frontiers | Frontiers in Nutrition

TYPE Original Research
PUBLISHED 02 August 2023
DOI 10.3389/fnut.2023.1215768

Regression calibration utilizing biomarkers developed from high-dimensional metabolites

Yiwen Zhang[1†], Ran Dai[2*†], Ying Huang[3], Ross L. Prentice[3] and Cheng Zheng[2*]

[1]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, United States, [2]Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE, United States, [3]Public Health Science Division, Fred Hutchinson Cancer Center, Seattle, WA, United States

OPEN ACCESS

EDITED BY
Alessandra Durazzo,
Council for Agricultural Research and Economics, Italy

REVIEWED BY
Mário Sousa-Pimenta,
Instituto Português de Oncologia, Portugal
Dimitris Panagoulias,
University of Piraeus, Greece

*CORRESPONDENCE
Ran Dai
✉ ran.dai@unmc.edu
Cheng Zheng
✉ cheng.zheng@unmc.edu

†These authors have contributed equally to this work and share first authorship

RECEIVED 02 May 2023
ACCEPTED 17 July 2023
PUBLISHED 02 August 2023

CITATION
Zhang Y, Dai R, Huang Y, Prentice RL and Zheng C (2023) Regression calibration utilizing biomarkers developed from high-dimensional metabolites. *Front. Nutr.* 10:1215768.
doi: 10.3389/fnut.2023.1215768

Addressing systematic measurement errors in self-reported data is a critical challenge in association studies of dietary intake and chronic disease risk. The regression calibration method has been utilized for error correction when an objectively measured biomarker is available; however, biomarkers for only a few dietary components have been developed. This paper proposes to use high-dimensional objective measurements to construct biomarkers for many more dietary components and to estimate the diet disease associations. It also discusses the challenges in variance estimation in high-dimensional regression methods and presents a variety of techniques to address this issue, including cross-validation, degrees-of-freedom corrected estimators, and refitted cross-validation (RCV). Extensive simulation is performed to study the finite sample performance of the proposed estimators. The proposed method is applied to the Women's Health Initiative cohort data to examine the associations between the sodium/potassium intake ratio and the total cardiovascular disease.

KEYWORDS

measurement error, regression calibration, feeding study, biomarker, high-dimensional data

1. Introduction

The field of nutritional epidemiology plays a crucial role in understanding the impact of dietary patterns on human health. The ongoing exploration of associations between dietary components and chronic disease risks continually uncovers valuable insights. For instance, the well-established link between obesity and cancer risk (1) serves as a testament to the significance of this research. In order to effectively prevent and control chronic diseases, it is imperative to acquire detailed information on how key energy balance factors associate with the risks of major chronic illnesses [World Cancer Research Fund/American Institute for Cancer Research (2)]. Investigating the complex working mechanisms of these energy balance factors necessitates a comprehensive examination of the connections between multiple dietary components and disease risks. Establishing such associations, however, is far from simple. A major challenge stems from biases in dietary assessment, which are notoriously difficult to address (3). Strong evidence (4) suggests that the misreporting of dietary energy intake is associated with individual characteristics, such as body mass index (BMI). These systematic measurement errors result in estimation biases that cannot be automatically rectified (5). Moreover, correcting measurement errors becomes increasingly challenging when attempting to model dietary components jointly in the context of their relationships with chronic diseases.

Correcting measurement errors has been an important subject in statistical methodology development, greatly influencing nutritional studies (6). Various strategies have been developed to address these errors (7–15). One notable method, regression calibration, is particularly useful for handling covariate-dependent measurement errors and offers ease of implementation (16). Studies within the Women's Health Initiative (WHI) have demonstrated the effectiveness of joint regression calibration approaches in addressing measurement errors when objective biomarkers are available for all modeled dietary intakes (4, 17–19). These biomarkers inform calibration equations for self-reported measurements of exposure variables, which then provide calibrated intake estimates to better assess associations between dietary exposures and disease risks.

