HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2024 April 30.

Published in final edited form as:

Stat Med. 2023 April 30; 42(9): 1398–1411. doi:10.1002/sim.9677.

Conditional Concordance-assisted Learning Under Matched Case-control Design for Combining Biomarkers for Population Screening

Wen Li[1], **Ruosha Li**[2], **Qingxiang Yan**[3], **Ziding Feng**[4], **Jing Ning**[5]

[1]Division of Clinical and Translational Sciences, Department of Internal Medicine, The University of Texas McGovern Medical School at Houston, TX, USA

[2]Department of Biostatistics and Data Science, The University of Texas School of Public Health, TX, USA

[3]F. Hoffmann-La Roche Ltd., ON, Canada

[4]Department of Biostatistics, Fred Hutchinson Cancer Research Center, WA, USA

[5]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, TX, USA

Summary

Incorporating promising biomarkers into cancer screening practices for early-detection is increasingly appealing because of the unsatisfactory performance of current cancer screening strategies. The matched case-control design is commonly adopted in biomarker development studies to evaluate the discriminative power of biomarker candidates, with an intention to eliminate confounding effects. Data from matched case-control studies have been routinely analyzed by the conditional logistic regression, although the assumed logit link between biomarker combinations and disease risk may not always hold. We propose a conditional concordance-assisted learning method, which is distribution-free, for identifying an optimal combination of biomarkers to discriminate cases and controls. We are particularly interested in combinations with a clinically and practically meaningful specificity to prevent disease-free subjects from unnecessary and possibly intrusive diagnostic procedures, which is a top priority for cancer population screening. We establish asymptotic properties for the derived combination and confirm its favorable finite sample performance in simulations. We apply the proposed method to the prostate cancer data from the Carotene and Retinol Efficacy Trial (CARET).

Correspondence: Ruosha Li, Department of Biostatistics and Data Science, The University of Texas School of Public Health, Houston, TX 77030, USA; Jing Ning, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. ruosha.li@uth.tmc.edu; jning@mdanderson.org.

Conflict of interest

None declared.

SUPPORTING INFORMATION

Additional supporting information could be found online in the Supporting Information section.

SOFTWARE

Software in the form of R and CPP codes that implement the proposed method are available online at https://github.com/liwenmoi/Conditional-Concordance-assisted-Learning.

Keywords

Conditional Concordance-assisted Learning; Conditional Logistic Regression; Matched Case-control Studies; Sensitivity; Specificity

1 | INTRODUCTION

Effective population screening can reduce the burden of cancer by detecting it at an early stage. However, the performance of current cancer screening strategies is still far from satisfactory for many types of cancer. For example, the sensitivity of the surveillance for hepatocellular carcinoma (i.e., an ultrasound every six months in cirrhosis patients) ranges from 32% to 65%,[1] and tumors diagnosed often have reached an advanced stage without curative treatment options. Biomarkers and their combinations are promising tools to complement current cancer screenings.[2] For biomarker development in early detection of cancer, Pepe et al.[3] has provided a comprehensive set of guidelines and recommended five phases from preclinical exploratory studies to cancer control studies. Case-control studies are commonly used in Phase 2 studies (clinical assay and validation) to assess the discriminative performance of biomarker candidates or their combinations to distinguish between cases and controls. Particularly, a matched case-control study is a popular design to reduce the confounding issue, in which each of cases is matched to one or more controls based on variables believed to be confounders. There are several advantages of matching. First, it allows one to assess the discriminative accuracy of the biomarkers beyond the contribution of the matching variables.[4] Second, a balanced number of cases and controls across the levels of the matching variables can reduce the variance for parameter estimation compared to an unmatched study with the same sample size.[5]

Matched case-control data have been routinely analyzed by the conditional logistic regression in literature. The combination of biomarkers for case-control discrimination, termed as composite score, is often derived by maximizing the conditional likelihood, a global fit criterion, under the logistic regression model. To quantify the discriminative ability of the derived composite score, sensitivity and specificity are two commonly used measures. They are associated with a specific cutoff and can be estimated by the percentage of positive results (e.g., composite score > the cutoff) among cases, and the percentage of negative results (e.g., composite score ≤ the cutoff) among controls, respectively. The cutoff can be determined by certain criterion such as the Youden's index.[6]

Maintaining a high specificity has been noted as a top priority for cancer population screening. Considering the low incidence rates of cancer in the general population, only screening tools with high specificity can prevent a large number of subjects from undergoing unnecessary and costly medical procedures and experiencing substantial psychological stress.[7] Taking ovarian cancer screening as an example, a clinically acceptable specificity should exceed 98%.[8] However, the commonly used statistical methods for combining biomarkers are often not tailored to this priority in cancer population screening. For example, the derived composite score by the conditional logistic regression often does not have optimal discriminative performance within the aforementioned clinically meaningful

region of specificity. Several methods that target a global criterion other than the logistic regression likelihood were proposed for unmatched data.[9,10,11] Another limitation of the conditional logistic regression is the parametric link function that associates the composite score and the risk of developing cancer. In practice, researchers often have limited information regarding the mathematical form of the true link function. A misspecified parametric link function may lead to a biased estimation and a suboptimal composite score.[12] It is desirable to leave the link function unspecified in the pursuit of the optimal biomarker combination.

Some recent works constructed composite scores by maximizing a local criterion[13,14,15,16]. For data from case-control studies, Meisner et al.[17], Zhang et al.[18], and Wang et al.[19] proposed to directly maximize the sensitivity under the constraint that the specificity is greater than a pre-specified threshold. The methods perform well and produce higher sensitivity than conditional logistic regression for general classification tasks. Nonetheless, these methods are designed for non-matched data, and they cannot be directly applied to individually matched data. When adapted to the analysis of matched case-control studies, their corresponding objective functions only include information from cases and ignore the information from controls. For the range of specificity of interest in cancer population screening, the resulting score often could not precisely maintain the pre-specified specificity when applied to external validation data. Yan et al.[20] alternatively derived the optimal score by maximizing the partial area under the receiver operating characteristic (ROC) curve, which is a trade-off between the local and global criteria. In this work, we aim to develop an optimal biomarker combination and the associated decision rule, which simultaneously maximizes the discriminative power and maintains the specificity at a level that is practically acceptable for cancer population screening. We leave the link function un-specified and propose a conditional concordance-assisted objective function based on the discriminative power of the composite score.

The remainder of the article is organized as follows: in Section 2, we first review existing methods and then introduce the proposed method and the corresponding estimation procedure. In Section 3, we conduct extensive simulation studies to assess finite sample performance of the proposed method and compare it with the existing methods. In Section 4, we apply the proposed method to the Carotene and Retinol Efficacy Trial (CARET), a matched case-control study at the Fred Hutchinson Cancer Research Center. A brief discussion is provided in Section 5.

