0\|_\infty > c_1).
\]

Combining (16) and (17) gives

\[
\Pr\{\mathbf{b}^\top \mathbf{M}_0 \leq \tau(\mathbf{b}) + \varepsilon/2\} \geq \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 < \tau(\mathbf{b}_*) - \varepsilon/2\}
\]

and subsequently \( \tau(\mathbf{b}) - \tau(\mathbf{b}_*) \geq -\varepsilon \). On the other hand, the same arguments lead to

\[
\Pr\{\mathbf{b}^\top \mathbf{M}_0 \leq \tau(\mathbf{b}) - \varepsilon/2\} < \rho \leq \Pr\{\mathbf{b}_*^\top \mathbf{M}_0 \leq \tau(\mathbf{b}_*)\}
\]

and subsequently, for \( \mathbf{b}_* \) sufficiently close to \( \mathbf{b} \), \( \tau(\mathbf{b}) - \tau(\mathbf{b}_*) \leq \varepsilon \). Therefore, \( \tau(\mathbf{b}) \) is continuous. With this and Condition 3, so is \( \phi(\mathbf{b}) \) by similar arguments. The existence result then follows from the compactness of \( \{\mathbf{b} : \|\mathbf{b}\|_1 = 1, \mathbf{b} \in \mathbb{R}^k\} \), by the extreme value theorem (e.g., Rudin, 1976, theorem 4.16).

Since the class of functions \( \{I(\mathbf{b}^\top \mathbf{M}_t \leq t) : \mathbf{b} \in \mathbb{R}^k, t \in \mathbb{R}\} \) is Donsker (e.g., Kosorok, 2008, lemma 9.12) and thus Glivenko–Cantelli,

\[
\sup_{\mathbf{b}, t} |\hat{F}_d(t; \mathbf{b}) - F_d(t; \mathbf{b})| = o(1),
\]
almost surely. Then, extending theorem 2.3.1 of Serfling (1980) under Condition 2, we obtain
\[ \sup_b |\hat{\tau}(t; b) - \tau(t; b)| = o(1), \]
amost surely. Meanwhile, Condition 3 in conjunction with the continuity of \( \tau(t; b) \) implies that \( F_1(t; b) \) is continuous at \( t = \tau(t; b) \) uniformly in \( b \) such that \( \|b\|_1 = 1 \), by similar arguments given earlier, for the continuity of \( \tau(t; b) \), in a proof by contradiction. Subsequently,
\[ \sup_b |\hat{\phi}(b) - \phi(b)| = o(1), \]
amost surely. Consequently, \( \hat{\phi}(\beta) - \phi(\beta) = o(1) \) and furthermore \( \hat{\phi}(\hat{\beta}) - \phi(\beta) = o(1) \), almost surely. Then, it follows that \( \hat{\phi}(\hat{\beta}) = \phi(\hat{\beta}) = o(1) \) and \( \hat{\phi}(\beta) = \phi(\beta) = o(1) \), almost surely.

Finally, under Condition 4, standard arguments can be used to establish the strong convergence of \( \hat{\beta} \) to \( \beta \) in light of the compact parameter space, continuity of \( \phi(b) \), and uniform strong convergence of \( \hat{\phi}(b) \) as established above.

\[ \square \]

APPENDIX B: PROOFS OF RESULTS IN SECTIONS 3.2 AND 3.3

As shown in the proof of Theorem 3.1, \( \tau(h) \) is continuous at \( \eta \). Therefore, we can make the \( \varepsilon \) in Lemma 3.2, Lemma 3.3, and Theorem 3.4 sufficiently small so as to restrict \( h \) to a neighborhood of \( \eta \), as we do so implicitly, such that \( \{h^\top, \tau(h)\}^\top \) is in the neighborhoods of \( \{\eta^\top, \tau(\eta)\}^\top \) implicated in Conditions 6 and 7.