There is a significant research gap in generating reliable calibrated estimates for numerous nutritional and physical activity variables using single objective measurements. Consequently, regression models with multiple predictors have been developed from feeding studies to obtain calibrated estimates (4, 20). For instance, in the WHI Nutrition and Physical Activity Assessment Study (NPAAS), a regression-based biomarker has been established for a single dietary component or energy balance factors (21). To address the systematic measurement errors in self-reported food frequency questionnaire (FFQ) data from a large cohort, blood and urine measurements were collected for a subgroup, while a feeding study (NPAAS-FS) was conducted on another smaller subgroup where both blood and urine measurements and assessed dietary intake information were collected. This novel feeding study design aimed to improve the accuracy of capturing measurement errors in the FFQ (21). However, there are some challenges in building the regression calibration method, as the classical measurement error assumption will be violated by the feeding study-based biomarker development procedure, which regresses the consumed nutrient on blood and urine measurements and personal characteristics. This issue arises because the residual of the regression model is independent of the predicted value instead of the actual one. Ignoring this violation results in biased estimates of the calibrated dietary intake and the diet-disease association due to Berkson-type errors (22). When developing biomarkers for objectively measured variables with low dimensions, new calibration methods have been developed to account for Berkson-type errors in association studies of univariate nutritional variables (20). Zhang et al. (23) have extended this approach to multivariate nutritional variables, providing consistent estimators for disease associations of a single dietary component and valid confidence intervals for disease association parameters under rare disease settings. Nevertheless, for some macronutrient intakes, suitable biomarkers cannot be developed from low-dimensional measurements. High-dimensional metabolites offer an opportunity to establish valid biomarkers, but it remains an open question on how to obtain valid inferences for such biomarkers.

In this paper, we concentrate on high-dimensional objective measurements for a univariate exposure of interest, where the sample size is smaller than the dimension of the variables, in constructing a biomarker model. High-dimensional variable selection constitutes a significant portion of the rapidly advancing statistical frontiers today. Over the past few decades, numerous studies have been dedicated to understanding the performance of various variable selection techniques. Frank and Friedman (24) first proposed a technique called bridge regression. Breiman (25) introduced the nonnegative garrote for shrinkage estimation and variable selection. Lasso, an l-1 regularized least squares method, was studied and introduced by (26) for variable selection. Nonconcave penalized likelihood estimators, such as smoothly clipped absolute deviation (SCAD), were proposed by (27) and (28). Efron et al. (29) presented the least angle regression for variable selection and introduced the LARS algorithm. Zou and Li (30) proposed one-step sparse estimates for nonconcave penalized likelihood models and introduced the local linear approximation algorithm for optimizing non-concave penalized likelihoods.

Building a biomarker model with high-dimensional sparse data requires predictive performance that can effectively address the challenges associated with such data. One issue that arises when working with high-dimensional models is the collinearity among covariates, which can result in spurious correlations between variables (31). Numerous researchers have explored penalized regression techniques, such as Lasso and SCAD, to handle high-dimensional sparse data. Alternatively, variable selection can also be done by ranking predictive powers using random forest (RF) (32). Variance estimation in high dimensional models presents its own challenges, due to factors such as collinearity among covariates and the presence of spurious correlations. A variety of techniques have been proposed to address the issue of variance estimation in high-dimensional regression methods. Cross-validation (CV), a popular resampling technique, has been widely applied to assess the performance of different models and obtain unbiased variance estimates (33). The bootstrap, another resampling method, has been employed to estimate the variability of regression parameters (34). The degrees-of-freedom corrected estimators, such as the generalized degrees of freedom and the effective degrees of freedom, provide better error variance estimates by accounting for the complexity of the models (35, 36). The refitted cross-validation (RCV) method is a modification of the standard cross-validation procedure that improves the estimation of error variance in high-dimensional regression (37).

The remaining of the paper is organized as follows. In Section 2, we introduce the framework of the present study and the notation. In Section 3, we introduce different methods and detail variance estimation procedures. In Section 4, we conduct extensive simulations to evaluate the finite sample performance of our proposed estimators. In Section 5, we apply our method to the WHI data to estimate the effect of macronutrient intakes on the risk of various chronic diseases. Finally, in Section 6, we present our conclusions and discussions.