2 | METHOD

2.1 | Notation

Consider a matched case-control study that allows multiple cases or controls in each stratum. Denote Y_{ki} as the disease status for the ith subject in the kth stratum, $k = 1,\ldots, \boldsymbol{K}$, where $Y_{ki} = 1$ means diseased (i.e., case) and $Y_{ki} = 0$ means non-diseased (i.e., control). Let n_{kD} and $n_{k\bar{D}}$ be the number of cases and matched controls in stratum k, and denote $n_k = n_{kD} + n_{k\bar{D}}$ as the stratum total. Then

$n_D = \sum_{k=1}^{K} n_{kD}$
and
$n_{\bar{D}} = \sum_{k=1}^{K} n_{k\bar{D}}$
are, respectively, the total numbers of cases and controls. For notational simplicity, we arrange the subjects in each stratum such that the first n_{kD} subjects are cases. Let $\boldsymbol{X}_{ki}$ be the p-dimensional vector of biomarkers for the ith subject in the kth stratum. We define the composite score as a linear combination $\boldsymbol{\beta}^T\boldsymbol{X}_{ki}$, where $\boldsymbol{\beta}$ is a vector of coefficients with the same dimension of $\boldsymbol{X}_{ki}$.

For given $\boldsymbol{\beta}$ and cutoff c, the sensitivity and study-specific specificity of the composite score can be estimated as

$$\widehat{Se}(\boldsymbol{\beta}, c) = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_{kD}} I\left(\boldsymbol{\beta}^T\boldsymbol{X}_{ki} > c\right)}{\sum_{k=1}^{K} n_{kD}}, \tag{1}$$

and

$$\widehat{Sp_s}(\boldsymbol{\beta}, c) = \frac{\sum_{k=1}^{K}\sum_{i=n_{kD}+1}^{K} I\left(\boldsymbol{\beta}^T\boldsymbol{X}_{ki} \le c\right)}{\sum_{k=1}^{K} n_{k\bar{D}}}. \tag{2}$$

The controls are typically sampled based on the matching variables of their associated cases instead of a random sampling, and thus they may not represent the general control population. Denote the sampling probability as p_{ki}, $i \in \{n_{kD}+1, \ldots, n_k\}$, which is the probability of being included in the matched study given the disease status and the matching variables in the source population.[21] The sampling probability can be estimated empirically or via a logistic regression model. Given the estimated sampling probability of controls, the population-level specificity can be estimated as follows:[22,23]

$$\widehat{Sp}(\boldsymbol{\beta}, c) = \frac{\sum_{k=1}^{K}\sum_{i=n_{kD}+1}^{n_k} \hat{p}_{ki}^{-1} I\left(\boldsymbol{\beta}^T\boldsymbol{X}_{ki} \le c\right)}{\sum_{k=1}^{K}\sum_{i=n_{kD}+1}^{n_k} \hat{p}_{ki}^{-1}}. \tag{3}$$

2.2 | Existing Methods and their Extensions to Matched Data

Data from matched case-control studies are routinely analyzed using the conditional logistic regression. The associated conditional likelihood is conditional on
$\{(n_{kD}, n_{k\bar{D}})\}_{k=1}^{K}$
, the total number of cases and the total number of subjects within each stratum,

$$\mathscr{L}_{CL}(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\prod_{i=1}^{n_{kD}} \exp\left(\boldsymbol{\beta}^T\boldsymbol{X}_{ki}\right)}{\sum_{J \in \mathscr{C}_k^D} \prod_{j \in J} \exp\left(\boldsymbol{\beta}^T\boldsymbol{X}_{kj}\right)}, \tag{4}$$

where
c_k^D
are all subsets of size n_{kD} from $C_k = \{1, \ldots, n_k\}$. One advantage of the conditional likelihood is that it avoids the estimation of stratum-specific nuisance parameters due to

matching. Denote
$\hat{\boldsymbol{\beta}}_{CL}$
as the estimator of $\boldsymbol{\beta}$, which maximizes the conditional likelihood in (4). Since a high population-level specificity is a top priority in cancer population screening, the cutoff value is usually determined by
$\hat{c}_{CL} = \inf\left\{c: \widehat{Sp}\left(\hat{\boldsymbol{\beta}}_{CL}, c\right) \geq \tau\right\}$
, where τ denotes the pre-specified level of specificity. Then the corresponding sensitivity is
$\widehat{Se}_{CL} = \widehat{Se}\left(\hat{\boldsymbol{\beta}}_{CL}, \hat{c}_{CL}\right)$
.

The direct method[17] can be extended to the matched data by controlling the population-level specificity instead of the study-specific one. In particular, we can obtain the estimators denoted as
$\left(\hat{\boldsymbol{\beta}}_{D}, \hat{c}_{D}\right)$
by

$$\arg\max_{\boldsymbol{\beta}, c} \widehat{Se}(\boldsymbol{\beta}, c), \tag{5}$$

subject to
$\widehat{Sp}(\boldsymbol{\beta}, c) \geq \tau$
. The corresponding sensitivity can be subsequently calculated by
$\widehat{Se}_{D} = \widehat{Se}\left(\hat{\boldsymbol{\beta}}_{D}, \hat{c}_{D}\right)$
. As expected, the derived composite score by the direct method often facilitates a higher sensitivity than that by the conditional logistic regression, since the sensitivity itself is the objective function to be maximized in the direct method. On the other hand, the objective function of the direct method only includes information from the cases and ignores information from the controls.

Yan et al.[20] developed a composite score by maximizing the partial area under the ROC curve, termed the pAUC method. The method was originally designed for non-matched case-control studies. For fair comparisons in simulation studies, we generalize this method to accommodate the data from the matched case-control studies as follows. Given the individually matched data, the density function of the control group can be estimated by incorporating the sampling probabilities,

$$\hat{f}_{\bar{D}}(s) = \frac{1}{\sum_{k=1}^{K}\sum_{i=n_{kD}+1}^{n_k}\left\{h_{\bar{D}}/\hat{p}_{ki}\right\}} \sum_{k=1}^{K}\sum_{i=n_{kD}+1}^{n_k} \hat{p}_{ki}^{-1} K\left(\frac{s - \boldsymbol{\beta}^T \boldsymbol{X}_{ki}}{h_{\bar{D}}}\right), \tag{6}$$

where $K(.)$ is a kernel function, and
$h_{\bar{D}}$
is a pre-specified bandwidth. The density function of the case group can be estimated in a similar fashion but without the sampling probability, denoted as
$\hat{f}_{D}(\cdot)$
. Then the estimated survival functions of the two groups are

$\widehat{S}_D(s) = \int_s^\infty \widehat{f}_D(t)dt$

and

$\widehat{S}_{\bar{D}}(s) = \int_s^\infty \widehat{f}_{\bar{D}}(t)dt$

. The kernel smoothed ROC and pAUC are then given by

$\widehat{\mathrm{ROC}_K}(t) = \widehat{S}_D\left\{\widehat{S}_{\bar{D}}^{-1}(t)\right\}$

and

$\widehat{\mathrm{pAUC}}_K(t_0) = \int_0^{1-t_0} \widehat{ROC}_K(t)dt$

, respectively, where $(t_0, 1)$ is the pre-specified range of interest for specificity. By maximizing

$\widehat{\mathrm{pAUC}}_K(t_0)$

, we can derive the coefficient estimates,

$\widehat{\boldsymbol{\beta}}_{pauc}$

, and the composite score.