Write \( F_{d,1}^{-1}(t) \) as the conditional distribution of \( M_{d,1} \) given \( M_{d,-1} \), and \( F_{d,1}^{-1}(t) \) as its density if it exists. For \( d = 0 \) under Condition 6 or 7, \( f_{d,1}^{-1}(t - h^\top M_{d,-1}) \) exists at \( t = \tau(h) \). Accordingly, \( F_d(t; h) = EF_{d,1}^{-1}(t - h^\top M_{d,-1}) \) has a density \( f_d(t; h) = EF_{d,1}^{-1}(t - h^\top M_{d,-1}) \) at \( t = \tau(h) \), which is bounded away from 0 for \( d = 0 \) by part (ii) of Condition 6. Thus, \( F_0\{\tau(h); h\} = \rho \). Taking derivatives on both sides yields the gradient
\[ \nabla \tau(h) = \frac{E[\tau(h) - h^\top M_{0,-1}] M_{0,-1}]}{\rho \{\tau(h); h\}} \]
which is bounded under Conditions 5 and 6.

Furthermore, under Condition 7 and for \( d = 1, 0 \), the derivative \( f_{d,1}^{-1}(t) \) of \( f_{d,1}^{-1} \) exists and \( F_d(t; h) \) has a second derivative:
\[ f_d(t; h) = EF_{d,1}^{-1}(t - h^\top M_{d,-1}) \]
at \( t = \tau(h) \). Subsequently, \( f_d(\tau(h); h) \) is in light of the compact parameter space, continuity of \( \phi(b) \), and uniform strong convergence of \( \hat{\phi}(b) \) as established above. In the following proofs of Lemmas 3.2 and 3.3, but not elsewhere, we restrict to the special case of nonnegative \( M_{d,-1} \), \( d = 1, 0 \); the result for the general case follows subsequently by the argument in Remark 4 via the biomarker splitting technique.

**Proof of Lemma 3.2.** By part (ii) of Condition 6, \( f_0\{\tau(h); h\} \) is bounded away from 0. Set finite constant \( c_2 = 2(k-1)^{1/2} \sup_{\|h-\eta\|_\infty \leq \varepsilon} f_0\{\tau(h); h\}^{-1} \). Following the proof of Serfling (1980, lemma 2.5.4.B), one obtains that, for each fixed \( h \) such that \( \|h-\eta\|_\infty \leq \varepsilon \),
\[ \Pr\{\|\hat{\tau}(h) - \tau(h)\| > c_2 n_0^{-1/2} (\log n_0)^{1/2} \} \leq 2n_0^{-2(k-1)} \]
for sufficiently large \( n_0 \). On the other hand, both \( \tau(h) \) and \( \hat{\tau}(h) \) are non-decreasing in each component of \( h \) as all components of \( M_{0,-1} \) are nonnegative. We shall exploit this monotonicity property in the extension of the pointwise result (20). Write the floor function as \([\cdot]\).

Impose an equally-spaced grid on each component of \( h \) with mesh size \( \varepsilon / \lfloor n_0 / 2 \rfloor \), centered at the corresponding component of \( \eta \). Thus, each \( h \) such that \( \| h - \eta \|_\infty \leq \varepsilon \) can be bracketed by \( h_- \) and \( h_+ \) in the sense that, componentwise, \( h_- \leq h \leq h_+ \) with \( h_- \) and \( h_+ \) being the adjacent grid points. Then,

\[
\hat{\tau}(h_-) - \tau(h_+) \leq \hat{\tau}(h) - \tau(h) \leq \hat{\tau}(h_+) - \tau(h_-).
\]

