The Canadian Journal of Statistics
Vol. xx, No. yy, 2022, Pages 1
La revue canadienne de statistique

Integrating Information from Existing Risk Prediction Models with No Model Details

Peisong Han[1]*, Jeremy M.G. Taylor[1] and Bhramar Mukherjee[1]

[1] *Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA*

Key words and phrases: Data integration; Empirical likelihood; Estimating equations; Estimation efficiency; External information.

MSC 2020: Primary 62F12; secondary 62J12

Abstract: Consider the setting where (i) individual-level data are collected to build a regression model for the association between an event of interest and certain covariates, and (ii) some risk calculators predicting the risk of the event using less detailed covariates are available, possibly as algorithmic black boxes with little information available about how they were built. We propose a general empirical-likelihood-based framework to integrate the rich auxiliary information contained in the calculators into fitting the regression model, to make the estimation of regression parameters more efficient. Two methods are developed, one using working models to extract the calculator information and one making a direct use of calculator predictions without working models. Theoretical and numerical investigations show that the calculator information can substantially reduce the variance of regression parameter estimation. As an application, we study the dependence of the risk of high grade prostate cancer on both conventional risk factors and newly identified molecular biomarkers by integrating information from the Prostate Biopsy Collaborative Group (PBCG) risk calculator, which was built based on conventional risk factors alone. *The Canadian Journal of Statistics* xx: 1–1; 2022 © 2022 Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–1; 2022 © 2022 Société statistique du Canada

1. INTRODUCTION

In the era of data science, it is common that data are collected from several sources and all provide useful information for answering the same scientific question. An analysis based on a single data source may yield biases in estimation or results that are not accurate enough. Integrating data from multiple sources becomes essential in order to pull together different pieces of information to draw more accurate conclusions and to make more insightful decisions. A common issue for data integration is that different sources usually provide information in different forms: some studies release the actual collected individual-level data whereas others only release aggregate data after the analysis. Different methods are needed to integrate different forms of data and to conduct inference.

This paper considers the setting where a current study collects individual-level data and builds a regression model to study the association between the risk of experiencing an event of interest and certain covariates, and where some risk prediction models for the same event using less detailed covariates have been built by previous studies and are accessible as risk calculators with little model detail released. These risk calculators contain rich information about the association of interest and thus it is highly desirable to integrate such information into fitting the regression model in order to improve the estimation efficiency and better understand how the risk is affected by different covariates, especially when the sample size of the current-study data is not large.

We consider the situation where some of the covariates are conventional risk factors known to be associated with the event and/or are typically adjusted for, and others are new potential factors whose association with the event has not been well studied. An example is our data application in Section 4 that studies the risk for high grade prostate cancer. Most existing studies on prostate cancer do not consider risk factors that are related to the molecular mechanisms of prostate cancer progression. However, it has been shown that prostate cancer antigen 3 (PCA 3) and TMPRSS2:ERG gene fusions are two biomarkers that have better specificity than prostate-specific antigen (PSA), a well-known conventional predictor, for early detection of prostate cancer (Tomlins et al., 2016). Therefore, in our data application we will build

* *Author to whom correspondence may be addressed.*
E-mail: peisong@umich.edu

CJS ???

an expanded regression model by including both biomarkers in addition to some conventional risk factors, such as age, race, PSA, digital rectal examination (DRE) findings, prior biopsy results, and family history. The setting we consider has a current study that collects individual-level data on both types of covariates. On the other hand, for the same event of interest, some existing studies have developed risk calculators that output the predicted risk of the event based on certain conventional factors. Many such risk calculators are accessible online. For instance, for prostate cancer, some widely used online risk calculators include the Prostate Cancer Prevention Trial (PCPT) risk calculator (Version 1, Thompson et al., 2006; Version 2, Ankerst et al., 2014) and the more recent Prostate Biopsy Collaborative Group (PBCG) risk calculator (Ankerst et al., 2018). Such risk calculators contain useful auxiliary information about the association of interest, but they are sometimes available as a black box with little detail released about the actual models used to build them, especially when they are built based on machine learning techniques (e.g. Mocellin et al., 2009).

The setting we consider is substantially different from the ones in the existing literature which assume that external study model details are available. Such details typically include the specification of the model, the estimated parameter values, and sometimes the corresponding standard errors (e.g. Imbens & Lancaster, 1994; Qin, 2000; Chatterjee et al., 2016; Cheng et al., 2018; Cheng et al., 2019; Han & Lawless, 2019; Huang and Qin, 2020; Sheng et al., 2021). It is also different from the settings where the population values of certain quantities are available and can be used as calibration factors to increase the efficiency of estimation (e.g. Deville & Sarndal, 1992; Chen & Qin, 1993; Chaudhuri et al., 2008; Lumley et al., 2011; Chen & Kim, 2014; Qin et al., 2015; Huang et al., 2016). The lack of such concrete information in our setting presents a unique challenge. Without assuming any detailed knowledge about the risk calculator models, we will develop methods to extract the auxiliary information contained in the risk calculators and integrate it into fitting the regression model of interest. Integrating such auxiliary information can considerably improve the estimation efficiency. Our setting is similar to that in Gu et al. (2019), who considered only one risk calculator and proposed a synthetic data method based on multiple imputation, fundamentally different from our development.

Our development is based on the empirical likelihood (EL) method (Owen, 1988, 2001; Qin & Lawless, 1994). The EL method has been widely adopted to integrate auxiliary information to improve the estimation efficiency, especially in survey sampling (e.g. Chen & Qin, 1993; Chen et al., 2002; Chaudhuri et al., 2008; Chen & Kim, 2014) and medicine and public health related research (e.g. Qin et al., 2015; Huang et al., 2016; Qin, 2017; Cheng et al., 2018; Han & Lawless, 2019). The constrained maximum likelihood estimation proposed in Chatterjee et al. (2016) has a strong connection to the EL method (Han & Lawless 2016). Compared to existing alternatives, such as those based on the generalized method of moments (e.g. Imbens & Lancaster, 1994), the generalized regression (e.g. Chen & Chen, 2000), weight calibration (e.g. Lumley et al., 2011), and others (e.g. Boonstra et al., 2013; Grill et al., 2015; Estes et al., 2018), the EL-type methods provide a likelihood-based framework for estimation and inference. Details about many superior properties of the EL method can be found in Owen (2001) and Qin (2017). Our development in this paper also extends the EL method to new settings of data integration and expands its applicability and effectiveness.

The rest of the paper is organized as follows. Section 2 contains our proposed methods, one based on postulating working models and one without. Section 3 gives simulation studies and Section 4 provides an application to prostate cancer research. Some discussions are given in Section 5. The Appendix includes some technical details for the results established in Section 2.

2. THE PROPOSED METHODS

2.1. Notation and Setup

The main interest is to study the association between the risk of experiencing an event and certain covariates, including both the conventional and the newly discovered risk factors. Let Y denote the binary outcome indicating if a subject experiences the event ($Y=1$ if yes and $Y=0$ if no), $\boldsymbol{X}$ the vector of conventional risk factors, and $\boldsymbol{Z}$ the vector of newly discovered risk factors. A regression model $P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ is specified for the risk $P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{\beta}$ is the vector of regression parameters and has the true value $\boldsymbol{\beta}_0$ such that $P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0) = P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z})$. A widely used regression model is the logistic regression

$$P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) = \exp(\beta_c + \boldsymbol{\beta}_X^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\beta}_Z^{\mathrm{T}}\boldsymbol{Z})/\{1 + \exp(\beta_c + \boldsymbol{\beta}_X^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\beta}_Z^{\mathrm{T}}\boldsymbol{Z})\}.$$

Let $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, denote the individual-level data collected based on a random sample of size n. Then the maximum likelihood estimator (MLE) for $\boldsymbol{\beta}_0$ is

$$\hat{\boldsymbol{\beta}}_{mle} \equiv \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}), \tag{1}$$

where $f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) = P(Y = 1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})^{I(Y=1)} P(Y = 0 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})^{I(Y=0)}$ and $I(\cdot)$ is the indicator function. Write $\boldsymbol{s}(\boldsymbol{\beta}) \equiv \boldsymbol{s}(Y, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) = \partial \log f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and $\boldsymbol{S}(\boldsymbol{\beta}) = E\{\boldsymbol{s}(\boldsymbol{\beta})\boldsymbol{s}(\boldsymbol{\beta})^{\mathrm{T}}\}$. From the theory of MLE (e.g. Lehmann & Casella, 2003), as $n \to \infty$, we know that $\hat{\boldsymbol{\beta}}_{mle} \xrightarrow{p} \boldsymbol{\beta}_0$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{mle} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{S}^{-1})$, and $\hat{\boldsymbol{\beta}}_{mle}$ has the optimal efficiency when no auxiliary information about $\boldsymbol{\beta}_0$ is available, where $\boldsymbol{S} \equiv \boldsymbol{S}(\boldsymbol{\beta}_0)$.

Previous studies on the same event of interest have usually produced rich auxiliary information about the association, and such information can be used to improve the estimation efficiency over the MLE $\hat{\boldsymbol{\beta}}_{mle}$. In this paper we consider the case where the results from previous studies are available through risk calculators. Suppose that there are J risk calculators for calculating the risk of experiencing the event. It is common that some risk factors collected in a current study were not used or available in previous studies, and, for those that were indeed used, the current study may take a finer measurement (e.g. a continuous measurement instead of a categorical one). To take these possibilities into consideration, let $\boldsymbol{Z}$ be the risk factors that are available only in the current study. For the jth calculator, $j = 1, \ldots, J$, let $\boldsymbol{X}_{(j)}$ be a possibly coarsened version of $\boldsymbol{X}$. For example, the variables in $\boldsymbol{X}_{(j)}$ may be a subset and/or a categorized version of those in $\boldsymbol{X}$. In other words, $\boldsymbol{X}_{(j)}$ is a many-to-one function of $\boldsymbol{X}$: the value of $\boldsymbol{X}_{(j)}$ can be completely determined by $\boldsymbol{X}$ but not the reverse. The jth calculator uses $\boldsymbol{X}_{(j)}$ as the predictors to calculate the risk of experiencing the event, $P(Y = 1 \mid \boldsymbol{X}_{(j)})$, and let $\hat{p}_{(j)}$ denote the corresponding predicted value for $P(Y = 1 \mid \boldsymbol{X}_{(j)})$. Thus, based on the collected individual-level data, the jth calculator produces data $(\boldsymbol{X}_{(j)i}, \hat{p}_{(j)i})$, $i = 1, \ldots, n$. To keep the generality of our development, we do not assume any other knowledge about these risk calculators.