2. Framework and notation

We aim to investigate the correlation between a particular type of dietary intake $Z \in \mathbb{R}$ [such as the (log-transformed) ratio of dietary sodium to potassium] and the timeframe, T, to the emergence of a specific chronic illness. Nevertheless, rather than directly observing Z, we only gather information on self-reported dietary intake $Q \in \mathbb{R}$, which may deviate from Z depending on individual characteristics:

$$Q = (1, Z, \boldsymbol{V}^\top)\boldsymbol{a} + \epsilon_q, \tag{1}$$

Here, $\boldsymbol{a} \in \mathbb{R}^{(2+q)}$ is an unidentified parameter vector, and ϵ_q is a random error with a mean of 0 that is independent of Z and $\boldsymbol{V}$. We also take into account potential confounding factors, referred to as personal characteristics $\boldsymbol{V} \in \mathbb{R}^q$ where q is the number of covariates. To model the hazard of the response, we employ a Cox model:

$$\lambda(t|Z, \boldsymbol{V}, Q) = \lambda(t|Z, \boldsymbol{V}) = \lambda_0(t)\exp((Z, \boldsymbol{V}^\top)\boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\theta} = (\theta_z, \boldsymbol{\theta}_v^\top)^\top \in \mathbb{R}^{(1+q)}$, θ_z is the parameter we are interested in, and $\lambda_0(t)$ represents a "baseline" hazard function.

In the NPAAS feeding study (NPAAS-FS), we furnish participants' meals with standardized food, which closely mimicking their regular diet, has well-documented nutrient content (21). The true unobservable dietary intake within the 2-week feeding period is denoted as

$$X = Z + \epsilon_x, \text{ where } \epsilon_x \sim \mathcal{N}(0, \sigma_x^2). \tag{3}$$

In our current model, we assume that ϵ_x is independent of Z and $\boldsymbol{V}$. However, condition (3) could be considered somewhat restrictive, given that the design of the feeding study is based on reported long-term dietary intake and not the actual diet. To address this, we have modified this assumption such that the true short-term unobserved diet X does not necessarily need to be centered around Z. Additional specifics can be found in Section 3. One intricate issue related to the feeding study is the measurement errors arising from food packaging. For example, a pack of chips labeled as 100 calories might in reality contain 101 calories. Consequently, the observed short-term dietary intake $\tilde{X}$ during the feeding study can be expressed as $\tilde{X} = X + \tilde{\epsilon}_x$, where $\tilde{\epsilon}_x \sim \mathcal{N}(0, \tilde{\sigma}_x^2)$ is independent of ϵ_x, Z, and $\boldsymbol{V}$.

The study is organized into three stages: the feeding study (Sample 1) for biomarker development, the biomarker sub-study (Sample 2) for calibration equation development, and the association study (Sample 3) using the complete cohort to establish the disease association.

When self-reported intake Q data from feeding study samples is available, the bias of self-reported dietary intake can be directly calibrated (refer to Section 3.4). However, self-reported dietary intake Q is usually not available concurrently in Sample 1. To acquire that data, a long-term feeding study would be necessary, wherein participants report their dietary intake provided over preceding months (e.g., 3 months). Furthermore, the Q value obtained just prior to the feeding period in NPAAS-FS is not collected at the same time as biomarker W, and it might be inappropriately highly correlated with $\tilde{X}$.

As an alternative, we could employ a high-dimensional biomarker $\boldsymbol{W} \in \mathbb{R}^p$, comprised of p blood and urine measurements obtained objectively, as a bridge between $\tilde{X}$ from the feeding study sample and Q from a separate, larger sample. We assume that the blood and urine measurements W are influenced by the short-term diet X, whereas the self-reported questionnaire data are directly impacted by the long-term diet Z. We assume that $\boldsymbol{W}$ is possibly high-dimensional and follows a parametric model:

$$\boldsymbol{W} = [(1, X, \boldsymbol{V}^\top)\boldsymbol{B}]^\top + \epsilon_w,$$

where $\boldsymbol{B} \in \mathbb{R}^{(2+q)\times p}$ is a matrix of unknown parameters and $\epsilon_w \sim \mathcal{N}(0, \sigma_w^2 I_p)$ is independent of $\epsilon_x, \tilde{\epsilon}_x, \epsilon_q, Z, \boldsymbol{V}$ and $\boldsymbol{B}$.