Therefore, given bounded \( \nabla \tau(h) \) as in (18),

\[
\sup_{\| h - \eta \|_\infty \leq \varepsilon} |\hat{\tau}(h) - \tau(h)| \leq \sup_{h_-} |\hat{\tau}(h_1) - \tau(h_1)| + \sup_{h_-} \{\tau(h_+) - \tau(h_-)\}
\]

\[
= \sup_{h_-} |\hat{\tau}(h_1) - \tau(h_1)| + O(n^{-1/2}),
\]

where \( h_1 \) is a grid point componentwise, taking up to \((2\lfloor n_0 / 2 \rfloor + 1)^k - 1\) different values. For sufficiently large \( n_0 \),

\[
\Pr \left\{ \sup_{h_-} |\hat{\tau}(h_1) - \tau(h_1)| > c_2 n_0^{-1/2} (\log n_0)^{1/2} \right\} \leq 2 n_0^{2(k-1)} (2\lfloor n_0 / 2 \rfloor + 1)^k - 1,
\]

following (20). By the Borel–Cantelli lemma,

\[
\sup_{h_-} |\hat{\tau}(h_1) - \tau(h_1)| = O(n^{-1/2} (\log n)^{1/2}),
\]

almost surely. So is \( \sup_{\| h - \eta \|_\infty \leq \varepsilon} |\hat{\tau}(h) - \tau(h)| \) subsequently. \( \square \)

**Proof of Lemma 3.3.** Consider \( h_i, i = 1, 2, \) and \( u \) such that \( \| h_i - \eta \|_\infty \leq \varepsilon, \| h_1 - h_2 \|_\infty = O(n^{-3/4}) \), and \( |u| \leq c_0 n_0^{-1/2} (\log n_0)^{1/2} \). Write \( h_\wedge = \min(h_1, h_2) \) and \( h_\vee = \max(h_1, h_2) \), where the minimization and maximization apply componentwise. Compute the variance,

\[
\text{(21)} \quad \text{var} \left[ I \{M_{d,1} + h_2^\top M_{d,-1} \leq \tau(h_1) + u\} - I \{M_{d,1} + h_1^\top M_{d,-1} \leq \tau(h_2)\} \right]
\]

\[
\leq E |I \{M_{d,1} + h_2^\top M_{d,-1} \leq \tau(h_1) + u\} - I \{M_{d,1} + h_1^\top M_{d,-1} \leq \tau(h_2)\}|
\]

\[
\leq E |I \{M_{d,1} + h_1^\top M_{d,-1} \leq \tau(h_2)\} - I \{M_{d,1} + h_1^\top M_{d,-1} \leq \tau(h_2)\}|
\]

\[
+ E |I \{M_{d,1} + h_2^\top M_{d,-1} \leq \tau(h_1) + u\} - I \{M_{d,1} + h_2^\top M_{d,-1} \leq \tau(h_2)\}|
\]

\[
\leq (h_\vee - h_\wedge)^\top E \int_0^1 M_{d,-1} f_{d,1}^\tau(\tau(h_2) - \{h_\wedge + r(h_\vee - h_\wedge)\}^\top M_{d,-1}) \, dr
\]

\[
+ f_{d}(t^*; h_2)(\tau(h_1) + u - \tau(h_2)),
\]

for some \( t^* \) in the line segment between \( \tau(h_1) + u \) and \( \tau(h_2) \). Thus, it is clear that this variance is bounded by \( c_3 n_0^{-1/2} (\log n_0)^{1/2} \) for some constant \( c_3 > 0 \) with large \( n_0 \), in light of bounded density \( f_d(t; h) \) and conditional density \( f_{d,1}^\tau(t - h^\top M_{d,-1}) \) around \( \tau(h) \), bounded gradient of \( \tau(h) \), and integrability of \( M_{d,-1} \) by Condition 5. Then, by a Bernstein’s inequality (Serfling, 1980, lemma 2.5.4.A),

\[
\text{(22)} \quad \Pr \left[ |\hat{F}_d(\tau(h_1) + u; h_2) - \hat{F}_d(\tau(h_2); h_1) - F_d(\tau(h_1) + u; h_2) + F_d(\tau(h_2); h_1)| \geq 3c_3^{1/2} k^{1/2} n_0^{-3/4} (\log n_0)^{3/4} \right]
\]
with sufficiently large $n_d$.