2.2. A Method Based on Working Models

To extract information from the data $(\boldsymbol{X}_{(j)i}, \hat{p}_{(j)i})$ produced by the jth calculator, we postulate a working model $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)}) \equiv P(Y = 1 \mid \boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$ for $P(Y = 1 \mid \boldsymbol{X}_{(j)})$ with parameter vector $\boldsymbol{\theta}_{(j)}$. Such a working model represents our knowledge/belief about how the association between Y and $\boldsymbol{X}_{(j)}$ should be modeled. This working model may be different from the actual unknown model used to build the calculator and it may be incorrectly specified for $P(Y = 1 \mid \boldsymbol{X}_{(j)})$, but this does not prevent the working model from providing a reasonable approximation to $P(Y = 1 \mid \boldsymbol{X}_{(j)})$ and extracting the calculator information.

A challenge when fitting the working model $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$ based on the data $(\boldsymbol{X}_{(j)i}, \hat{p}_{(j)i})$ is that the $\hat{p}_{(j)}$ is a fixed function of $\boldsymbol{X}_{(j)}$ determined by the jth calculator with no randomness. If the output from the jth calculator were a random binary value $\hat{Y}_{(j)} \sim$ Bernoulli$(\hat{p}_{(j)})$, we could fit the working model $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$ by maximizing the likelihood

$$\prod_{i=1}^{n} p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})^{\hat{Y}_{(j)i}} \{1 - p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})\}^{1-\hat{Y}_{(j)i}},$$

which would lead to estimating $\boldsymbol{\theta}_{(j)}$ by solving the corresponding score equation

$$\sum_{i=1}^{n} \frac{\hat{Y}_{(j)i} - p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})}{p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})\{1 - p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})\}} \frac{\partial p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})}{\partial \boldsymbol{\theta}_{(j)}} = \boldsymbol{0}.$$

With the fixed value $\hat{p}_{(j)}$, we replace the $\hat{Y}_{(j)}$ in the above equation by $\hat{p}_{(j)}$ and estimate $\boldsymbol{\theta}_{(j)}$ by $\hat{\boldsymbol{\theta}}_{(j)}$ that solves

$$\sum_{i=1}^{n} \frac{\hat{p}_{(j)i} - p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})}{p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})\{1 - p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})\}} \frac{\partial p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})}{\partial \boldsymbol{\theta}_{(j)}} = \boldsymbol{0}. \tag{2}$$

Therefore, the information contained in the jth calculator is summarized by the working model $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$ and the parameter estimate $\hat{\boldsymbol{\theta}}_{(j)}$ that satisfies (2).

Another justification for fitting the model $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$ by solving (2) is to consider the weighted least squares discrepancy

$$\sum_{i=1}^{n} \left(\frac{\hat{p}_{(j)i} - p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})}{\sqrt{\hat{p}_{(j)i}(1 - \hat{p}_{(j)i})}} \right)^2$$

between $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$ and $\hat{p}_{(j)}$, where $\hat{p}_{(j)}(1 - \hat{p}_{(j)})$ is an estimate of the variance of the binary outcome Y at $\boldsymbol{X}_{(j)}$. Minimizing this discrepancy amounts to solving

$$\sum_{i=1}^{n} \frac{\hat{p}_{(j)i} - p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})}{\hat{p}_{(j)i}(1 - \hat{p}_{(j)i})} \frac{\partial p_{(j)}(\boldsymbol{X}_{(j)i}; \boldsymbol{\theta}_{(j)})}{\partial \boldsymbol{\theta}_{(j)}} = \mathbf{0}.$$

Equation (2) is the same as this equation with $\hat{p}_{(j)}(1 - \hat{p}_{(j)})$ replaced by $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})\{1 - p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})\}$, which is an estimate of the variance of Y based on the working model. Thus, $\hat{\boldsymbol{\theta}}_{(j)}$ actually makes $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$, the prediction of $P(Y = 1 \mid \boldsymbol{X}_{(j)})$ based on the working model, as close as possible to $\hat{p}_{(j)}$, the prediction based on the calculator.

To transform the summarized information into a form that can be integrated into the estimation of $\boldsymbol{\beta}_0$, consider for a moment the hypothetical scenario where the calculators output the exact true risk of experiencing the event; i.e. $\hat{p}_{(j)} \equiv P(Y = 1 \mid \boldsymbol{X}_{(j)})$. Then, because $\hat{\boldsymbol{\theta}}_{(j)}$ solves (2), we have $E_{(Y,\boldsymbol{X},\boldsymbol{Z})}\{\boldsymbol{h}_{(j)}(Y, \boldsymbol{X}_{(j)}; \boldsymbol{\theta}^*_{(j)})\} = \mathbf{0}$, where $\hat{\boldsymbol{\theta}}_{(j)} \xrightarrow{p} \boldsymbol{\theta}^*_{(j)}$ as $n \to \infty$ and

$$\boldsymbol{h}_{(j)}(Y, \boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)}) = \frac{Y - p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})}{p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})\{1 - p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})\}} \frac{\partial p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})}{\partial \boldsymbol{\theta}_{(j)}}.$$

Here and after, we occasionally use a subscript to explicitly indicate under which distribution is the expectation taken; e.g., $E_{(Y,\boldsymbol{X},\boldsymbol{Z})}(\cdot)$ is taken under the joint distribution of $(Y, \boldsymbol{X}, \boldsymbol{Z})$. Write

$$\boldsymbol{u}_{(j)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}_{(j)}) = \frac{P(Y = 1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) - p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})}{p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})\{1 - p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})\}} \frac{\partial p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})}{\partial \boldsymbol{\theta}_{(j)}},$$

we then have

$$\begin{aligned} E_{(Y,\boldsymbol{X},\boldsymbol{Z})}\{\boldsymbol{h}_{(j)}(Y, \boldsymbol{X}_{(j)}; \boldsymbol{\theta}^*_{(j)})\} &= E_{(\boldsymbol{X},\boldsymbol{Z})}[E_{(Y \mid \boldsymbol{X},\boldsymbol{Z})}\{\boldsymbol{h}_{(j)}(Y, \boldsymbol{X}_{(j)}; \boldsymbol{\theta}^*_{(j)}) \mid \boldsymbol{X}, \boldsymbol{Z}\}] \\ &= E_{(\boldsymbol{X},\boldsymbol{Z})}\{\boldsymbol{u}_{(j)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*_{(j)})\}, \end{aligned}$$

and thus

$$E_{(\boldsymbol{X},\boldsymbol{Z})}\{\boldsymbol{u}_{(j)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*_{(j)})\} = \mathbf{0}. \tag{3}$$

Therefore, through fitting the working model $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$, (3) summarizes the information regarding $\boldsymbol{\beta}_0$ contained in the jth calculator as a moment constraint under the marginal distribution of $(\boldsymbol{X}, \boldsymbol{Z})$.

To take the moment constraint (3) into account when estimating $\boldsymbol{\beta}_0$, we consider a discrete distribution $q_i = dF(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ on the data points $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, and propose to estimate $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}_{el1}$ defined through

$$\begin{aligned}
&\max_{\boldsymbol{\beta}, q_1, \ldots, q_n} \prod_{i=1}^{n} \{f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) q_i\} \quad \text{subject to} \\
&q_i > 0, \quad \sum_{i=1}^{n} q_i = 1, \quad \sum_{i=1}^{n} q_i \boldsymbol{u}_{(j)}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_{(j)}) = \mathbf{0} \quad (j = 1, \ldots, J).
\end{aligned} \tag{4}$$

Compared to (1), the maximization in (4) is over the joint distribution of $(Y, \boldsymbol{X}, \boldsymbol{Z})$, where the conditional distribution of $Y \mid (\boldsymbol{X}, \boldsymbol{Z})$ is parametrically modeled and the marginal distribution of $(\boldsymbol{X}, \boldsymbol{Z})$ is nonparametrically modeled subject to certain constraints that are a data version of (3). Therefore, $\hat{\boldsymbol{\beta}}_{el1}$ integrates the auxiliary information about $\boldsymbol{\beta}_0$ and thus should have higher efficiency compared to the MLE $\hat{\boldsymbol{\beta}}_{mle}$. The maximization (4) is similar to that in Qin (2000).

With J calculators, write

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_{(1)} \\ \boldsymbol{\theta}_{(2)} \\ \vdots \\ \boldsymbol{\theta}_{(J)} \end{pmatrix} \quad \text{and} \quad \boldsymbol{u}(\boldsymbol{\beta}, \boldsymbol{\theta}) \equiv \boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{u}_{(1)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}_{(1)}) \\ \boldsymbol{u}_{(2)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}_{(2)}) \\ \vdots \\ \boldsymbol{u}_{(J)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}_{(J)}) \end{pmatrix}.$$

In the Appendix we show that $\hat{\boldsymbol{\beta}}_{el1}$ is the component of $(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\rho}})$ that satisfies

$$\sum_{i=1}^{n} \boldsymbol{s}_i(\hat{\boldsymbol{\beta}}_{el1}) + \sum_{i=1}^{n} \frac{\partial \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\theta}}) / \partial \boldsymbol{\beta}^{\mathrm{T}}}{1 - \hat{\boldsymbol{\rho}}^{\mathrm{T}} \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\theta}})} \hat{\boldsymbol{\rho}} = \mathbf{0}, \tag{5}$$

$$\sum_{i=1}^{n} \frac{\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\theta}})}{1 - \hat{\boldsymbol{\rho}}^{\mathrm{T}} \boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\theta}})} = \mathbf{0}, \tag{6}$$

Equivalently, $\hat{\boldsymbol{\beta}}_{el1}$ is also the component of $(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\theta}})$ that satisfies (5), (6) and (2) simultaneously.

When studying the asymptotic properties of $\hat{\boldsymbol{\beta}}_{el1}$, a complication is that, for each risk calculator, the output risk prediction is based on a fixed function of the input covariates and does not accommodate the uncertainty associated with the model fitting when building the risk calculators. If the fixed function of $\boldsymbol{X}_{(j)}$ used by the jth calculator is not the same as $P(Y = 1 \mid \boldsymbol{X}_{(j)})$, $\hat{\boldsymbol{\beta}}_{el1}$ will not be theoretically consistent for $\boldsymbol{\beta}_0$, although the bias may be small if the output $\hat{p}_{(j)}$ is close to $P(Y = 1 \mid \boldsymbol{X}_{(j)})$. This is intuitive because the calculator information needs to compatible with that of the current study to yield improvements after data integration. Incompatible information will bias the estimation. Therefore, for the purpose of establishing asymptotic properties of $\hat{\boldsymbol{\beta}}_{el1}$, we assume that

$$\hat{p}_{(j)} = P(Y = 1 \mid \boldsymbol{X}_{(j)}), \quad j = 1, \ldots, J. \tag{7}$$

We then have the following theorem on $\hat{\boldsymbol{\beta}}_{el1}$.