In practical terms, our best option is to utilize the baseline Q gathered at a separate time (for instance, at baseline for Sample 3) for Sample 1. This baseline Q has been effectively used in studies concerning various dietary components [e.g., protein and carbohydrate; (22)]. However, a time gap exists between the data collection for this baseline Q and the timing of $(\tilde{X}, \boldsymbol{V}, \boldsymbol{W}, Z)$ measurements in Sample 1. Consequently, there's a concern that the conditional distribution $(Q|\tilde{X}, \boldsymbol{V}, \boldsymbol{W}, Z)$ in Sample 1 may differ from Samples 2 and 3 for specific dietary components. Even when Q is available, the feeding study's sample size is usually restricted, which could lead to less than optimal efficiency for disease association estimates. In such instances, we consider Q as unavailable in Sample 1 and use $\boldsymbol{W}$ to predict $\tilde{X}$.

The process of estimating the association between Z and T is divided into three stages, each utilizing distinct, non-overlapping samples derived from the same fundamental population: 1. the biomarker creation phase, 2. the calibration phase, and 3. the phase assessing the association. Each stage employs a different sample. The size of the sample used in stage k is denoted as n_k. In Stage 1, there are n_1 samples, and for each individual i, we have access to data $(\tilde{X}_i, \boldsymbol{W}_i^\top, \boldsymbol{V}_i^\top)$ and possibly Q_i; in Stage 2, n_2 samples are available, and for each individual i, we have $(Q_i, \boldsymbol{W}_i^\top, \boldsymbol{V}_i^\top)$; in Stage 3, there are n_3 samples, and for each individual i, we have $(Q_i, \boldsymbol{V}_i^\top)$ and the composite outcome $[T_i^* = T_i \wedge C_i, \Delta_i = I(T_i \leq C_i)]$, where T_i is the time of disease occurrence, and C_i is a potential censoring time. Conventionally, T_i and C_i are assumed to be independent given $(Q_i, \boldsymbol{V}_i^\top)$.

During the first stage, we utilize data from the biomarker creation phase to develop the biomarker. This model can be constructed by regressing the observed short-term dietary intakes $\tilde{X}$ on one of the following:

(i) blood/urine measurements $\boldsymbol{W}$ and personal characteristics $\boldsymbol{V}$;

(ii) blood/urine measurements $\boldsymbol{W}$, self-reported dietary intake Q, and personal characteristics $\boldsymbol{V}$;

(iii) self-reported dietary intake Q and personal characteristics $\boldsymbol{V}$.

As earlier indicated, self-reported dietary intake Q may be deemed unavailable during Stage 1. If that's the case, we treat Q as unavailable and opt for choice (i) in Stage 1. When Q is accessible in Stage 1, choice (ii) might enhance the estimation of $\tilde{X}$. If the biomarker W is not available, option (iii) directly models $\tilde{X}$ based on Q and V, but the effectiveness might be hampered by the limited sample size n_1. In Stage 2, a calibration equation is developed using self-reported log-transformed dietary intake Q and personal characteristics $\boldsymbol{V}$ to predict actual intake X if options (i) or (ii) are implemented in Stage 1. If option (iii) is chosen, Stage 2 is bypassed, as the equation is already established in Stage 1. However, option (i) introduces Berkson-type error, which impacts regression calibration, thus necessitating new methodologies to address this characteristic. In Stage 3, we calibrate the self-reported dietary intake utilizing the Stage 2 calibration equation, conduct disease association analyses with the available data on Q, $\boldsymbol{V}$, and the composite survival outcome (T^{*}, Δ).

In summary, the high-dimensional regression calibration procedure has three stages: biomarker construction, calibration, and estimation. In Stage 1, the relationship between the true dietary intake $\boldsymbol{X}$ and high-dimensional biomarker $\boldsymbol{W}$ is established. If self-reported dietary intakes $\boldsymbol{Q}$ are not available, option (i) can be used. If $\boldsymbol{Q}$ is available in Stage 1, whether or not $\boldsymbol{W}$ is also available, relationships between $\boldsymbol{X}$ and $\boldsymbol{Q}$ can be directly established with option (iii). If both $\boldsymbol{W}$ and $\boldsymbol{Q}$ are available for Stage 1, one of the options from (i), (ii), and (iii) can be used. As discussed, (i) might lead to Berkson type error (38) and (iii) might have low efficiency. For Stage 2, we developed bias correction methods to account for the bias introduced by the Berkson type error. For Stage 3, we can use a multivariate approach to jointly study the associations between multiple dietary components and the disease risks.