2. 3 | Proposed Conditional Concordance-Assisted Learning

To ensure the robustness of the optimal composite score, we minimize model assumptions and leave the link function unspecified. We propose a conditional concordance-assisted learning (CCAL) method for combining biomarkers. Based on the discriminative ability of the composite score within each stratum, we construct the following conditional concordance-assisted function (CCAF) tailored to the unique structure of matched case-control studies:

$$\mathscr{L}(\boldsymbol{\beta}, c) = \prod_{k=1}^{K} \frac{\prod_{i=1}^{n_{kD}} I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} > c) \prod_{i=n_{kD}+1}^{n_k} \{1 - I(\boldsymbol{\beta}^T \boldsymbol{X}_{ki} > c)\}}{\sum_{J \in C_k^D} \left[\prod_{j \in J} I(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} > c) \prod_{j \in \mathscr{C}_k \setminus J} \{1 - I(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} > c)\}\right]}. \quad (7)$$

Here, the concordance means that cases are more likely than controls to be classified as being screening positive by the score in a matched stratum. Given

$\{(n_{kD}, n_k\bar{D})\}_{k=1}^K$

, the denominator of (7) describes all possible classifications that render n_{kD} positives and

$n_{k\bar{D}}$

negatives, and the numerator is the correct classification, namely whether the classification at threshold c is concordant with the true case-control status. This CCAF has a close connection with the decision rule to determine screening positive or negative subjects, and to calculate the sensitivity and specificity, while avoiding the need to specify a parametric link function. For a case-control pair in stratum k, the numerator is maximized when $\boldsymbol{\beta}^T\boldsymbol{X}_{k1} > c$ and $\boldsymbol{\beta}^T\boldsymbol{X}_{k2} \le c$, where subscripts 1 and 2 denote case and control, respectively. Thus, CCAF is likely maximized at a $(\boldsymbol{\beta}, c)$ that renders $\boldsymbol{\beta}^T\boldsymbol{X}_{k1} > c$ and $\boldsymbol{\beta}^T\boldsymbol{X}_{k2} \le c$. Note that sensitivity is defined as the proportion of cases with $\boldsymbol{\beta}^T\boldsymbol{X}_{ki} > c$, and it is naturally closely related to the CCAF. The connection between CCAF and specificity can be similarly assessed.

Similar to the conditional likelihood, the CCAF characterizes the discriminative ability of the composite score within each case-control stratum, while eliminating the need to estimate stratum-specific nuisance parameters. The objective function of the direct method

in (5) often produces a higher sensitivity given a prespecified specificity in the training data compared to other methods, since the sensitivity itself is the objective function to be maximized. However, its objective functions only include information from cases and ignore the information from controls, and therefore its optimal performance may not be stable and often cannot be transferred to the validation data. Different from the direct method, CCAF unitizes the information from both cases and controls and targets to optimize the discrimination power of the resulting biomarker combination. The optimal discrimination performance by the fuller use of information from both cases and controls can better control for the specificity and sensitivity on the validation data, which are confirmed through simulation studies. To ensure identifiability, we set the Euclidean norm $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{i=1}^{p} \beta_i^2}$ to be one. When maximizing this objective function, we add a tiny constant ϵ to the product in the numerator and denominator to avoid zeros. The preliminary simulation studies (unreported) confirm that the estimation is insensitive to the value of ϵ.

In current screening practice, the score is usually for the work up for further clinical diagnosis and to prevent disease-free subjects from undergoing more expensive/invasive test procedures. Maintaining a high specificity is essential in screening for diseases with low incidences such as cancer, because a high specificity with a reasonable sensitivity can achieve a feasible positive predictive value for population screening.[7] Acknowledging this top priority in cancer screening, we search for the optimal score within the clinically meaningful region of sensitivity, and maximize the CCAF in (7) subject to the constraint of $\widehat{Sp}(\boldsymbol{\beta}, c) \geq \tau$. The threshold τ is pre-specified and should be tailored to the disease incidence and the target population. For example, a threshold of 80% might be reasonable in a study of high-risk subjects, but a much higher threshold (e.g., 95%) is usually required for general population screening.

We propose a stable and computationally efficient algorithm to maximize (7) under the constraint based on the profiling approach. For any given $\boldsymbol{\beta}$, we obtain an estimate of c, denoted as $\hat{c}(\boldsymbol{\beta})$, by finding the τth quantile of $\boldsymbol{\beta}^T\boldsymbol{X}$ among controls as $\inf\{c : W_n(\boldsymbol{\beta}, c) \geq 0\}$, where $W_n(\boldsymbol{\beta}, c) = \frac{1}{n_{\bar{D}}} \sum_{k=1}^{n_D} \sum_{j=n_{kD}+1}^{n_k} \left\{ \hat{p}_{kj}^{-1} I\left(\boldsymbol{\beta}^T \boldsymbol{X}_{kj} \leq c\right) - \tau \right\}$. We then plug $\hat{c}(\boldsymbol{\beta})$ into equation (7) and maximize the profiled conditional concordance function $\mathscr{L}\{\boldsymbol{\beta}, \hat{c}(\boldsymbol{\beta})\}$ with respect to $\boldsymbol{\beta}$. This approach offers simultaneous estimates for $\boldsymbol{\beta}$ and the associated cutoff $\hat{c}(\boldsymbol{\beta})$. Given these estimates by the CCAL, the sensitivity and specificity can be calculated by equations (1), (2), and (3).

The proposed objective function (i.e., CCAF) is not continuous with respect to the unknown parameters. With a small number of biomarkers, we can implement the Nelder-Mead method with multiple starting values to identify the global maximizers. However, this method may become computationally burdensome when there are many biomarkers. An alternative solution is to use a continuous kernel function,

$\int_{-\infty}^{\beta^T x_{ki}-c} K(u;h_n)du$

, to approximate the indicator function $\boldsymbol{I}(\boldsymbol{\beta}^T\boldsymbol{X}_{ki} > c)$, where $K(\cdot, h_n)$ is a symmetric kernel function and h_n is the bandwidth.[24,25,12] Accordingly, we have the following kernel-smoothed CCAF,

$$\mathscr{L}_{KS}(\boldsymbol{\beta}, c) = \prod_{k=1}^{K} \frac{\prod_{i=1}^{n_{kD}} \int_{-\infty}^{\beta^T x_{ki}-c} K(u;h_n)du \prod_{i=n_{kD}+1}^{n_k} \left\{1 - \int_{-\infty}^{\beta^T x_{ki}-c} K(u;h_n)du\right\}}{\sum_{J \in C_k^D} \left[\prod_{j \in J} \int_{-\infty}^{\beta^T x_{kj}-c} K(u;h_n)du \prod_{j \in C_k \setminus J} \left\{1 - \int_{-\infty}^{\beta^T x_{kj}-c} K(u;h_n)du\right\}\right]}. \quad (8)$$

Theoretically, many smooth and symmetric probability density functions can be used as the kernel function, and the standard normal distribution is a popular choice in practice. Then the estimation can be accomplished by existing programs such as the *Rsolnp* package in R.