Theorem 1

Under Regularity Conditions 1 specified in the Appendix, assuming (7) and as $n \to \infty$, we have (i) $\hat{\boldsymbol{\beta}}_{el1} \xrightarrow{p} \boldsymbol{\beta}_0$, and (ii) $\sqrt{n}(\hat{\boldsymbol{\beta}}_{el1} - \boldsymbol{\beta}_0)$ has an asymptotic normal distribution with mean $\mathbf{0}$ and variance

$$\left\{\boldsymbol{S} + \boldsymbol{G}_1^T \boldsymbol{\Omega}^{-1} \boldsymbol{G}_1 - \boldsymbol{G}_1^T (\boldsymbol{\Omega}^{-1} \boldsymbol{U}_1 - \boldsymbol{I})(\boldsymbol{G}_1 \boldsymbol{S}^{-1} \boldsymbol{G}_1^T + \boldsymbol{U}_1 \boldsymbol{\Omega}^{-1} \boldsymbol{U}_1)^{-1} (\boldsymbol{U}_1 \boldsymbol{\Omega}^{-1} - \boldsymbol{I}) \boldsymbol{G}_1\right\}^{-1}, \tag{8}$$

where $\boldsymbol{G}_1 \equiv E\{\partial \boldsymbol{u}(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*)/\partial \boldsymbol{\beta}\}$, $\boldsymbol{\Omega} \equiv E\left[\boldsymbol{u}(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*) - \tilde{E}\{\boldsymbol{u}(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*)\}\right]^{\otimes 2}$ *with*

$$\tilde{E}\{\boldsymbol{u}(\boldsymbol{\beta}, \boldsymbol{\theta})\} \equiv \begin{pmatrix} E\{\boldsymbol{u}_{(1)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}_{(1)}) \mid \boldsymbol{X}_{(1)}\} \\ \vdots \\ E\{\boldsymbol{u}_{(J)}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}_{(J)}) \mid \boldsymbol{X}_{(J)}\} \end{pmatrix}$$

and $E\boldsymbol{A}^{\otimes 2} = E(\boldsymbol{A}\boldsymbol{A}^T)$ *for any matrix* $\boldsymbol{A}$, $\boldsymbol{U}_1 \equiv E\{\boldsymbol{u}(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*)\boldsymbol{u}(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*)^T\}$, $\boldsymbol{I}$ *is the identity matrix, and* $\boldsymbol{\theta}^* \equiv (\boldsymbol{\theta}^*_{(1)}, \ldots, \boldsymbol{\theta}^*_{(J)})$.

The proof of Theorem 1 is given in the Appendix. Due to the complexity of (8), there is no general clear comparison to $\boldsymbol{S}^{-1}$, the asymptotic variance of the MLE $\hat{\boldsymbol{\beta}}_{mle}$. But an efficiency improvement over $\hat{\boldsymbol{\beta}}_{mle}$ can be anticipated when the calculators are not poorly built and when the working models are reasonably postulated. Indeed, in such settings comprehensive simulation studies have shown that $\hat{\boldsymbol{\beta}}_{el1}$ provides a substantial efficiency gain over $\hat{\boldsymbol{\beta}}_{mle}$. In the next subsection we will propose an alternative estimator that is guaranteed to improve over $\hat{\boldsymbol{\beta}}_{mle}$.

2.3. A Method Without Working Models

The working model approach summarizes the auxiliary information contained in the calculators by fitting a working model to the data $(\boldsymbol{X}_{(j)i}, \hat{p}_{(j)i})$. As seen from (8), the estimation of $\boldsymbol{\theta}_{(j)}$ introduced by the working models has a very complex effect on the asymptotic variance of $\hat{\boldsymbol{\beta}}_{el1}$. In this section we consider an alternative method without working models, the efficiency of which will not be compromised by the estimation of any nuisance parameters.

For any arbitrary vector function $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ of $\boldsymbol{X}_{(j)}$, assuming all relevant moments exist, we have

$$E_{(\boldsymbol{X},\boldsymbol{Z})}[\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})\{P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}) - P(Y=1 \mid \boldsymbol{X}_{(j)})\}] = \boldsymbol{0}. \tag{9}$$

Since $P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}) = P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)$ and the jth calculator outputs $\hat{p}_{(j)}$ as a prediction for $P(Y=1 \mid \boldsymbol{X}_{(j)})$, (9) provides an alternative to (3) to summarize the auxiliary information contained in the jth calculator, again in the form of a moment equality under the distribution of $(\boldsymbol{X}, \boldsymbol{Z})$. To integrate this summary information into the estimation of $\boldsymbol{\beta}_0$, let $q_i = dF(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ and define another estimator $\hat{\boldsymbol{\beta}}_{el2}$ through

$$\max_{\boldsymbol{\beta}, q_1, \ldots, q_n} \prod_{i=1}^n \{f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta})q_i\} \quad \text{subject to} \quad q_i > 0, \quad \sum_{i=1}^n q_i = 1,$$
$$\sum_{i=1}^n q_i \boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)i})\{P(Y_i = 1 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) - \hat{p}_{(j)i}\} = \boldsymbol{0} \quad (j = 1, \ldots, J). \tag{10}$$

This estimator of $\boldsymbol{\beta}_0$ directly uses the outputs from the calculators instead of postulating working models. For each calculator, a vector function $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ needs to be chosen. For a chosen $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$, the constraints in (10) are similar to those in (4) when the working model $p_{(j)}(\boldsymbol{X}_{(j)}; \boldsymbol{\theta}_{(j)})$ is the logistic regression

$$\frac{\exp\{\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})^{\mathrm{T}}\boldsymbol{\theta}_{(j)}\}}{1 + \exp\{\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})^{\mathrm{T}}\boldsymbol{\theta}_{(j)}\}}, \tag{11}$$

because in this case the last constraint in (4) becomes

$$\sum_{i=1}^n q_i \boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)i})\{P(Y_i = 1 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) - p_{(j)}(\boldsymbol{X}_{(j)i}; \hat{\boldsymbol{\theta}}_{(j)})\},$$

and thus the only difference between (4) and (10) is in which prediction of $P(Y=1 \mid \boldsymbol{X}_{(j)})$ is used, the $p_{(j)}(\boldsymbol{X}_{(j)}; \hat{\boldsymbol{\theta}}_{(j)})$ from the working model or the $\hat{p}_{(j)}$ from the calculator. In this case, intuitively, because the maximization in (10) does not involve estimation of any nuisance parameters, $\hat{\boldsymbol{\beta}}_{el2}$ should have a higher efficiency compared to $\hat{\boldsymbol{\beta}}_{el1}$, as the asymptotic variance of $\hat{\boldsymbol{\beta}}_{el2}$ will not have a component coming from the estimation of $\boldsymbol{\theta}_{(j)}$. This intuition is confirmed by our theoretical results below.

For ease of notation, write

$$\boldsymbol{u}(\boldsymbol{\beta}) \equiv \boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{d}_{(1)}(\boldsymbol{X}_{(1)})\{P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) - \hat{p}_{(1)}\} \\ \boldsymbol{d}_{(2)}(\boldsymbol{X}_{(2)})\{P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) - \hat{p}_{(2)}\} \\ \vdots \\ \boldsymbol{d}_{(J)}(\boldsymbol{X}_{(J)})\{P(Y=1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) - \hat{p}_{(J)}\} \end{pmatrix}.$$

Based on derivations similar to those leading to (5) and (6), $\hat{\boldsymbol{\beta}}_{el2}$ is the component of $(\hat{\boldsymbol{\beta}}_{el2}, \hat{\boldsymbol{\rho}})$ that satisfies (5) and (6) but with $\boldsymbol{u}(\boldsymbol{\beta}, \boldsymbol{\theta})$ replaced by $\boldsymbol{u}(\boldsymbol{\beta})$. Following the same proof as that for Theorem 1, the properties of $\hat{\boldsymbol{\beta}}_{el2}$ are given below.

Theorem 2

Under Regularity Conditions 2 specified in the Appendix, assuming (7) and as $n \to \infty$, we have (i) $\hat{\boldsymbol{\beta}}_{el2} \xrightarrow{p} \boldsymbol{\beta}_0$, and (ii) $\sqrt{n}(\hat{\boldsymbol{\beta}}_{el2} - \boldsymbol{\beta}_0)$ has an asymptotic normal distribution with mean $\mathbf{0}$ and variance

$$\left\{\boldsymbol{S} + \boldsymbol{G}_2^T \boldsymbol{U}_2^{-1} \boldsymbol{G}_2\right\}^{-1}, \tag{12}$$

where $\boldsymbol{G}_2 \equiv E\{\partial \boldsymbol{u}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}\}$ and $\boldsymbol{U}_2 \equiv E\{\boldsymbol{u}(\boldsymbol{\beta}_0)\boldsymbol{u}(\boldsymbol{\beta}_0)^T\}$.

It is clear that (12) is smaller than $\boldsymbol{S}^{-1}$, the asymptotic variance of the MLE $\hat{\boldsymbol{\beta}}_{mle}$. Thus, $\hat{\boldsymbol{\beta}}_{el2}$ is expected to be more efficient than $\hat{\boldsymbol{\beta}}_{mle}$ when the calculators were not poorly built. There is no general comparison between (8) and (12) because (8) depends on the working models postulated and (12) depends on the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ chosen. However, in the setting where all working models for $\hat{\boldsymbol{\beta}}_{el1}$ are logistic regression as in (11) and all vector functions $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ for $\hat{\boldsymbol{\beta}}_{el2}$ are taken to be the same as those in (11), a direct comparison between (8) and (12) is possible. In this case, some calculation shows that $\boldsymbol{G}_1 = \boldsymbol{G}_2$ and $\boldsymbol{\Omega} = \boldsymbol{U}_2$. Therefore, (12) is smaller than (8) because the third term that is subtracted in (8) is semi-positive definite. In this case $\hat{\boldsymbol{\beta}}_{el2}$ has a higher efficiency than $\hat{\boldsymbol{\beta}}_{el1}$. Because of the popularity of logistic regression, this efficiency comparison is widely applicable in practice. Another important observation from (12) is that, since $\boldsymbol{G}_2^{\mathrm{T}} \boldsymbol{U}_2^{-1} \boldsymbol{G}_2$ becomes larger in the positive-definite sense as the dimension of $\boldsymbol{u}(\boldsymbol{\beta})$ increases, (12) becomes smaller as more calculators are integrated, as long as the calculators output correct predictions and do not use exactly the same predictors.