3. Methods

we first consider the case where $\tilde{\sigma}_x$ is known. We propose methods to estimate $\tilde{\sigma}_x$ in the discussion section. In the real data analysis where $\tilde{\sigma}_x$ is not available, we vary this parameter to perform sensitivity analysis.

With high-dimensional data on urine measurements ($\boldsymbol{W}$), we first need to obtain estimated coefficients among n_1 subjects in the biomarker discovery sample of the observed short-term dietary intake $\tilde{X}$ on high-dimensional blood and urine measurements ($\boldsymbol{W}$) as well as subject characteristics ($\boldsymbol{V}$). Three different approaches including Lasso, SCAD, and RF are used to conduct variable selection in high-dimensional statistical inference. We will describe each approach explicitly for every method in the following subsections.

3.1. Method 1: the naïve three-step approach with multiple exposure

In the first step, we need to fit a linear regression of $\tilde{X}$ on $\boldsymbol{W}$ and $\boldsymbol{V}$:

$$\mathbb{E}(\tilde{X}|\boldsymbol{W},\boldsymbol{V}) = (1,\boldsymbol{W},\boldsymbol{V})\boldsymbol{\beta}_1.$$

With Lasso approach, the coefficients, $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$, minimize the penalized least squares (PL_{Lasso}) as below:

$$PL_{Lasso} = \sum_{i=1}^{n}(\tilde{X}_i - (1,\boldsymbol{W}_i^{\top},\boldsymbol{V}_i^{\top})\boldsymbol{\beta}_1)^2 + \lambda\sum_{j=1}^{p}|\beta_{1j}|.$$

Lasso performs variable selection by shrinking coefficient estimates toward zero leading to a sparse model. The tuning parameter λ is selected through cross-validation.

With the SCAD approach, a nonconvex penalty is given by:

$$PL_{SCAD}(\beta_{1j}) = \begin{cases} \lambda|\boldsymbol{\beta}_1| \quad if\ |\beta_{1j}| < \lambda \\ 2a\lambda|\boldsymbol{\beta}_1|^2 - 2a\lambda|\boldsymbol{\beta}_1| \quad if\ \lambda < |\beta_{1j}| < a\lambda \\ (a+1)\lambda^2/2 \quad if\ |\beta_{1j}| > a\lambda \end{cases}$$

The first derivatives of $PL_{SCAD}(\beta_{1j})$ is continuous and is given by

$$PL_{SCAD}(\boldsymbol{\beta}_1)' = \lambda\{I(\boldsymbol{\beta}_1 < \lambda) + \frac{(a\lambda - \boldsymbol{\beta}_1)_+}{(a-1)\lambda}I(\boldsymbol{\beta}_1 > \lambda)\}$$

for some $a > 2$ and $\beta_1 > 0$. Similar to Lasso, λ in SCAD is selected through cross-validation based on the smallest mean square error (MSE) whereas a is set to be 3.7 based on simulation results and Bayesian statistical point of view from (27).

Other than penalized regression as we described above, RF is another choice for variable selection. The basic concept is to grow regression trees in the general form below:

$$\mathbb{E}[\tilde{X}|\boldsymbol{W},\boldsymbol{V}] = \sum_{m=1}^{M} c_m 1_{(\boldsymbol{W},\boldsymbol{V})\in R_m}$$

where $R_1,\ldots,R_M$ denotes a partition of feature space. Then we can repeat this procedure to build the RF by considering the approximate square root of the total number of predictors each time. The advantage of RF is we can see the contribution of each variable to the regression tree and their relative importance.

For each method, we did direct selection and post selection. For direct selection, we applied an estimated model from each approach to predict the long-term dietary intake straightly. For post selection, we performed linear regression afterward with selected variables ($\hat{S}$) from each approach. Specifically, for Lasso and SCAD, we have:

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = argmin\{\sum_{i=1}^{n}(\tilde{X}_i - (1,\boldsymbol{W}_i^{\top},\boldsymbol{V}_i^{\top})\boldsymbol{\beta}_1)_2^2\}.$$

$$\boldsymbol{\beta}_{\mathbf{1}} \in \mathbb{R}^P\ and\ \hat{\beta}_{1j} = 0\ \forall\, j \notin \hat{S}$$