Let

$(\hat{\boldsymbol{\beta}}, \hat{c}) \equiv \left\{\hat{\boldsymbol{\beta}}, \hat{c}(\hat{\boldsymbol{\beta}})\right\}$

denote the estimator by maximizing the CCAF in (7), we establish asymptotic properties of

$(\hat{\boldsymbol{\beta}}, \hat{c})$

, and the plug-in estimator

$\widehat{Se}(\hat{\boldsymbol{\beta}}, \hat{c})$

of the sensitivity. Denote the limiting values of these parameters by

$(\tilde{\boldsymbol{\beta}}, \tilde{c})$

and

$\widetilde{Se} = \widetilde{Se}(\tilde{\boldsymbol{\beta}}, \tilde{c})$

, whose formal definition is provided in Section 1 of Supplementary Materials. The main technical challenge is the discontinuity of $\mathscr{L}(\boldsymbol{\beta}, c)$ with respect to

$(\hat{\boldsymbol{\beta}}, \hat{c})$

due to the indicator functions in (7), and therefore standard methods requiring the smoothness and differentiability are not applicable. Under the mild regularity conditions given in Supplementary Materials, we apply the empirical process techniques to show that

$\lim_{n\to\infty} \| (\hat{\boldsymbol{\beta}}, \hat{c}) - (\tilde{\boldsymbol{\beta}}, \tilde{c}) \| \xrightarrow{p} 0$

, as well as the consistency of the associated sensitivity estimates.

3 | SIMULATION STUDIES

We conducted simulation studies to evaluate the finite sample performance of the proposed CCAL method and compared it to that of three methods reviewed in Section 2.2: the conditional logistic regression, the direct method, and the pAUC method.

3.1 | Simulation Settings

We considered four different scenarios for the performance evaluation, where the parametric assumption of conditional logistic regression was satisfied in Scenario 1, but not in other scenarios.

Scenario 1. We generated two independent biomarkers, X_1 and X_2, from the standard normal distribution. Two matching variables Z_1 and Z_2 were generated from the *Bernoulli*(0.3) and *Bernoulli*(0.1) independently. The matching group membership $\mathscr{S}$ was based on the values of Z_1 and Z_2: $\mathscr{S} = 1$ if $Z_1 = 0$ and $Z_2 = 0$; $\mathscr{S} = 2$ if $Z_1 = 1$ and $Z_2 = 0$; $S = 3$ if $Z_1 = 0$ and $Z_2 = 1$; and $\mathscr{S} = 4$ otherwise. The disease status followed a *Bernoulli* distribution with a diseased probability of $\text{logit}^{-1}\{(X_1 + 3X_2 + 0.5Z_1 + 4Z_2)/1.5 - 7\}$, where $\text{logit}(t) = \log\{t/(1-t)\}$.

Scenario 2. We generated two biomarkers (X_1 and X_2) and the matching variable Z_1 from a multivariate normal distribution conditional on the disease status. Among controls, X_1 followed $N(0, 3)$, and both X_2 and Z_1 followed $N(0, 1)$. They were pairwise correlated with a correlation coefficient of 0.3. Among cases, X_1, X_2, and Z_1 independently followed $N(3, 3)$, $N(3, 5)$, and $N(3, 5)$, respectively. Hence, both means and covariance matrices of the biomarkers and the matching variable depended on the disease status, and the covariance matrices were disproportional for cases and controls. The matching group membership was defined as $\mathscr{S} = I\{Z_1 \geq \Phi^{-1}(1/4)\} + I\{Z_1 \geq \Phi^{-1}(1/2)\} + I\{Z_1 \geq \Phi^{-1}(3/4)\} + 1$, where Φ is the standard normal cumulative distribution function.

Scenario 3. We used the same data generation scheme as that in Scenario 2, except that the correlation between biomarkers among cases was set to 0.9.

Scenario 4. We considered different correlation directions between controls and cases. Specifically, X_1, X_2, and Z_1 were negatively correlated with a correlation coefficient of −0.3 among controls, whereas they were positively correlated with a correlation coefficient of 0.3 among cases. Among controls, X_1 followed $N(0, 3)$, and X_2 and Z_1 followed $N(0, 1)$. Among cases, X_1 followed $N(0, 3)$, and X_2 and Z_1 followed $N(0, 5)$.

Under all four scenarios, we used 1:1 matching to construct the matched case-control data. To ensure locating of the global maximum of the proposed objective function, we used 20 sets of starting values around the coefficient estimates by the conditional logistic regression. The maximization converged fairly quickly for our method, and therefore the use of multiple starting points was not computationally intensive. For fair comparison, the same 20 sets of starting values were used for all four methods. We adopted the bootstrap method for variance estimation. In particular, we resampled the strata with replacement 200 times and calculated the bootstrap standard deviation.

The sample size

$n_D = n_{\bar{D}}$

varied from 50 to 400, and the pre-specified threshold of specificity τ varied from 0.70 to 0.98. When implementing the kernel-smoothing method, we adopted bandwidth $h_n = C_h(n_D)^{-1/3}$, where $(n_D)^{-1/3}$ is the optimal bandwidth recommended by Jones[24] and $C_h = 0.2$, 1, or

5. For each setting, we conducted 1000 simulation replicates and summarized the results. We calculated the sensitivities and specificities of the composite score by the aforementioned four methods using independent external validation data sets with a large sample size of 20,000, such that the variability due to the external data sets was negligible.[26,20] All specificities reported in the simulation studies were in the population-level. The specificity range of interest for the pAUC method was set to (0.7, 1), or $t_0 = 0.7$, following the simulation settings in Yan et al.[20]

3.2 | Simulation Results

Figures 1 and 2 show the average values and empirical standard errors (ESE) of estimated sensitivities (± ESE) and specificities (± ESE) at various prespecified specificities τ (0.70, 0.75, 0.85, 0.90, 0.95, and 0.98). Here, the composite scores and the cutoffs were estimated using the training data sets and then tested using the large validation data. To better differentiate the results of the four different methods, the error bars corresponding to different τ's were shifted slightly along the x-axis. The corresponding summary tables are presented in Tables S1–S4 in the Supplementary Materials.