Now, with all components of $M_{d-1}$ being nonnegative, both $F_d\{\tau(h_1) + u; h_2\}$ and $\hat{F}_d\{\tau(h_1) + u; h_2\}$ are non-decreasing in each component of $h_1$ and $u$, and are non-increasing in each component of $h_2$. Impose a grid on each component of $h$ to bracket $h$ by $h_-$ and $h_+$ in the same fashion as that in the proof of Lemma 3.2, however, with a different mesh size $\varepsilon/\lfloor n_d^{3/4} \rfloor$. Similarly, impose an equally-spaced grid on $u$ centered at 0 with mesh size $c_0 n_d^{-1/2} (\log n_d)^{1/2}/\lfloor n_d^{3/4} (\log n_d)^{1/2} \rfloor$, to bracket $u$ by the adjacent points $u_-$ and $u_+$ on the grid. Therefore,

$$\hat{F}_d\{\tau(h_-) + u_-; h_+\} \leq \hat{F}_d\{\tau(h_-) + u_+; h_-\} - \hat{F}_d\{\tau(h_-) + u_+; h_+\} - F_d\{\tau(h_-) + u_+; h_+\} - F_d\{\tau(h_-) + u_-; h_+\} + F_d\{\tau(h_-) + u_-; h_-\} + F_d\{\tau(h_-) + u_+; h_-\} - F_d\{\tau(h_-) + u_+; h_+\} + F_d\{\tau(h_-) + u_-; h_-\} + F_d\{\tau(h_-) + u_-; h_+\}$$

Let $\{h_1, h_2\}$ be either $\{h_-, h_+\}$ or $\{h_+, h_-\}$, and $\{u_1, u_2\}$ be either $\{u_-, u_+\}$ or $\{u_+, u_-\}$. Then,

$$\sup_{\|h - \eta\|_\infty \leq \varepsilon} \hat{F}_d\{\tau(h) + u; h\} - F_d\{\tau(h) + u; h\} = \hat{F}_d\{\tau(h_1) + u_1; h_2\} - F_d\{\tau(h_1) + u_1; h_2\} + F_d\{\tau(h_1) + u_2; h_1\} - F_d\{\tau(h_2) + u_2; h_1\}$$

where $\{h_1, h_2\}$ and $u_1$ take up to $2(2\lfloor n_d^{3/4} \rfloor)^{k-1} + 2\lfloor n_d^{3/4} (\log n_d)^{1/2} \rfloor + 1$ different values, respectively. Given (22), the probability that the first maximum above exceeds $3\varepsilon \lfloor n_d^{3/4} (\log n_d)^{3/4} \rfloor$ is no larger than $4n_d^{-d/2}(2\lfloor n_d^{3/4} \rfloor)^2(k-1)\lfloor n_d^{3/4} (\log n_d)^{1/2} \rfloor + 1$ for sufficiently large $n_d$. Then, the first maximum is $O\{n^{-3/4}(\log n)^{3/4}\}$ almost surely by the Borel–Cantelli lemma. On the other hand, the second maximum is $O(n^{-3/4})$ by arguments similar to those for (21). Together, they lead to the assertion.

PROOF OF THEOREM 3.4. By Lemmas 3.2 and 3.3, uniformly in $\|h - \eta\|_\infty \leq \varepsilon$,

$$\hat{F}_d\{\tau(h); h\} - F_d\{\tau(h); h\} = f_d\{\tau(h); h\} \cdot \{\tau(h) - \tau(h)\} + O(n^{-1} \log n),$$

almost surely. Meanwhile, under Condition 7, $F_d(t; h)$ has a bounded second partial derivative with respect to $t$ around $\tau(h)$. Then, a Taylor expansion along with Lemma 3.2 gives that, uniformly in $\|h - \eta\|_\infty \leq \varepsilon$,

$$F_d\{\tau(h); h\} - F_d\{\tau(h); h\} = f_d\{\tau(h); h\} \cdot \{\tau(h) - \tau(h)\} + O(n^{-1} \log n),$$