2.4. Some Discussion and Remarks

Because of the efficiency properties of $\hat{\boldsymbol{\beta}}_{el2}$ and the direct use of the output probabilities $\hat{p}_{(j)}$ from the calculators without introducing any nuisance parameters, it is more desirable to implement $\hat{\boldsymbol{\beta}}_{el2}$ in practice, which requires specifying the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$. In theory, the asymptotic variance (12) becomes smaller as more functions are included in $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$. In practice, however, a large dimension of $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ may jeopardize the numerical performance. The specification of the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ can be guided by first building working models, such as logistic regression models, and then taking the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ to be the corresponding regressors. The working models should be specified based both on existing scientific knowledge about the association of interest and on the calculator data $(\boldsymbol{X}_{(j)}, \hat{p}_{(j)})$. When partial information is available about the original models used to build the calculators, such as the inclusion of certain interactions, it should be accommodated when specifying the working models. Intuitively, working models that are close to the possibly unknown true calculator models should lead to good final performance, because the true models contain all the calculator information. After the working models are specified as in (11), the corresponding $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ can be used to implement

$\hat{\boldsymbol{\beta}}_{el2}$. The resulting $\hat{\boldsymbol{\beta}}_{el2}$ is guaranteed to be more efficient than both $\hat{\boldsymbol{\beta}}_{mle}$ and the $\hat{\boldsymbol{\beta}}_{el1}$ based on the specified working models. In this way the working models are specified only as a guideline for choosing the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$, and the efficiency of $\hat{\boldsymbol{\beta}}_{el2}$ is not affected by the estimation of any nuisance parameters.

The large sample properties of $\hat{\boldsymbol{\beta}}_{el1}$ and $\hat{\boldsymbol{\beta}}_{el2}$ are established under the assumption that $\hat{p}_{(j)} = P(Y = 1 \mid \boldsymbol{X}_{(j)})$. In reality, due to model specification and random errors when building the calculators, $\hat{p}_{(j)}$ is not the same as $P(Y = 1 \mid \boldsymbol{X}_{(j)})$. For example, the model of interest $P(Y = 1 \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ implicitly imposes restrictions on modeling $P(Y = 1 \mid \boldsymbol{X}_{(j)})$, which may not be met by the calculator models. However, the proposed methods should still lead to small finite-sample bias and considerable efficiency gains compared to $\hat{\boldsymbol{\beta}}_{mle}$ if $\hat{p}_{(j)}$ is a good approximation to $P(Y = 1 \mid \boldsymbol{X}_{(j)})$. This should be the case when the calculators were built based on carefully specified models and decent sample sizes so that they captured most of the association between the risk of interest and corresponding covariates. In practice, there are ways to quickly check the quality of this approximation. One way is to compare the data $(\boldsymbol{X}_{(j)i}, \hat{p}_{(j)i})$ produced by the calculators to the data $(\boldsymbol{X}_{(j)i}, Y_i)$ using some simple quantities, such as the means of $\hat{p}_{(j)}$ and Y within each level of $\boldsymbol{X}_{(j)}$. Another way is to compare the coefficients of the regression of Y_i on $\boldsymbol{X}_{(j)i}$ to the coefficients of the regression of $\hat{p}_{(j)i}$ on $\boldsymbol{X}_{(j)i}$. The comparison can be made by constructing confidence intervals for the former coefficients and check if they cover the latter coefficients. In our data application in Section 4 we carry out such a comparison.

The proposed methods are very flexible in the sense that they apply to black-box-type calculators where little information is available about how the calculators were built, especially when they were built based on machine learning techniques. Some extra care may be needed when there are multiple calculators based on similar covariates. In this case, the auxiliary information provided by these calculators may be similar because the information is about the association between the outcome of interest and the corresponding covariates. Although in theory $\hat{\boldsymbol{\beta}}_{el2}$ will keep gaining efficiency when integrating more calculators, using all these calculators simultaneously might jeopardize numerical performance because some constraints in (4) and (10) may become highly correlated. Therefore, in the presence of multiple calculators with similar inputs, we would recommend using the one(s) based on large sample sizes and based on populations similar to the current study, which may be checked by the procedures mentioned in the previous paragraph.

A simple way to implement the proposed methods might seem to be to solve the equations in (5) and (6). However, this procedure is not recommended owing to its unstable behavior: Equation (6), viewed as an equation for $\boldsymbol{\rho}$ for fixed $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, typically has many roots (Han & Wang, 2013). Here we need the $\hat{\boldsymbol{\rho}}$ such that the $\hat{q}_i$ maximizing (4) or (10) are between 0 and 1. Solving (5) and (6) directly can lead to an unwanted root. Please refer to the derivation of (5) and (6) in the Appendix for an expression for $\hat{q}_i$. A more reliable implementation is to follow the Newton-Raphson-type algorithm provided in Han & Lawless (2019).

3. SIMULATION STUDIES

In this section we carry out simulation studies to investigate the finite-sample performance of the proposed methods. The covariates $\boldsymbol{X} = (X_1, X_2)$ and Z are generated from a three dimensional multivariate normal distribution, where the means are all zero, the variances are all one, and the correlations are all 0.4. Given the covariates, the response Y is generated from a Bernoulli distribution with

$$P(Y = 1 \mid \boldsymbol{X}, Z) = \text{expit}(0.5 - 0.5X_1 - 0.5X_2 + 0.5Z + 0.5X_2Z),$$

where $\text{expit}(x) = e^x/(1 + e^x)$. Therefore we have $\boldsymbol{\beta}_0 = (\beta_c, \beta_{X_1}, \beta_{X_2}, \beta_Z, \beta_{X_2Z}) = (0.5, -0.5, -0.5, 0.5, 0.5)$. We consider Calculators 1, 2 and 3 constructed based on the generated data using covariates $X_{(1)} = X_1$, $\boldsymbol{X}_{(2)} = (X_1, \tilde{X}_2)$, and $\boldsymbol{X}_{(3)} = (X_1, X_2)$, respectively, where $\tilde{X}_2 = I(X_2 > 0)$. Calculator 1 for $P(Y = 1 \mid X_{(1)})$ is based on the semiparametric single index model in Klein & Spady (1993) implemented in R (R Core Team, 2021) with package *np* (Hayfield & Racine, 2008). Because there is only one covariate, this semiparametric model becomes nonparametric. Calculator 2 for $P(Y = 1 \mid \boldsymbol{X}_{(2)})$ is based on the same nonparametric model but for $\tilde{X}_2 = 1$ and $\tilde{X}_2 = 0$ separately. Calculator 3 for $P(Y = 1 \mid \boldsymbol{X}_{(3)})$ is based on the random forest method (Breiman, 2001) implemented in the R package *randomForest* (Liaw & Wiener, 2002). The nonparametric nature of these models requires a large sample size to achieve a good quality for risk prediction, and we used a sample size of 50000 to build these calculators to ensure their good quality.

First, we compare estimators $\hat{\boldsymbol{\beta}}_{el1}$ based on working models, $\hat{\boldsymbol{\beta}}_{el2}$ without using working models, and the MLE $\hat{\boldsymbol{\beta}}_{mle}$ without integrating calculator information. In this comparison we fix the working models or the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ for each calculator $j = 1, 2, 3$. Specifically, for $\hat{\boldsymbol{\beta}}_{el1}$, the postulated working models for the calculators are all logistic regression as specified in (11), where $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ contains the intercept and corresponding main effects; i.e., $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)}) = (1, \boldsymbol{X}_{(j)})$, $j = 1, 2, 3$. These same $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ are used for $\hat{\boldsymbol{\beta}}_{el2}$. For both $\hat{\boldsymbol{\beta}}_{el1}$ and $\hat{\boldsymbol{\beta}}_{el2}$, we consider seven versions, $\hat{\boldsymbol{\beta}}_{el}$-1, $\hat{\boldsymbol{\beta}}_{el}$-2, $\hat{\boldsymbol{\beta}}_{el}$-3, $\hat{\boldsymbol{\beta}}_{el}$-12, $\hat{\boldsymbol{\beta}}_{el}$-13, $\hat{\boldsymbol{\beta}}_{el}$-23, and $\hat{\boldsymbol{\beta}}_{el}$-123, where the numbers indicate which calculators are incorporated into the estimation.

Tables 1 and 2 contain the simulation results based on $n = 400$ and $n = 1000$, respectively, both using 1000 replications. Compared to $\hat{\boldsymbol{\beta}}_{mle}$, the empirical standard errors of $\hat{\boldsymbol{\beta}}_{el1}$ and $\hat{\boldsymbol{\beta}}_{el2}$ for those covariates also used by the corresponding calculators are substantially smaller, confirming our theoretical conclusion of efficiency gains by integrating calculator information. Note that, for $\hat{\boldsymbol{\beta}}_{el1}$-23 and $\hat{\boldsymbol{\beta}}_{el1}$-123, the empirical standard errors corresponding to those covariates not used by the calculators (i.e. Z and $X_2 Z$) become larger compared to $\hat{\boldsymbol{\beta}}_{mle}$. This is caused by the estimation of nuisance parameters in the working models. As seen from (8), the effect of estimating nuisance parameters on the efficiency of $\hat{\boldsymbol{\beta}}_{el1}$ is quite complex. Although in general we anticipate efficiency gains for $\hat{\boldsymbol{\beta}}_{el1}$, the gains are typically small for the coefficients that do not appear in the calculators and may even be negative due to estimation of nuisance parameters.

The comparison of empirical standard errors between $\hat{\boldsymbol{\beta}}_{el1}$ and $\hat{\boldsymbol{\beta}}_{el2}$ confirms that the latter is more efficient, especially when more than one calculator is used. Empirical standard errors for $\hat{\boldsymbol{\beta}}_{el2}$ either decrease or stay about the same as more calculators are integrated, consistent with our theory, whereas those for $\hat{\boldsymbol{\beta}}_{el1}$ may increase considerably due to the estimation of more nuisance parameters. It is also seen that, for all considered estimators, the mean of estimated standard errors over 1000 replications is very close to the corresponding empirical standard error, especially when $n = 1000$, confirming the derived asymptotic variances for both $\hat{\boldsymbol{\beta}}_{el1}$ and $\hat{\boldsymbol{\beta}}_{el2}$.