Under Scenario 1, the logistic regression model is the underlying true model. When the sample size of the training data was small
($n_D = n_{\bar{D}} = 50$)
, all composite scores by the four methods could not precisely maintain the pre-specified specificity on the validation data sets, as shown in Figure 1(B). The direct method consistently resulted in the lowest specificities, and the performance of the remaining three methods were similar for most of the τ's considered. When the sample size increased to
$n_D = n_{\bar{D}} = 100$
, all methods except for the direct method were able to well maintain the pre-specified specificities, as seen in Figure 2(B). On the other hand, the direct method had slightly higher sensitivities compared with the other three methods.

Under Scenarios 2–4, the relationship between biomarkers and disease status cannot be captured by a simple parametric model such as the logistic regression model. As expected, the composite score by the conditional logistic regression performed poorly and had markedly lower sensitivities compared with other three methods. With
$\widehat{Se}$
denoting estimates of the sensitivity by the other three methods, the *relative percentage difference*, defined as (mean
$\widehat{Se}$
- mean
$\widehat{Se}_{CL}$
)/mean
$\widehat{Se}_{CL} \times 100\%$
, ranged from 32% to 124% at $\tau = 0.98$.

We also observed that the CCAL method produced the highest specificities when evaluated on the validation data. When the sample size of the training data was moderate or large (e.g.,

$n_D = n_{\bar{D}} \geq 100$
), the specificities from the CCAL method were close to the pre-specified level of τ, and even slightly higher than τ in some settings. Specifically, the difference between the average of estimated specificities and the prespecified τ was between −0.02 and 0.01. On the other hand, the direct method again failed to preserve the specificity. For example, under Scenario 3 with n_D of 100 and τ of 0.80, the difference between the average of estimated specificities and the prespecified level was as large as 0.06; see Figure 2(F). Similar to the direct method, the pAUC method could not maintain the specificity for most of the τ considered.

Interestingly, although the estimated specificities by the four methods had similar variance, the estimated sensitivities had quite different variation. The estimated sensitivities by the conditional logistic regression had the largest standard errors, while those by the pAUC method had the smallest standard errors. Such a statistical efficiency difference can be partially explained by the fact that the three semiparametric methods were based on the local or sub-global performance by focusing on a clinically-relevant region of the specificity; while the conditional logistic regression maximized the global performance including those clinically-irrelevant specificities, such as $\tau = 0.3$. On the other hand, the pAUC method used more data information around the pre-specified specificity and then produced more stable estimates.

The estimated sensitivities on the training data are summarized in Table 1 and Table S5 in Supplementary Materials. The ESEs and the average of the estimated standard errors (ASEs) by the bootstrap method agreed well, indicating the bootstrap method can accurately capture the variability of the proposed method.

We also implemented the CCAL method by maximizing the kernel-smoothed CCAF in (8), and we summarized the results in Table S6 in Supplementary Materials. The results were similar to those by the original CCAF, suggesting the kernel-smoothed method is a reasonable alternative. We compared the results by using the three different values of C_h, and found that the kernel-smoothed method was quite robust to the choice of the bandwidth in the settings considered. Besides the above simulation studies that focused on 1:1 matching, we also conducted simulations using 1:3 matched under Scenario 2 (see Table S7 and Figure S1 in Supplementary Materials). All methods had improvements in preserving the specificities compared to results with 1:1 matching, especially under the setting with a small sample size (e.g., $n_D = 50$). The performance comparisons among these methods had similar patterns to those reported with 1:1 matching.

Additional simulation studies were conducted to compare the performance of the proposed CCAF to that of the unconditional methods, including the conventional logistic regression with the full likelihood and an unconditional objective function $\mathcal{L}_{full}(\boldsymbol{\beta}, c)$ defined in the Supplementary Materials. As seen in Table S8 in the Supplementary Materials, even though the data was generated from a logistic regression model, the conventional logistic regression with the full likelihood had lower specificities and lower sensitivities compared to other methods. The performance of biomarker combinations by $\mathcal{L}_{full}(\boldsymbol{\beta}, c)$ was also worse than that of the proposed CCAF, although better than the full likelihood. Detailed simulation settings and results are presented in Section 2.1 of the Supplementary Materials.

Last, we evaluated the Youden Index of the different methods and summarized results in Tables S9–S12 of the Supplementary Materials. Under Scenario 1, there was no one method that uniformly outperformed others when the evaluation metric puts equal importance to the sensitivity and specificity. Under Scenarios 2–4, the proposed method showed a better discrimination capacity in terms of Youden Index than the conditional logistic regression.

4 | APPLICATION

We illustrate the proposed CCAL method for disease status discrimination on a prostate cancer data set in CARET, a randomized trial that enrolled 18,314 subjects at high risk for prostate cancer to evaluate the efficacy of the combination of beta-carotene and retinol on reducing prostate cancer risk. During the intervention phase of CARET, blood samples were collected and stored, providing invaluable resources for future research.

Within CARET, a matched case-control study was conducted, including 71 prostate cancer cases diagnosed between 1988 and 1995, and 71 controls matched by age and number of blood samples.[27] Two biomarkers for prostate cancer, the total prostate specific antigen (tPSA), and the free prostate-specific antigen (fPSA), were measured from the stored blood samples of the subjects in the data set. The details of the study are given in Etzioni et al.[27] The analytic data set included 68 matched pairs of cases and controls from this existing matched case-control study due to missing information. Our goal was to compose a risk score using the biomarkers to distinguish cases from controls under a matched study design. We performed bootstrap validation with a bootstrap sample size of 10,000.[28] Due to the small sample size, we only focused on τ ranging from 0.70 to 0.95. As shown in Figure 3 (blue squares), the CCAL method steadily kept the specificity at or above the pre-specified threshold.

Since the sampling probabilities of the controls in the prostate cancer data were unavailable, the estimated cutoff, sensitivity, and specificity are study-specific, and as a result cannot be generalized to the general population directly. To control the population-level specificity, one solution is to combine the current matched case-control data with the Census data. However, the population from the Census data differs systematically from the at-risk screening population, and thus it is not an ideal source for this study. Instead, we can borrow information from the intervention arm of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial by comparing the age distribution of the controls in the prostate cancer data and the age distribution of the participants in the PLCO trial. Sampling probabilities were calculated using an approach similar to the propensity score method. The validation results by controlling the population-level specificity are also summarized in Figure 3 (red circles). The performance of the different methods on the prostate cancer data in CARET is reported in Table S13 in the Supplementary Materials. In general, the proposed method outperformed the conditional logistic regression method. For example, when controlling the specificity at 95%, the proposed method could identify 72% of cases, while only 63% could be identified by the conditional logistic regression method. The proposed method also showed advantages over the pAUC method in terms of optimizing sensitivity. The proposed and direct methods had almost identical performance for this particular data.