almost surely. Almost surely at most $k$ independent observations of $M_0$ may simultaneously satisfy $M_{0,1} + h^T M_{0,-1} = t$ for $(h^T, t)^T$ in a neighborhood of $(\eta^T, \tau(\eta))^T$ where the conditional density $f_{0,1|t}(t - h^T M_{0,-1})$ is bounded under part (i) of Condition 6. Then, with
the consistency of \( \hat{\tau}(h) \), given by Lemma 3.2, and the continuity of \( \tau(h) \), uniformly in \( \{h : \|h - \eta\|_\infty \leq \varepsilon\} \),

\[
|\hat{F}_0\{\hat{\tau}(h) ; h\} - \rho| \leq k/n_0,
\]

almost surely. Equation (7) then follows from equations (23), (24), and (25), as \( f_0\{\tau(h) ; h\} \) is bounded away from 0 under part (ii) of Condition 6.

Equations (23), (24), and (7) give rise to

\[
\hat{\phi}(h) = 1 - \hat{F}_1\{\tau(h) ; h\} + f_1\{\tau(h) ; h\} f_0\{\tau(h) ; h\}^{-1} |\hat{F}_0\{\tau(h) ; h\} - \rho|
+ O\{n^{-3/4}(\log n)^{3/4}\},
\]

almost surely. Following equation (7) and Lemma 3.2, uniformly in \( \{h : \|h - \eta\|_\infty \leq \varepsilon\} \),

\[
|\hat{F}_0\{\tau(h) ; h\} - \rho| = O\{n^{-1/2}(\log n)^{1/2}\},
\]

almost surely. Meanwhile, since the gradient \( \nabla f_d\{\tau(h) ; h\} \) given in (19) is bounded, \( f_1\{\tau(h) ; h\} f_0\{\tau(h) ; h\}^{-1} \) has a bounded gradient at \( \eta \). Thus, equation (8) follows.

\[\Box\]

**Proof of Theorem 3.5.** Our proof follows the general framework of Kim and Pollard (1990), although their main theorem does not apply to our problem. Rather than working with \( \hat{\phi}(h) \) directly, we tackle the problem through its approximation \( \overline{\phi}(h) \), which linearly combines two independent random components, \( \hat{F}_d\{\tau(h) ; h\}, d = 1, 0 \). Write \( \hat{F}_d\{\tau(h) ; h\} - \hat{F}_d\{\tau(\eta) ; \eta\} = \hat{E}_g(M_d; h) \), where

\[
g(m; h) = I\{m_1 + h^T m_{-1} \leq \tau(h)\} - I\{m_1 + \eta^T m_{-1} \leq \tau(\eta)\}
\]

with \( m \equiv (m_1, m_{-1}^T) \). Consider the class of such functions,

\[\mathcal{G}_\varepsilon = \{g(\cdot; h) : \|h - \eta\|_\infty \leq \varepsilon\},\]

with envelope

\[
G_\varepsilon(m) = \sup_{\|h - \eta\|_\infty \leq \varepsilon} |g(m; h)|
= I\left\{ \min_{\|h - \eta\|_\infty \leq \varepsilon} \tau(h) - h^T m_{-1} \leq m_1 \leq \max_{\|h - \eta\|_\infty \leq \varepsilon} \tau(h) - h^T m_{-1} \right\}.
\]

Since the subgraphs of functions in \( \mathcal{G}_\infty \) form a VC class and \( G_\varepsilon(m) \) is bounded, \( \mathcal{G}_\varepsilon \) is uniformly manageable (Kim and Pollard, 1990, section 3). When \( M_{d-1} \) is nonnegative, \( EG_\varepsilon(M_d)^2 = O(\varepsilon) \) as \( \varepsilon \downarrow 0 \) for both \( d = 1, 0 \), by arguments similar to those for (21). This result then holds generally, i.e., when \( M_{d-1} \) is not necessarily nonnegative, by the biomarker splitting technique described in Remark 4. The same approach can be used to establish \( E[g(M_d; h_1) - g(M_d; h_2)] = O(\|h_1 - h_2\|_\infty) \) for \( h_1 \) and \( h_2 \) near \( \eta \). Furthermore, since \( G_\varepsilon(m) \leq 1, E[G_\varepsilon(M_d)^2 I\{G_\varepsilon(M_d) > 1\}] = 0 \). With these properties of \( \mathcal{G}_\varepsilon \), we can now utilize the results in Kim and Pollard (1990).