Second, we vary the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ for each calculator to assess its effects. We also make a comparison of the proposed method to that in Gu et al. (2019). Since $\hat{\boldsymbol{\beta}}_{el2}$ is superior to $\hat{\boldsymbol{\beta}}_{el1}$ both theoretically and numerically, we now only include $\hat{\boldsymbol{\beta}}_{el2}$. We focus on the three versions $\hat{\boldsymbol{\beta}}_{el2}$-1, $\hat{\boldsymbol{\beta}}_{el2}$-2 and $\hat{\boldsymbol{\beta}}_{el2}$-3 since the method in Gu et al. (2019) deals with one calculator. The three versions of Gu et al.'s estimator are $\hat{\boldsymbol{\beta}}_{GTCM}$-1, $\hat{\boldsymbol{\beta}}_{GTCM}$-2 and $\hat{\boldsymbol{\beta}}_{GTCM}$-3. For our estimators we consider the following specifications of the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$. For $\hat{\boldsymbol{\beta}}_{el2}$-1, take $\boldsymbol{d}_{(1)}(\boldsymbol{X}_{(1)})$ to be $(1, X_{(1)})$ and $(1, X_{(1)}, X_{(1)}^2)$, resulting in $\hat{\boldsymbol{\beta}}_{el2}$-1-1 and $\hat{\boldsymbol{\beta}}_{el2}$-1-2, respectively. For $\hat{\boldsymbol{\beta}}_{el2}$-2, take $\boldsymbol{d}_{(2)}(\boldsymbol{X}_{(2)})$ to be $(1, X_1, \tilde{X}_2)$, $(1, X_1, \tilde{X}_2, X_1\tilde{X}_2)$ and $(1, X_1, \tilde{X}_2, X_1\tilde{X}_2, X_1^2)$, resulting in $\hat{\boldsymbol{\beta}}_{el2}$-2-1 and $\hat{\boldsymbol{\beta}}_{el2}$-2-2 and $\hat{\boldsymbol{\beta}}_{el2}$-2-3, respectively. For $\hat{\boldsymbol{\beta}}_{el2}$-3, take $\boldsymbol{d}_{(3)}(\boldsymbol{X}_{(3)})$ to be $(1, X_1, X_2)$, $(1, X_1, X_2, X_1X_2)$ and $(1, X_1, X_2, X_1X_2, X_1^2, X_2^2)$, resulting in $\hat{\boldsymbol{\beta}}_{el2}$-3-1 and $\hat{\boldsymbol{\beta}}_{el2}$-3-2 and $\hat{\boldsymbol{\beta}}_{el2}$-3-3, respectively.

Table 3 contains the simulation results based on $n = 400$ and $n = 1000$ using 1000 replications. The MLE $\hat{\boldsymbol{\beta}}_{mle}$ is also included as the benchmark for comparison. When varying the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ for the proposed estimator $\hat{\boldsymbol{\beta}}_{el2}$ based on either Calculator 2 or Calculator 3, an inclusion of the interaction between the calculator covariates (i.e. $X_1\tilde{X}_2$ for Calculator 2 or X_1X_2 for Calculator 3) in addition to the main effects leads to further efficiency gains for estimating $\beta_{X_2 Z}$ (i.e. comparing EL2-2-2 to EL2-2-1 or comparing EL2-3-2 to EL2-3-1). A further inclusion of the quadratic effects does not lead to additional efficiency gains. These observations suggest that interactions should be included in $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ if the current study model includes interactions that involve the calculator covariates. We leave a detailed theoretical and numerical study of this as a future research topic. Overall, the estimators GTCM-1, GTCM-2 and GTCM-3 based on Gu et al. (2019) sometimes have small biases, and the latter two have roughly the same or slightly larger empirical standard errors compared to our estimators EL2-2-2 and EL2-3-2 that include interactions in $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$. For all the estimators under comparison, the efficiency for estimating β_Z does not change much from the MLE. This observation makes sense because the calculators provide little auxiliary information about the regression coefficients for the covariates not used by the calculators. This observation is also in full agreement with findings in the existing literature (e.g. Chatterjee et al., 2016; Cheng et al., 2018; Han & Lawless, 2019).

4. DATA APPLICATION

As an application, we fit an expanded regression model for high grade prostate cancer by including two biomarkers, prostate cancer antigen 3 (PCA 3) and TMPRSS2:ERG gene fusions, measured from urine, in addition to the conventional risk factors: prostate-specific antigen (PSA), age, race, digital rectal examination (DRE) findings, prior biopsy results, and family history. To improve efficiency, we will integrate information from an online accessible risk calculator, the Prostate Biopsy Collaborative Group (PBCG) risk calculator (Ankerst et al., 2018).

The PBCG risk calculator was built as a state-of-the-art risk prediction tool, an alternative to the widely used Prostate Cancer Prevention Trial (PCPT) risk calculator (Version 1, Thompson et al., 2006; Version 2, Ankerst et al., 2014). The PCPT calculator was the first online prostate cancer risk assessment tool and was built based on data collected in the 1990s from 5519 men, mostly white, in the placebo group of the PCPT. In contrast, the PBCG risk calculator was built on data from 15611 men undergoing prostate biopsies during 2006-2017 at eight North American institutions and three European institutions participating in the PBCG. The heterogeneity in the study cohorts in PBCG endows the PBCG risk calculator with a much wider applicability compared to the PCPT calculator.

The data set we use is the validation cohort from Tomlins et al. (2016), who investigated whether including two additional biomarkers, PCA 3 and TMPRSS2:ERG, could give a more accurate risk prediction compared to the PCPT calculator. This cohort consists of 1244 men presenting for diagnostic biopsy at seven community clinics throughout the United States. Since the PBCG risk calculator requires an input PSA level between 2 and 50 ng/ml and an input age between 40 and 90 years old, in our analysis we remove subjects with PSA and age outside those range. The final analysis is based on 1014 men.

The logistic regression model of interest is

$$\log \frac{\pi}{1-\pi} = \beta_c + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 Z_1 + \beta_8 Z_2,$$

where π is the probability of observing high grade prostate cancer, $X_1 = \log_2(\text{PSA})$ (log transformation of the PSA level with base 2), $X_2 =$ age, $X_3 =$ DRE (a binary indicator of an abnormal digital rectal exam), $X_4 =$ biopsy (a binary indicator of prior negative biopsy), $X_5 =$ race (a binary indicator of African Ancestry) and $X_6 =$ family history (a binary indicator of first-degree family history of prostate cancer) are the conventional risk predictors, and $Z_1 = \log_2(\text{PCA3} + 1)$ and $Z_2 = \text{T2:ERG}$ are the two biomarkers. Here, following Cheng et al. (2019), we take the $\log_2$ transformation of PCA3 and dichotomize TMPRSS2:ERG by splitting at the median. The PBCG risk calculator uses X_1 through X_6 as the input and the exact formula for predicting the risk of high grade prostate cancer is given in the supplementary material of Ankerst et al. (2018). In the same supplementary Ankerst et al. (2018) also gave formulas for risk prediction when some or all of X_3, X_4 and X_6 are not available. Therefore, as an illustration, in our application we also consider predictions based only on X_1, X_2 and X_5 as input, and refer to this prediction as calculator PBCG-2. Note that, since these two calculators were built using the same data and the input of PBCG-2 is a subset of PBCG, we do not anticipate an efficiency improvement by integrating PBCG-2 in addition to PBCG.

As a checking of the assumption that $\hat{p}_{(j)}$ should be close to $P(Y = 1 \mid \boldsymbol{X}_{(j)})$, we construct 95% confidence intervals for the regression coefficients of $\boldsymbol{X}_{(j)}$ based on a logistic regression of Y on $\boldsymbol{X}_{(j)}$, and then check if these intervals cover the corresponding coefficients based on a logistic model for $\hat{p}_{(j)}$ conditional on $\boldsymbol{X}_{(j)}$, where the latter coefficients are computed by solving equation (2). We found coverage for all coefficients for both calculators. When looking at the 68% confidence intervals instead, we found coverage for all coefficients but not that of X_1 for both calculators. These findings show that it is reasonable to assume $\hat{p}_{(j)}$ is close to $P(Y = 1 \mid \boldsymbol{X}_{(j)})$.

Table 4 contains the analysis results based on both $\hat{\boldsymbol{\beta}}_{mle}$ and $\hat{\boldsymbol{\beta}}_{el2}$. For $\hat{\boldsymbol{\beta}}_{el2}$, the $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ is taken to be $(1, X_1, X_2, X_3, X_4, X_5, X_6)$ and $(1, X_1, X_2, X_5)$ for the PBCG and PBCG-2 calculators, respectively. Results based on $\hat{\boldsymbol{\beta}}_{el1}$ are very similar to those based on $\hat{\boldsymbol{\beta}}_{el2}$ and are thus omitted. A major observation is that, after integrating information from the PBCG calculator, the standard errors corresponding to X_1 through X_6, the covariates also used by the PBCG calculator, reduce to a third or a quarter of those for the MLE. Standard errors corresponding to Z_1 and Z_2 also become smaller. Similar observations can be made when integrating the PBCG-2 calculator. These are in full agreement with our theoretical results. Integrating both calculators simultaneously produces more or less the same results as integrating the PBCG calculator alone, which is what we anticipated. Another major observation is that integrating information from the PBCG calculator reveals a marginal significance of African Ancestry (p-value$= 0.062$) and a high significance of family history (p-value< 0.001) for their association with high grade prostate cancer. This significance is not detected without integrating the PBCG information.

5. DISCUSSION

We have proposed two methods for integrating information contained in existing risk calculators into estimation of regression parameters. The first method relies on working models to extract information contained in the calculators and the second method directly uses the risk predictions without working models. The second method is recommended in practice because its efficiency gain is not compromised by the estimation of any nuisance parameters introduced by working models. Given that many risk calculators have been developed for various diseases and many of them are of black-box-type, our proposed methods have a broad range of applications.