To provide more insight into the performance of the optimal score provided by the proposed method on the prostate cancer screening, we further examined the positive predictive value (PPV), a key characteristic of a screening test. PPV is defined as the probability that a subject has the condition given that the subject's test is positive subject is test positive, and it depends on the characteristics of the test, as well as the population of interest. We calculated PPV based on the estimated population-level results and the prevalence of prostate cancer among US men aged 50 and up, which is 5.92%.[29] When restricting specificity to be at least 0.95, the PPV is 51.7%, meaning that around half of the men who had a positive test result detected by the proposed optimal score actually had prostate cancer. This is better than the PPV of the prostate specific antigen-based screening test used in several cohort studies, which is approximately 30%.[30]

5 | DISCUSSION

In this paper, we proposed an alternative semiparametric method to the conditional logistic regression given the data from matched case-control studies. We developed a CCAL method to avoid the need to estimate stratum-specific parameters. In the meanwhile, instead of using parametric link functions as in the conditional logistic regression, we directly used the decision rule on the construction of the CCAF. We maximized the proposed CCAF with a constraint of achieving a clinically acceptable specificity, based on the general guidance in cancer population screening practice. Different from the objective function of the direct method, the CCAF used information from both cases and controls, and it was shown to be advantageous to maintain the pre-specified specificity in the independent validation data.

Maintaining specificity is a pre-requisite for a good screening tool, since even tiny loss in specificity has severe consequences. For instance, considering the low incidence of liver cancer, each 1% drop in specificity of screening results in 1000 more subjects getting false positive results, experiencing psychological trauma, and even going through biopsy for diagnosis in a population screening program of 100,000 subjects.[31] Thus, being able to keep specificity on external validation data makes the proposed method more appealing than other existing methods in population screening. Although the focus of this paper is individually matched data, the proposed method can straightforwardly be extended to studies that use frequency matching (e.g., case and control groups have similar proportion of smokers, females in a lung cancer study) by post-hoc forming strata.

Of note, we maximized the proposed CCAF by using 20 different sets of initial values to minimize the possibility that the algorithm converged to a local maximum of the objective function depending on the starting values. Even though we applied multiple starting values, the computation burden was not heavy. For example, in a 100-run simulation with a sample size of 400 under Scenario 1, the CPU time of a desktop with 3.30GHz CPU was 0.86 minutes. In the presence of large number of risk factors, we can then use the kernel-smoothed method, which has satisfactory performance as shown in our simulation studies.

Although the true optimal score may not be a linear combination of the biomarkers, we only consider and identify the optimal score within the class of linear combinations throughout

this paper. One advantage of such combinations is the computational simplicity and ease of interpretation/communication with clinicians compared to nonlinear functions. The method development on how to implicitly detect an optimal combination with a potentially complex form and simultaneously maintain a pre-specified specificity is beyond the scope of this paper, but it is worthy of future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was partially supported by the National Institutes of Health (NIH) grants R01CA269696 (JN), U24CA230144(JN, ZF) and R01DK117209(RL), and CPRIT grant RP200633 (JN). We acknowledge the support provided by the Biostatistics/Epidemiology/ Research Design (BERD) component of the Center for Clinical and Translational Sciences (CCTS) for this project that is currently funded through a grant (UL1TR003167), funded by the National Center for Advancing Translational Sciences (NCATS), awarded to the University of Texas Health Science Center at Houston. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCATS. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the results reported in this paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://research.fredhutch.org/diagnostic-biomarkers-center/en/datasets.html.

References

1. Singal AG, Conjeevaram HS, Volk ML, et al. Effectiveness of hepatocellular carcinoma surveillance in patients with cirrhosis. Cancer Epidemiology and Prevention Biomarkers 2012.
2. Schutte K, Schulz C, Link A, Malfertheiner P Current biomarkers for hepatocellular carcinoma: Surveillance, diagnosis and prediction of prognosis. World journal of hepatology 2015; 7(2): 139. [PubMed: 25729470]
3. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. Journal of the National Cancer Institute 2001; 93(14): 1054–1061. [PubMed: 11459866]
4. Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. Biometrics 2008; 64(1): 1–9. [PubMed: 17501939]
5. Rose S, Laan V. d MJ. Why match? Investigating matched case-control study designs with causal effect estimation. The international journal of biostatistics 2009; 5(1).
6. Unal I. Defining an optimal cut-point value in ROC analysis: an alternative approach. Computational and mathematical methods in medicine 2017; 2017.
7. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. Journal of the National Cancer Institute 2008; 100(20): 1432–1438. [PubMed: 18840817]
8. Skates SJ, Horick N, Yu Y, et al. Preoperative sensitivity and specificity for early-stage ovarian cancer when combining cancer antigen CA-125II, CA 15–3, CA 72–4, and macrophage colony-stimulating factor using mixtures of multivariate normal distributions. Journal of clinical oncology 2004; 22(20): 4059–4066. [PubMed: 15381683]
9. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. Journal of the American Statistical Association 1993; 88(424): 1350–1355.
10. Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. Biometrics 2006; 62(1): 221–229. [PubMed: 16542249]