We first establish \( \overline{\eta} - \eta = O_p(n^{-1/3}) \). By lemma 4.1 of Kim and Pollard (1990), there exists \( \varepsilon > 0 \) such that, for \( \|h - \eta\|_\infty \leq \varepsilon \), with each \( \delta > 0 \)

\[
|\hat{E}_g(M_d; h) - E_g(M_d; h)| \leq \delta \|h - \eta\|_\infty^2 + O_p(n^{-2/3}), \quad d = 1, 0.
\]

Give that the Hessian matrix of \( -E_g(M_1; h) \) at \( \eta \) is negative definite, we can choose \( \delta \) such that \( E_g(M_1; h) \geq (2 + \lambda) \delta \|h - \eta\|_\infty^2 \) in this neighborhood, provided that \( \varepsilon \) is sufficiently small. When \( \eta \) is in this neighborhood,

\[
\overline{\phi}(\eta) - \overline{\phi}(\eta) = -\hat{E}_g(M_1; \eta) + \lambda \hat{E}_g(M_0; \eta)
\leq -\delta \|\eta - \eta\|_\infty^2 + O_p(n^{-2/3}).
\]
On the other hand,
\[
\bar{\phi}(\eta) - \bar{\phi}(\eta) = \{\hat{\phi}(\eta) - \hat{\phi}(\eta)\} + \{\hat{\phi}(\eta) - \hat{\phi}(\eta)\} - \{\hat{\phi}(\eta) - \hat{\phi}(\eta)\} + \{\hat{\phi}(\eta) - \hat{\phi}(\eta)\}
\]
\[
\geq ||\eta - \eta||_{\infty}o_p(n^{-1/3}) + o_p(n^{-2/3}),
\]
where equation (8) or more specifically (9) has been used. Combining the two gives
\[
\delta ||\eta - \eta||_{\infty} \leq ||\eta - \eta||_{\infty}o_p(n^{-1/3}) + O_p(n^{-2/3}),
\]
which shows the $O_p(n^{-1/3})$ convergence rate of $\eta$.

Now, we work with the rescaled process
\[
n_1^{2/3}\{\bar{\phi}(\eta + n_1^{-1/3}a) - \bar{\phi}(\eta)\} = -n_1^{2/3}\hat{E}g(M_1;\eta + n_1^{-1/3}a) + \lambda n_1^{2/3}\hat{E}g(M_0;\eta + n_1^{-1/3}a),
\]
which is a linear combination of two independent processes. Weak convergence of each follows from the results of Kim and Pollard (1990, lemma 4.5, lemma 4.6, and theorem 4.7). Compute
\[
H = -\nabla^2 F_1(\tau(h); h)|_{h=\eta}
\]
\[
= -E[f'_{1,1}]^{-1}\{\eta - \eta^T M_{1,-1}\}\{\nabla \tau(\eta_1) - \tau(\eta_1)\} \otimes \tau_1^2
\]
\[
+ \lambda E[f'_{0,1}]^{-1}\{\eta - \eta^T M_{0,-1}\}\{\nabla \tau(\eta_1) - \tau(\eta_1)\} \otimes \tau_1^2,
\]
where $v^{\otimes 2} \equiv vv^T$. Let
\[
V_d(a_1, a_2) = \lim_{\alpha \to \infty} \alpha E\left\{\prod_{i=1}^{2} g(M_d; \eta + \alpha^{-1}a_i)\right\}
\]
\[
= E\left(\min_{i=1,2} \{\nabla \tau(\eta) - M_{d,-1}\}^T a_i \mid f_{d,1}^{-1}\{\tau(\eta) - \eta^T M_{d,-1}\}
\]
\[
\times I \left[\prod_{i=1}^{2} \{\nabla \tau(\eta) - M_{d,-1}\}^T a_i > 0\right]\right),
\]
and define
\[
V(a_1, a_2) = V_1(a_1, a_2) + \gamma \lambda^2 V_0(a_1, a_2).
\]