A potential issue when integrating information from multiple calculators is that these calculators may target different populations, either because of an original design or because of the sample based on which they were built. When the population from which the current study sample is taken is different from the populations the existing calculators can be applied to, $\hat{p}_{(j)}$ may no longer be a good approximation to $P(Y = 1 \mid \boldsymbol{X}_{(j)})$, and in this case integrating information from such calculators will introduce bias. Thus, it is crucial to use calculators whose target population is the same as the current one. This indeed was the consideration in our data application when choosing the PBCG calculator over the PCPT calculator, as the former was built based on multiple heterogeneous cohorts and thus has a wider applicability. Recently Estes et al. (2018) and Sheng et al. (2021) proposed methods to address the population heterogeneity problem, under the setting where external models and parameter estimates are directly available. These methods can be combined with the techniques developed in this paper to deal with black-box calculators in the presence of population heterogeneity. New methods dealing with population heterogeneity could also be developed for data integration purposes. For example, as pointed out by one referee, the instance weighting method for domain adaption in the natural language processing literature may be adopted (e.g. Jiang & Zhai, 2007).

In this paper we considered binary outcomes that indicate whether experiencing an event or not. The proposed methods can be directly applied to continuous outcomes under linear regression models with calculators predicting the outcome values. A desirable extension is to the setting of survival outcomes where calculators are available that give the, say, 5-year survival probabilities based on a set of risk factors. Because of the importance of survival outcomes in many areas, including medicine and public health, such an extension is highly desirable and will be investigated in our future research.

ACKNOWLEDGEMENTS

We thank the Editor, Associate Editor, and two referees for their constructive comments that have helped improve the quality of the paper. The research of the first author was partially supported by a University of Michigan start-up grant, the research of the second author was partially supported by the US National Institutes of Health grant CA129102, and the research of the third author was partially supported by the National Science Foundation grant DMS 1712933 and National Institutes of Health grant R01-HG008773-01.

Appendix

Derivation of (5) and (6)

Using the Lagrange multiplier method, the Lagrangian corresponding to the constrained optimization problem (4) is

$$\mathcal{L} = \sum_{i=1}^{n} \log f_i(\boldsymbol{\beta}) + \sum_{i=1}^{n} \log q_i + n\boldsymbol{\rho}^{\mathrm{T}} \sum_{i=1}^{n} q_i \boldsymbol{u}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}) - \mu(\sum_{i=1}^{n} q_i - 1),$$

where $\boldsymbol{\rho}$ and μ are the Lagrange multipliers. At the solution $\hat{\boldsymbol{\beta}}_{el1}$ and $\hat{q}_i$ we must have $\partial\mathcal{L}/\partial q_i = 0$ and $\partial\mathcal{L}/\partial\boldsymbol{\beta} = \mathbf{0}$ for some $\hat{\boldsymbol{\rho}}$ and $\hat{\mu}$. Multiplying both sides of $\partial\mathcal{L}/\partial q_i = 1/\hat{q}_i + n\hat{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\theta}}) - \hat{\mu} = 0$ by $\hat{q}_i$ and summing over i, the constraints in (4) lead to $\hat{\mu} = n$, which, combined with $\partial\mathcal{L}/\partial q_i = 0$ yields $\hat{q}_i = 1/[n\{1 - \hat{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\theta}})\}]$. Then $\partial\mathcal{L}/\partial\boldsymbol{\beta} = \mathbf{0}$ gives (5) and the constraint $\sum_{i=1}^n q_i\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\theta}}) = \mathbf{0}$ gives (6).

Regularity conditions

Regularity Conditions 1. (1) The parameter spaces $\mathcal{B}$ for $\boldsymbol{\beta}$ and Θ for $\boldsymbol{\theta}$ are compact. $\boldsymbol{\beta}_0$ is an interior point of $\mathcal{B}$ and $\boldsymbol{\theta}^*$ is an interior point of Θ. (2) $E[\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})]$ is uniquely maximized at $\boldsymbol{\beta}_0$ and $E[\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta})] = \mathbf{0}$ has a unique solution $\boldsymbol{\theta}^*$. (3) $E[\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{B}\times\Theta} \|\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta})\|^{\alpha}] < \infty$ for some $\alpha > 2$. (4) $E[\partial^2 \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T]$ and $E\{\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*)\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*)^T\}$ are non-singular. (5) $\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{B}\times\Theta} n^{-1/2}\sum_{i=1}^n\{\boldsymbol{l}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) - E[\boldsymbol{l}(\boldsymbol{\beta}, \boldsymbol{\theta})]\} = O_p(1)$ for $\boldsymbol{l}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ and $\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta})$. (6) $\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}, \boldsymbol{\theta})$ is continuously differentiable, $E[\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{B}\times\Theta} \|\partial\boldsymbol{u}(\boldsymbol{\beta}, \boldsymbol{\theta})/\partial\boldsymbol{\beta}\|] < \infty$, and $E[\sup_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{B}\times\Theta} \|\partial\boldsymbol{u}(\boldsymbol{\beta}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\|] < \infty$. (7) $\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ is twice continuously differentiable and $E[\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\partial\boldsymbol{s}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\|] < \infty$.

Regularity Conditions 2. (1) The parameter space $\mathcal{B}$ for $\boldsymbol{\beta}$ is compact. $\boldsymbol{\beta}_0$ is an interior point of $\mathcal{B}$. (2) $E[\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})]$ is uniquely maximized at $\boldsymbol{\beta}_0$. (3) $E[\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\|^{\alpha}] < \infty$ for some $\alpha > 2$. (4) $E[\partial^2 \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T]$ and $E\{\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}_0)^T\}$ are non-singular. (5) $\sup_{\boldsymbol{\beta}\in\mathcal{B}} n^{-1/2}\sum_{i=1}^n\{\boldsymbol{l}_i(\boldsymbol{\beta}) - E[\boldsymbol{l}(\boldsymbol{\beta})]\} = O_p(1)$ for $\boldsymbol{l}(\boldsymbol{\beta}) = \log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ and $\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$. (6) $\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ is continuously differentiable and $E[\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\partial\boldsymbol{u}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\|] < \infty$. (7) $\log f(Y|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ is twice continuously differentiable and $E[\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\partial\boldsymbol{s}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\|] < \infty$.

Proof of Theorem 1

For (i), because $E\{\boldsymbol{u}(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*)\} = \mathbf{0}$, an application of the M-estimator theory (e.g., van der Vaart 1998) to (5) and (6) leads to $(\hat{\boldsymbol{\beta}}_{el1}, \hat{\boldsymbol{\rho}}) \xrightarrow{p} (\boldsymbol{\beta}_0, \mathbf{0})$ as $n \to \infty$. Thus $\hat{\boldsymbol{\beta}}_{el1} \xrightarrow{p} \boldsymbol{\beta}_0$.

For (ii), applying the mean-value theorem to (5) and (6) around $(\boldsymbol{\beta}_0, \boldsymbol{\rho} = \mathbf{0})$ leads to

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \boldsymbol{s}_i(\boldsymbol{\beta}_0) \\ \boldsymbol{u}_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}) \end{pmatrix} + \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \frac{\partial\boldsymbol{s}_i(\bar{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}, & \frac{\partial\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1},\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\beta}^{\mathrm{T}}}{1-\hat{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1},\hat{\boldsymbol{\theta}})} \\ \frac{\partial\boldsymbol{u}_i(\bar{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\beta}}{1-\bar{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1},\hat{\boldsymbol{\theta}})}, & \frac{\boldsymbol{u}_i(\bar{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}})\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1},\hat{\boldsymbol{\theta}})^{\mathrm{T}}}{\{1-\bar{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{u}_i(\hat{\boldsymbol{\beta}}_{el1},\hat{\boldsymbol{\theta}})\}^2} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{el1} - \boldsymbol{\beta}_0 \\ \hat{\boldsymbol{\rho}} \end{pmatrix},$$

where $\bar{\boldsymbol{\beta}}$ is some value between $\hat{\boldsymbol{\beta}}_{el1}$ and $\boldsymbol{\beta}_0$ and $\bar{\boldsymbol{\rho}}$ is some value between $\hat{\boldsymbol{\rho}}$ and $\mathbf{0}$. Then we have

$$\begin{aligned}
&\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\beta}}_{el1} - \boldsymbol{\beta}_0 \\ \hat{\boldsymbol{\rho}} \end{pmatrix} \\
&= -\begin{pmatrix} -\boldsymbol{S}, & \boldsymbol{G}_1^{\mathrm{T}} \\ \boldsymbol{G}_1, & \boldsymbol{U}_1 \end{pmatrix}^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} \boldsymbol{s}_i(\boldsymbol{\beta}_0) \\ \boldsymbol{u}_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\theta}}) \end{pmatrix} + o_p(1) \\
&= -\begin{pmatrix} -\boldsymbol{S}, & \boldsymbol{G}_1^{\mathrm{T}} \\ \boldsymbol{G}_1, & \boldsymbol{U}_1 \end{pmatrix}^{-1} \left\{ \frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} \boldsymbol{s}_i(\boldsymbol{\beta}_0) \\ \boldsymbol{u}_i(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*) \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \left\{\frac{1}{n}\sum_{i=1}^n \frac{\partial\boldsymbol{u}_i(\boldsymbol{\beta}_0,\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\right\}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \end{pmatrix} \right\} + o_p(1) \\
&= -\begin{pmatrix} -\boldsymbol{S}, & \boldsymbol{G}_1^{\mathrm{T}} \\ \boldsymbol{G}_1, & \boldsymbol{U}_1 \end{pmatrix}^{-1} \left\{ \frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} \boldsymbol{s}_i(\boldsymbol{\beta}_0) \\ \boldsymbol{u}_i(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*) \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{Q}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \end{pmatrix} \right\} + o_p(1)
\end{aligned}$$

where $\boldsymbol{Q} \equiv E\{\partial\boldsymbol{u}(\boldsymbol{\beta}_0, \boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\}$ and $\bar{\boldsymbol{\theta}}$ is some value between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$. On the other hand, since $\hat{\boldsymbol{\theta}}_{(j)}$ solves (2) and we assume $\hat{p}_{(j)} = P(Y = 1 \mid \boldsymbol{X}_{(j)})$, it is seen that $\hat{\boldsymbol{\theta}}$ actually solves $\sum_{i=1}^n \tilde{E}_i\{\boldsymbol{u}(\boldsymbol{\beta}_0, \boldsymbol{\theta})\} = \mathbf{0}$. Then some algebra shows that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) = -\boldsymbol{Q}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{E}_i\{\boldsymbol{u}(\boldsymbol{\beta}_0,\boldsymbol{\theta}^*)\} + o_p(1).$$

Therefore we have

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{el1}-\boldsymbol{\beta}_0\\ \hat{\boldsymbol{\rho}}\end{pmatrix} = -\begin{pmatrix}-\boldsymbol{S},\, \boldsymbol{G}_1^{\mathrm{T}}\\ \boldsymbol{G}_1,\, \boldsymbol{U}_1\end{pmatrix}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\boldsymbol{s}_i(\boldsymbol{\beta}_0)\\ \boldsymbol{u}_i(\boldsymbol{\beta}_0,\boldsymbol{\theta}^*)-\tilde{E}_i\{\boldsymbol{u}(\boldsymbol{\beta}_0,\boldsymbol{\theta}^*)\}\end{pmatrix} + o_p(1).$$

Some calculation leads to

$$\mathrm{Var}\begin{pmatrix}\boldsymbol{s}(\boldsymbol{\beta}_0)\\ \boldsymbol{u}(\boldsymbol{\beta}_0,\boldsymbol{\theta}^*)-\tilde{E}\{\boldsymbol{u}(\boldsymbol{\beta}_0,\boldsymbol{\theta}^*)\}\end{pmatrix} = \begin{pmatrix}\boldsymbol{S} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Omega}\end{pmatrix},$$

and thus from the central limit theorem we have

$$\mathrm{asyVar}\left\{\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{el1}-\boldsymbol{\beta}_0\\ \hat{\boldsymbol{\rho}}\end{pmatrix}\right\} = \begin{pmatrix}-\boldsymbol{S},\, \boldsymbol{G}_1^{\mathrm{T}}\\ \boldsymbol{G}_1,\, \boldsymbol{U}_1\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{S} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Omega}\end{pmatrix}\begin{pmatrix}-\boldsymbol{S},\, \boldsymbol{G}_1^{\mathrm{T}}\\ \boldsymbol{G}_1,\, \boldsymbol{U}_1\end{pmatrix}^{-1}.$$

Some further calculation then yields (8).