11. Yin J, Tian L. Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. Statistics in medicine 2014; 33(8): 1426–1440. [PubMed: 24311111]
12. Shen W, Ning J, Yuan Y, Lok AS, Feng Z. Model-free scoring system for risk prediction with application to hepatocellular carcinoma study. Biometrics 2018; 74(1): 239–248. [PubMed: 28742219]
13. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. Biostatistics 2000; 1(2): 123–140. [PubMed: 12933515]
14. Wang Z, Chang YCI. Marker selection via maximizing the partial area under the ROC curve of linear risk scores. Biostatistics 2011; 12(2): 369–385. [PubMed: 20729218]
15. Hsu MJ, Hsueh HM. The linear combinations of biomarkers which maximize the partial area under the ROC curves. Computational Statistics 2013; 28(2): 647–666.
16. Yu W, Park T. Two simple algorithms on linear combination of multiple biomarkers to maximize partial area under the ROC curve. Computational Statistics & Data Analysis 2015; 88: 15–27.
17. Meisner A, Carone M, Pepe MS, Kerr KF. Combining biomarkers by maximizing the true positive rate for a fixed false positive rate. Biometrical Journal 2021; 63(6): 1223–1240. [PubMed: 33871887]
18. Zhang Z, Lu Y, Tian L. On feature ensemble optimizing the sensitivity and partial ROC curve. Statistica Sinica 2019; 29: 1395–1418.
19. Wang Y, Zhao YQ, Zheng Y. Learning-based biomarker-assisted rules for optimized clinical benefit under a risk constraint. Biometrics 2020; 76(3): 853–862. [PubMed: 31833561]
20. Yan Q, Bantis LE, Stanford JL, Feng Z. Combining multiple biomarkers linearly to maximize the partial area under the ROC curve. Statistics in medicine 2018; 37(4): 627–642. [PubMed: 29082535]
21. Xu H, Qian J, Paynter NP, et al. Estimating the receiver operating characteristic curve in matched case control studies. Statistics in medicine 2019; 38(3): 437–451. [PubMed: 30467878]
22. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association 1952; 47(260): 663–685.
23. Huang Y. Evaluating and comparing biomarkers with respect to the area under the receiver operating characteristics curve in two-phase case-control studies. Biostatistics 2016; 17(3): 499–522. [PubMed: 26883772]
24. Jones M. The performance of kernel density functions in kernel distribution function estimation. Statistics & Probability Letters 1990; 9(2): 129–132.
25. Zeng D, Lin DY. Efficient Estimation for the Accelerated Failure Time Model. Journal of the American Statistical Association 2007; 102: 1387–1396.
26. Payne R, Yang M, Zheng Y, Jensen MK, Cai T. Robust risk prediction with biomarkers under two-phase stratified cohort design. Biometrics 2016; 72(4): 1037–1045. [PubMed: 27037494]
27. Etzioni R, Pepe M, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. Medical Decision Making 1999; 19(3): 242–251. [PubMed: 10424831]
28. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology 2001; 54(8): 774–781. [PubMed: 11470385]
29. SEER*Stat Database. Surveillance, Epidemiology, and End Results (SEER) Program. www.seer.cancer.gov; 2019.
30. Wilbur J. Prostate cancer screening: the continuing controversy. American family physician 2008; 78(12): 1377–1384. [PubMed: 19119557]
31. U.S. Cancer Statistics Working Group. U.S. cancer statistics data visualizations tool, based on November 2018 submission data (1999–2016). www.cdc.gov/cancer/dataviz; 2019.

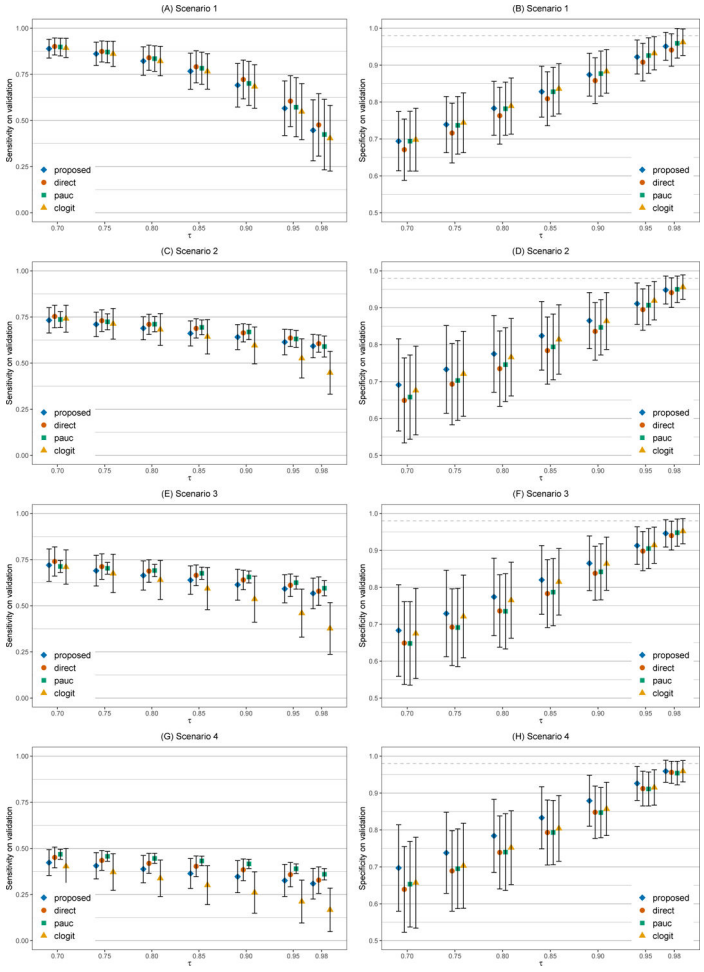

FIGURE 1.

Visualization of simulation results on validation data when the sample size of the training data is

$n_D = n_{\bar{D}} = 50$

. clogit: conditional logistic regression; τ: the pre-specified threshold of specificity; Gray dashed line: y-axis at 0.98.

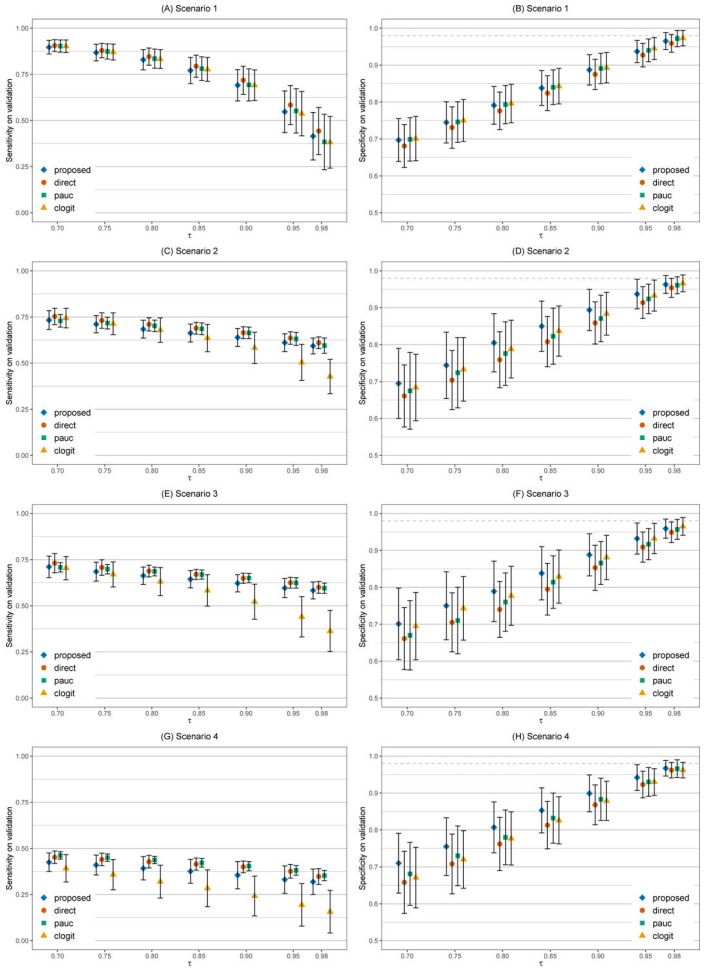

FIGURE 2.

Visualization of simulation results on validation data when the sample size of the training data is

$n_D = n_{\bar{D}} = 100$

. clogit: conditional logistic regression; τ: the pre-specified threshold of specificity; Gray dashed line: y-axis at 0.98.