Then, $n_1^{2/3}\{\bar{\phi}(\eta + n_1^{-1/3}a) - \bar{\phi}(\eta)\}$ converges weakly to the corresponding combination of the two limiting processes, which is a Gaussian process $Z(a)$ with continuous sample paths, mean $a^T H a/2$, and covariance kernel $V$. So does $n_1^{2/3}\{\hat{\phi}(\eta + n_1^{-1/3}a) - \hat{\phi}(\eta)\}$, following equation (8) or (9).

As in Kim and Pollard (1990, example 6.2), one could show that the variance of $Z(a_1) - Z(a_2)$ is $V(a_1 - a_2, a_1 - a_2)$. Thus, the Gaussian process $Z$ has nondegenerate increments provided $V(a, a) \neq 0$ for $a \neq 0$. Following Kim and Pollard (1990, theorem 2.7), $n_1^{1/3}(\bar{\eta} - \eta)$ then converges in distribution to $\arg\max_a Z(a)$. Subsequently, the weak convergence of $\phi(\bar{\eta})$ follows with a Taylor expansion argument. □

APPENDIX C: PROOFS OF RESULTS IN SECTION 4

Proof of Corollary 4.1. First, consider $\sigma = o(1)$. With inequality (14), the same arguments in the proof of Theorem 3.1 lead to $\sup_{b} |\hat{\phi}_{s-1, \sigma}(b) - \phi(b)| = o(1)$ and subsequently $\sup_{b} |\hat{\phi}_{s-1, \sigma}(b) - \phi(b)| = o(1)$, almost surely. It then follows that $\hat{\phi}(\hat{\beta}_{\sigma}) - \phi(\hat{\beta}) = o(1)$ almost surely.
Now, switch to the circumstance of the first biomarker being the anchor and take \( \sigma = \sigma_p(n^{-2/3}) \). Since \( \hat{\beta}_1, \sigma \) converges in probability to \( \beta_1 > 0, \hat{\beta}_1, \sigma > 0 \) and thus identity (13) holds with probability tending to 1. Inequality (14) leads to
\[
\hat{\phi}(\hat{\sigma}^{-1}(h)) = 1 - \hat{F}_1(\hat{\tau}(h) + \sigma_p(n^{-2/3}); h) \\
= 1 - F_1(\tau(h); h) - F_1(\hat{\tau}(h) + \sigma_p(n^{-2/3}); h) + F_1(\tau(h); h) \\
+ O_p\bigl(n^{-3/4}(\log n)^{3/4}\bigr) \\
= \hat{\phi}(h) + o_p(n^{-2/3}),
\]
uniformly in \( \{h : \|h - \eta\|_\infty \leq \varepsilon\} \) from some \( \varepsilon > 0 \), where the second equality follows from Lemmas 3.2 and 3.3 and the last one from (23) after a Taylor expansion. Then, \( \hat{\phi}(\hat{\eta}_\sigma) - \hat{\phi}(\eta) = o_p(n^{-2/3}) \) is implied. \( \square \)

Acknowledgments. The authors thank the reviewers for their helpful comments and suggestions, in particular the Associate Editor for pointing out several mistakes in previous versions of the paper, and Dattatraya H. Patil for assistance in arranging the prostate cancer dataset analyzed in Section 5.2.

Funding. The authors were supported in part by NIH Grants R01 CA230268, U01 CA113913, and P30 AI050409.

REFERENCES