BIBLIOGRAPHY

1 D. P. Ankerst, J. Hoefler, S. Bock, P. J. Goodman, A. Vickers, J. Hernandez, L. J. Sokoll, M. G. Sanda, J. T. Wei, R. J. Leach, and I. M. Thompson. The prostate cancer prevention trial risk calculator 2.0 for the prediction of low- versus high-grade prostate cancer. *Urology*, 83:1362–1367, 2014.

2 D. P. Ankerst, J. Straubinger, K. Selig, L. Guerrios, A. D. Hoedt, J. Hernandez, M. A. Liss, and et al. A contemporary prostate biopsy risk calculator based on multiple heterogeneous cohorts. *European Urology*, 74:197–203, 2018.

3 P. S. Boonstra, J. M. G. Taylor, and B. Mukherjee. Incorporating auxiliary information for improved prediction in high-dimensional datasets: an ensemble of shrinkage approaches. *Biostatistics*, 14:259–272, 2013.

4 L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

5 N. Chatterjee, Y. H. Chen, P. Maas, and R. J. Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111:107–117, 2016.

6 S. Chaudhuri, M. S. Handcock, and M. S. Rendall. Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *Journal of the Royal Statistical Society Series B*, 70:311–328, 2008.

7 J. Chen and J. Qin. Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80:107–116, 1993.

8 J. Chen, R. R. Sitter, and C. Wu. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89:230–237, 2002.

9 S. Chen and J. K. Kim. Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 24:335–355, 2014.

10 Y. H. Chen and H. Chen. A unified approach to regression analysis under double sampling design. *Journal of the Royal Statistical Society, Series B*, 62:449–460, 2000.

11 W. Cheng, J. M. G. Taylor, T. Gu, S. A. Tomlins, and B. Mukherjee. Informing a risk prediction model for binary outcomes with external coefficient information. *Journal of the Royal Statistical Society: Series C*, 68:121–139, 2019.

12 W. Cheng, J. M. G. Taylor, P. S. Vokonas, S. K. Park, and B. Mukherjee. Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in medicine*, 37:1515–1530, 2018.

13 J. Deville and C. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

14 J. P. Estes, B. Mukherjee, and J. M. G. Taylor. Empirical bayes estimation and prediction using summary-level information from external big data sources adjusting for violations of transportability. *Statistics in Biosciences*, 10:568–586, 2018.

15 S. Grill, M. Fallah, R. Leach, I. Thompson, K. Hemminki, and D. Ankerst. A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation. *Journal of Clinical Epidemiology*, 68:563–573, 2015.

16 T. Gu, J. M. G. Taylor, W. Cheng, and B. Mukherjee. Synthetic data method to incorporate external information into a current study. *Canadian Journal of Statistics*, 47:580–603, 2019.

17 P. Han and J. F. Lawless. Discussion of "constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources". *Journal of the American Statistical Association*, 111:118–121, 2016.

18 P. Han and J. F. Lawless. Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica*, 29:1321–1342, 2019.

19 P. Han and L. Wang. Estimation with missing data: beyond double robustness. *Biometrika*, 100:417–430, 2013.

20 T. Hayfield and J.S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

21 C-Y. Huang and J. Qin. A unified approach for synthesizing population-level covariate effect information in semiparametric estimation with survival data. *Statistics in Medicine*, 39:1573–1590, 2020.

22 C-Y. Huang, J. Qin, and H-T. Tsai. Efficient estimation of the cox model with auxiliary subgroup survival information. *Journal of the American Statistical Association*, 111:787–799, 2016.

23 G. W. Imbens and T. Lancaster. Combining micro and macro data in microeconometric models. *Review of Economic Studies*, 61:655–680, 1994.

24 J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, 2007.

25 R. W. Klein and R. H. Spady. An efficient semiparametric estimator for binary response models. *Econometrica*, 61:387–421, 1993.

26 E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer New York, 2003.

27 A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

28 T. Lumley, P. A. Shaw, and J. Y. Dai. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79:200–220, 2011.

29 S. Mocellin, J. F. Thompson, S. Pasquali, M. C. Montesco, P. Pilati, D. Nitti, R. P. Saw, R. A. Scolyer, J. R. Stretch, and C. R. Rossi. Sentinel node status prediction by four statistical models: results from a large bi-institutional series (n = 1132). *Annals of Surgery*, 250:964–969, 2009.

30 A. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249, 1988.

31 A. Owen. *Empirical Likelihood*. Chapman & Hall/CRC Press, New York, 2001.

32 J. Qin. Combining parametric and empirical likelihoods. *Biometrika*, 87:484–490, 2000.

33 J. Qin. *Biased sampling, over-identified parameter problems and beyond*. Springer Nature, 2017.

34 J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22:300–325, 1994.

35 J. Qin, H. Zhang, P. Li, D. Albanes, and K. Yu. Using covariate specific disease prevalence information to increase the power of case-control study. *Biometrika*, 102:169–180, 2015.

36 Y. Sheng, Y. Sun, C.-Y. Huang, and M. Kim. Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics*, page 10.1111/biom.13429, 2021.

37 R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, 2021.

38 I. M. Thompson, D. P. Ankerst, C. Chi, P. J. Goodman, C. M. Tangen, M. S. Lucia, Z. Feng, H. L. Parnes, and C. A. Coltman. Assessing prostate cancer risk: results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98:529–534, 2006.

39 S. A. Tomlins, J. R. Day, R. J. Lonigro, D. H. Hovelson, J. Siddiqui, L. P. Kunju, R. L. Dunn, S. Meyer, P. Hodge, J. Groskopf, J. T. Wei, and A. M. Chinnaiyan. Urine tmprss2:erg plus pca3 for individualized prostate cancer risk assessment. *European Urology*, 70:45–53, 2016.

40 A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

TABLE 1: Simulation results for comparisons between methods with and without using working models based on $n = 400$ and 1000 replications. emp: empirical standard error. est: mean of estimated standard errors over 1000 replications. cov: percentage over 1000 replications that the 95% confidence intervals constructed based on asymptotic distributions cover the true value.

		Method 1: with working models					Method 2: without working models				
		β_c	β_{X_1}	β_{X_2}	β_Z	β_{X_2Z}	β_c	β_{X_1}	β_{X_2}	β_Z	β_{X_2Z}
MLE	bias	0.003	-0.005	-0.014	0.006	0.015	0.003	-0.005	-0.014	0.006	0.015
	emp	0.123	0.137	0.139	0.140	0.143	0.123	0.137	0.139	0.140	0.143
	est	0.120	0.135	0.139	0.139	0.141	0.120	0.135	0.139	0.139	0.141
	cov	95.0	94.8	95.1	95.6	94.8	95.0	94.8	95.1	95.6	94.8
EL-1	bias	-0.001	0.018	-0.014	0.006	0.015	-0.001	0.019	-0.014	0.006	0.015
	emp	0.052	0.071	0.139	0.140	0.143	0.052	0.071	0.139	0.140	0.143
	est	0.052	0.073	0.138	0.138	0.140	0.052	0.073	0.138	0.138	0.140
	cov	95.4	94.2	94.9	95.3	94.8	95.4	94.0	95.0	95.3	94.7
EL-2	bias	-0.000	0.020	-0.035	0.006	0.016	0.000	0.020	-0.035	0.006	0.015
	emp	0.049	0.059	0.061	0.140	0.142	0.049	0.059	0.061	0.140	0.142
	est	0.048	0.062	0.059	0.138	0.140	0.048	0.061	0.059	0.138	0.140
	cov	95.4	92.3	94.4	95.4	95.0	95.5	92.0	94.1	95.4	95.0
EL-3	bias	-0.006	-0.002	-0.017	0.006	0.015	-0.005	-0.002	-0.015	0.006	0.013
	emp	0.064	0.072	0.076	0.140	0.143	0.063	0.071	0.075	0.140	0.143
	est	0.061	0.074	0.076	0.138	0.140	0.060	0.073	0.074	0.138	0.140
	cov	95.3	95.5	95.3	95.2	94.9	95.5	94.4	94.7	95.3	94.7
EL-12	bias	0.003	0.021	-0.036	0.005	0.015	0.000	0.022	-0.036	0.005	0.015
	emp	0.050	0.060	0.061	0.140	0.142	0.049	0.059	0.061	0.140	0.142
	est	0.049	0.062	0.060	0.138	0.140	0.048	0.061	0.059	0.138	0.140
	cov	94.9	91.7	94.1	95.4	95.4	95.8	91.5	93.8	95.4	95.3
EL-13	bias	0.002	0.004	-0.011	0.005	0.009	-0.001	0.017	-0.017	0.006	0.014
	emp	0.074	0.072	0.082	0.140	0.143	0.049	0.062	0.075	0.140	0.143
	est	0.073	0.074	0.085	0.138	0.140	0.049	0.062	0.074	0.138	0.140
	cov	94.6	94.8	96.1	95.2	95.1	95.5	92.5	94.5	95.3	94.6
EL-23	bias	-0.001	-0.003	-0.023	0.024	0.026	-0.003	0.016	-0.032	0.005	0.022
	emp	0.072	0.072	0.078	0.146	0.159	0.048	0.060	0.060	0.141	0.141
	est	0.068	0.072	0.079	0.141	0.149	0.047	0.061	0.058	0.138	0.138
	cov	93.6	95.1	95.0	94.9	92.7	95.5	92.3	94.3	94.8	94.5
EL-123	bias	0.003	-0.002	-0.019	0.027	0.020	-0.003	0.018	-0.033	0.005	0.022
	emp	0.078	0.075	0.080	0.149	0.166	0.048	0.060	0.059	0.141	0.141
	est	0.077	0.073	0.080	0.142	0.154	0.047	0.060	0.058	0.138	0.138
	cov	93.6	94.6	95.2	93.9	92.8	95.9	91.7	94.2	94.9	94.6

TABLE 2: Simulation results for comparisons between methods with and without using working models based on $n = 1000$ and 1000 replications. emp: empirical standard error. est: mean of estimated standard errors over 1000 replications. cov: percentage over 1000 replications that the 95% confidence intervals constructed based on asymptotic distributions cover the true value.