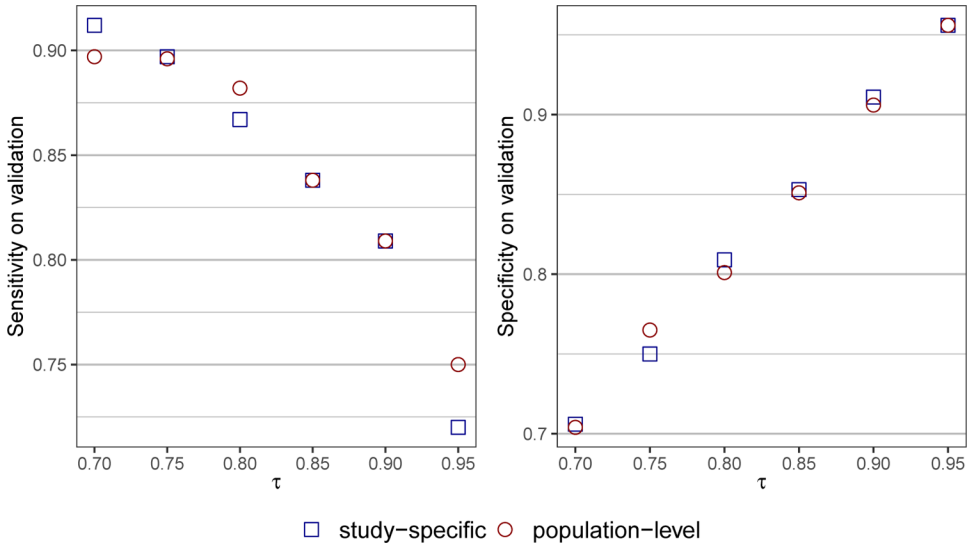

FIGURE 3.

Study-specific and population-level results of the proposed method applied to the prostate cancer data. τ: prespecified threshold of specificity.

TABLE 1

Summary statistics of estimated sensitivities on the training data. K: number of strata; τ: prespecified specificity; Clogit: conditional logistic regression; Mean: empirical mean sensitivity; ESE: empirical standard error; ASE: average of estimated standard errors.

Scenario	K	τ	Proposed			Direct		pAUC		Clogit	
			Mean	ESE	ASE	Mean	ESE	Mean	ESE	Mean	ESE
1	50	.70	.911	.061	.067	.925	.052	.901	.064	.898	.066
		.75	.885	.073	.079	.900	.065	.873	.074	.867	.080
		.80	.848	.087	.094	.867	.075	.838	.086	.829	.093
		.85	.794	.107	.113	.820	.093	.789	.102	.773	.110
		.90	.717	.130	.132	.752	.111	.709	.131	.692	.133
		.95	.597	.158	.140	.636	.146	.582	.171	.555	.166
		.98	.476	.172	.130	.506	.174	.435	.205	.413	.187
	100	.70	.914	.043	.046	.924	.038	.908	.043	.906	.044
		.75	.885	.053	.056	.898	.045	.880	.050	.876	.053
		.80	.846	.066	.068	.865	.055	.841	.061	.837	.064
		.85	.791	.081	.083	.815	.069	.788	.073	.783	.077
		.90	.711	.095	.103	.739	.084	.701	.096	.696	.094
		.95	.565	.121	.123	.603	.111	.558	.128	.540	.127
		.98	.432	.136	.120	.461	.132	.389	.157	.384	.146
2	50	.70	.767	.091	.090	.797	.077	.747	.078	.753	.095
		.75	.743	.092	.092	.774	.078	.736	.079	.725	.105
		.80	.720	.086	.092	.751	.077	.725	.078	.693	.108
		.85	.691	.091	.092	.725	.075	.707	.078	.653	.113
		.90	.668	.090	.093	.699	.075	.682	.078	.604	.122
		.95	.640	.093	.093	.665	.076	.641	.081	.533	.130
		.98	.615	.092	.091	.632	.080	.597	.087	.452	.137
	100	.70	.754	.067	.070	.781	.057	.733	.057	.748	.067
		.75	.729	.065	.071	.758	.055	.721	.057	.717	.072
		.80	.702	.066	.070	.734	.056	.705	.058	.683	.082
		.85	.678	.067	.070	.711	.054	.690	.058	.640	.092
		.90	.653	.067	.070	.686	.054	.667	.057	.586	.100
		.95	.623	.068	.069	.654	.054	.632	.059	.506	.112
		.98	.605	.063	.065	.624	.055	.598	.064	.432	.107
3	50	.70	.746	.105	.108	.770	.094	.709	.072	.714	.111
		.75	.714	.105	.109	.741	.091	.701	.071	.680	.124
		.80	.685	.103	.109	.713	.088	.690	.072	.643	.129
		.85	.659	.101	.108	.688	.085	.675	.074	.596	.138
		.90	.633	.106	.107	.660	.087	.657	.078	.540	.149
		.95	.609	.103	.107	.629	.092	.630	.080	.468	.154
		.98	.588	.109	.108	.598	.104	.601	.083	.385	.163
	100	.70	.726	.075	.081	.749	.067	.706	.048	.705	.078

Scenario	K	τ	Proposed			Direct		pAUC		Clogit	
			Mean	ESE	ASE	Mean	ESE	Mean	ESE	Mean	ESE
		.75	.697	.069	.078	.723	.061	.697	.049	.671	.085
		.80	.674	.067	.077	.702	.056	.685	.051	.634	.093
		.85	.655	.066	.077	.682	.053	.671	.051	.587	.102
		.90	.632	.068	.078	.661	.053	.653	.050	.528	.112
		.95	.607	.071	.078	.637	.055	.627	.054	.448	.125
		.98	.594	.068	.077	.612	.058	.598	.055	.373	.127
4	50	.70	.483	.097	.111	.522	.082	.515	.054	.459	.117
		.75	.466	.097	.112	.503	.080	.505	.055	.427	.121
		.80	.445	.099	.113	.484	.080	.494	.054	.390	.123
		.85	.418	.105	.116	.463	.081	.481	.053	.350	.128
		.90	.399	.109	.118	.443	.084	.463	.054	.305	.138
		.95	.378	.112	.122	.413	.092	.435	.055	.252	.143
		.98	.358	.109	.121	.380	.099	.403	.059	.200	.146
	100	.70	.465	.065	.081	.499	.052	.496	.036	.431	.087
		.75	.447	.071	.084	.484	.052	.486	.036	.396	.095
		.80	.429	.076	.088	.470	.050	.476	.037	.355	.104
		.85	.411	.079	.092	.455	.051	.460	.040	.316	.113
		.90	.389	.087	.096	.438	.049	.439	.042	.271	.123
		.95	.364	.088	.097	.413	.053	.417	.042	.219	.130
		.98	.351	.082	.094	.384	.059	.384	.045	.177	.130