		Method 1: with working models					Method 2: without working models				
		β_c	β_{X_1}	β_{X_2}	β_Z	β_{X_2Z}	β_c	β_{X_1}	β_{X_2}	β_Z	β_{X_2Z}
MLE	bias	0.004	-0.004	-0.003	0.003	0.008	0.004	-0.004	-0.003	0.003	0.008
	emp	0.075	0.087	0.087	0.086	0.091	0.075	0.087	0.087	0.086	0.091
	est	0.076	0.085	0.087	0.087	0.088	0.076	0.085	0.087	0.087	0.088
	cov	95.0	94.0	94.6	95.0	94.5	95.0	94.0	94.6	95.0	94.5
EL-1	bias	0.005	0.008	-0.003	0.003	0.008	0.005	0.008	-0.003	0.003	0.007
	emp	0.031	0.046	0.087	0.086	0.091	0.031	0.046	0.087	0.086	0.091
	est	0.032	0.045	0.087	0.087	0.088	0.032	0.045	0.087	0.087	0.088
	cov	95.7	93.8	94.6	95.1	94.6	95.7	93.8	94.6	95.1	94.6
EL-2	bias	0.005	0.016	-0.012	0.003	0.008	0.005	0.017	-0.012	0.003	0.008
	emp	0.029	0.039	0.035	0.086	0.090	0.029	0.039	0.035	0.086	0.090
	est	0.030	0.038	0.036	0.087	0.087	0.030	0.038	0.036	0.087	0.087
	cov	95.8	89.7	95.2	94.8	94.2	95.6	89.6	95.2	94.8	94.2
EL-3	bias	-0.001	-0.002	0.007	0.003	0.007	-0.001	-0.002	0.006	0.003	0.007
	emp	0.038	0.047	0.045	0.086	0.091	0.037	0.046	0.044	0.086	0.091
	est	0.038	0.046	0.047	0.087	0.087	0.037	0.045	0.046	0.087	0.087
	cov	95.5	94.6	95.2	94.9	94.2	95.3	93.9	94.7	95.0	94.3
EL-12	bias	0.002	0.016	-0.012	0.003	0.008	0.005	0.016	-0.012	0.003	0.007
	emp	0.030	0.039	0.035	0.086	0.090	0.029	0.039	0.035	0.086	0.090
	est	0.030	0.038	0.036	0.087	0.087	0.030	0.038	0.036	0.087	0.087
	cov	95.2	90.6	95.4	94.7	94.3	95.4	89.7	95.4	94.8	94.3
EL-13	bias	0.001	0.001	0.010	0.002	0.005	0.004	0.005	0.006	0.003	0.007
	emp	0.045	0.047	0.051	0.086	0.091	0.030	0.040	0.045	0.086	0.091
	est	0.046	0.046	0.052	0.087	0.087	0.030	0.039	0.046	0.087	0.087
	cov	95.7	94.2	93.6	95.0	93.8	95.8	93.6	94.7	95.1	94.3
EL-23	bias	-0.001	-0.001	0.001	0.009	0.028	0.001	0.014	-0.010	0.001	0.016
	emp	0.043	0.046	0.050	0.087	0.096	0.028	0.039	0.034	0.086	0.089
	est	0.043	0.044	0.049	0.088	0.093	0.029	0.038	0.036	0.087	0.086
	cov	94.7	93.1	94.5	95.1	94.2	95.6	90.5	96.0	94.8	93.9
EL-123	bias	-0.003	0.000	0.004	0.009	0.027	0.002	0.014	-0.009	0.001	0.016
	emp	0.046	0.047	0.050	0.088	0.098	0.028	0.039	0.034	0.086	0.089
	est	0.049	0.044	0.049	0.088	0.096	0.029	0.038	0.036	0.087	0.086
	cov	95.0	93.1	93.6	94.7	94.9	95.2	90.4	95.8	94.9	94.0

TABLE 3: Simulation results for different choices of $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$ and comparisons to Gu et al. (2019) based on 1000 replications. The proposed method here is Method 2 without using working models. emp: empirical standard error. GTCM-1,2 or 3: the method in Gu et al. (2019) using Calculator 1, 2, or 3, respectively. EL2-j-1,2 or 3: the proposed Method 2 without working models using Calculator j and different specifications of $\boldsymbol{d}_{(j)}(\boldsymbol{X}_{(j)})$.

		$n = 400$					$n = 1000$				
		β_c	β_{X_1}	β_{X_2}	β_Z	$\beta_{X_2 Z}$	β_c	β_{X_1}	β_{X_2}	β_Z	$\beta_{X_2 Z}$
MLE	bias	0.010	-0.005	-0.013	0.006	0.016	0.006	-0.006	-0.003	0.006	0.002
	emp	0.122	0.136	0.144	0.140	0.139	0.076	0.084	0.087	0.083	0.085
GTCM-1	bias	0.036	0.015	0.008	0.044	-0.087	0.040	0.009	0.013	0.047	-0.093
	emp	0.059	0.081	0.138	0.134	0.096	0.037	0.049	0.084	0.082	0.060
EL2-1-1	bias	-0.000	0.018	-0.013	0.006	0.016	0.007	0.009	-0.003	0.006	0.002
	emp	0.052	0.076	0.144	0.140	0.139	0.033	0.045	0.087	0.083	0.085
EL2-1-2	bias	-0.000	0.018	-0.013	0.006	0.016	0.007	0.009	-0.002	0.006	0.003
	emp	0.050	0.075	0.144	0.141	0.130	0.031	0.044	0.087	0.084	0.078
GTCM-2	bias	0.028	-0.017	0.087	0.044	-0.089	0.032	-0.016	0.098	0.048	-0.093
	emp	0.054	0.071	0.076	0.133	0.095	0.035	0.042	0.047	0.081	0.060
EL2-2-1	bias	0.000	0.019	-0.033	0.006	0.015	0.006	0.018	-0.012	0.006	0.002
	emp	0.048	0.065	0.062	0.141	0.138	0.030	0.037	0.035	0.083	0.085
EL2-2-2	bias	-0.009	0.011	-0.034	0.008	0.049	-0.002	0.011	-0.013	0.008	0.028
	emp	0.042	0.060	0.062	0.141	0.104	0.026	0.034	0.036	0.084	0.061
EL2-2-3	bias	-0.010	0.011	-0.034	0.008	0.051	-0.002	0.011	-0.013	0.009	0.029
	emp	0.042	0.060	0.062	0.142	0.104	0.026	0.034	0.036	0.084	0.061
GTCM-3	bias	0.031	0.001	0.001	0.044	-0.087	0.034	-0.001	0.018	0.047	-0.093
	emp	0.064	0.081	0.081	0.134	0.095	0.041	0.049	0.049	0.082	0.060
EL2-3-1	bias	-0.005	0.000	-0.012	0.006	0.014	0.000	-0.003	0.007	0.006	0.001
	emp	0.059	0.076	0.075	0.140	0.139	0.039	0.044	0.045	0.083	0.085
EL2-3-2	bias	-0.008	-0.001	-0.013	0.006	0.024	-0.002	-0.006	0.005	0.007	0.013
	emp	0.057	0.073	0.073	0.140	0.114	0.036	0.043	0.044	0.083	0.069
EL2-3-3	bias	-0.010	-0.001	-0.017	0.008	0.032	0.000	-0.008	0.012	0.005	0.005
	emp	0.057	0.073	0.070	0.140	0.111	0.036	0.043	0.040	0.084	0.068

TABLE 4: Prostate cancer study analysis results based on logistic regression ($n = 1014$). est: estimated value. s.e.: standard error. p: p-value. DRE: a binary indicator of an abnormal digital rectal exam. biopsy: a binary indicator of prior negative biopsy. race: a binary indicator of African Ancestry. history: a binary indicator of first-degree family history of prostate cancer. PCA3: $\log_2(\text{PCA3} + 1)$. T2ERG: dichotomized TMPRSS2:ERG split at the median

	Without calculators			With PBCG calculator			With PBCG-2 calculator			With both calculators		
	est	s.e.	p	est	s.e.	p	est	s.e.	p	est	s.e.	p
Intercept	-7.421	0.769	< 0.001	-6.802	0.227	< 0.001	-7.276	0.315	< 0.001	-6.826	0.229	< 0.001
$\log_2(\text{PSA})$	0.595	0.119	< 0.001	0.756	0.033	< 0.001	0.774	0.051	< 0.001	0.751	0.033	< 0.001
age	0.037	0.011	0.001	0.027	0.004	< 0.001	0.036	0.005	< 0.001	0.028	0.004	< 0.001
DRE	0.635	0.210	0.003	0.769	0.057	< 0.001	0.631	0.199	0.001	0.762	0.057	< 0.001
biopsy	-0.895	0.245	< 0.001	-0.938	0.058	< 0.001	-0.901	0.216	< 0.001	-0.943	0.058	< 0.001
race	0.090	0.341	0.793	0.192	0.103	0.062	-0.050	0.120	0.675	0.181	0.103	0.079
history	0.248	0.209	0.235	0.413	0.049	< 0.001	0.249	0.191	0.193	0.414	0.049	< 0.001
PCA3	0.347	0.062	< 0.001	0.347	0.057	< 0.001	0.347	0.056	< 0.001	0.345	0.057	< 0.001
T2ERG	0.565	0.183	0.002	0.567	0.166	0.001	0.568	0.165	0.001	0.582	0.166	< 0.001