For RF, the 10 most important variables are considered as the final selected variables. For both direct and post selection, we considered two ways to deal with $\boldsymbol{W}$ and $\boldsymbol{V}$; one is to consider both $\boldsymbol{W}$ and $\boldsymbol{V}$ in the approach of variable selection while the other is to consider only $\boldsymbol{W}$. To be more specific, in Lasso and SCAD, the penalization will be applied to $(\boldsymbol{W}, \boldsymbol{V})$ or to only $\boldsymbol{W}$, respectively. In RF, the decision trees will be built by considering $(\boldsymbol{W}, \boldsymbol{V})$ or only $\boldsymbol{W}$, respectively.

With estimated $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ we had in the prior step, we can then compute $\hat{X}_{1i} = (1,\boldsymbol{W}_i^{\top},\boldsymbol{V}_i^{\top})\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ to predict the long-term dietary intake (Z) among the n_2 calibration samples and run a regression of $\hat{X}_1$ on self-reported food frequency questionnaire data (Q) and $\boldsymbol{V}$ to build calibration equation using the n_2 calibration samples to estimate the parameter

$$\widehat{\gamma}_1 = \left\{\sum_{i=n_1+1}^{n_1+n_2}(1,Q_i,\boldsymbol{V}_i^{\top})^{\top}(1,Q_i,\boldsymbol{V}_i^{\top})\right\}^{-1}\left\{\sum_{i=n_1+1}^{n_1+n_2}(1,Q_i,\boldsymbol{V}_i^{\top})^{\top}\widehat{X}_{1i}\right\}.$$

Using the Stage 3 sample, we then estimate Z as $\widehat{Z}_{1i} = (1,Q_i,\boldsymbol{V}_i^{\top})\widehat{\gamma}_1$ for $i = n_1+n_2+1,\cdots,n_1+n_2+n_3$. Finally, we estimate the association between Z and the time-to-event endpoint (T^{*}, Δ) by solving the score equation for Cox model:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3}\int_0^{\tau}\left[\begin{pmatrix}\widehat{Z}_{1i}\\ \boldsymbol{V}_i\end{pmatrix} - \sum_j \frac{Y_j(t)\exp\left\{(\widehat{Z}_{1j},\boldsymbol{V}_j^{\top})\theta\right\}}{\sum_k Y_k(t)\exp\left\{(\widehat{Z}_{1k},\boldsymbol{V}_k^{\top})\theta\right\}}\begin{pmatrix}\widehat{Z}_{1j}\\ \boldsymbol{V}_j\end{pmatrix}\right]dN_i(t).$$

where τ is a pre-specified large number and we assume $P(C > \tau) > 0$, $N_i(t) = I(\Delta_i = 1, T_i^* \leq t)$ and $Y_i(t) = I(T_i^* \geq t)$. In application, τ is typically defined as the largest follow-up time in the Stage 3 sample.

3.2. Method 2: three-step with bias correction

As shown in (20), for low-dimension setting, Method 1 will lead to a bias factor in $\hat{Z}_1$ when using $\hat{X}_1$ and a bias-corrected estimator has been proposed. so for this high dimensional setting, we propose a similar bias-corrected estimator $\hat{X}_2 = \hat{X}_1 \widehat{BF}^{-1}$ where

$$\widehat{BF} = \hat{R}^2_{1|\boldsymbol{V}} = 1 - \frac{\widehat{Var}(\tilde{X}|\boldsymbol{V}, \boldsymbol{W}) - \tilde{\sigma}_x^2}{\widehat{Var}(\tilde{X}|\boldsymbol{V}) - \tilde{\sigma}_x^2}$$

is an estimated version of the bias factor.

For direct selection, we used K-fold cross-validated errors to compute the $\widehat{Var}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s)$ in penalized regression and RF to obtain $\widehat{BF}$. Denote the predicted value for the $k - th$ fold when using regression parameters from the other $K - 1$ training datasets as $\widehat{\tilde{X}}_{1k}$ when using $(\boldsymbol{W}, \boldsymbol{V})$ as predictors and as $\widehat{\tilde{X}}_{2k}$ when using $\boldsymbol{V}$ as predictors, then we have

$$\widehat{\tilde{X}}_1 = (\widehat{\tilde{X}}_{11}, \widehat{\tilde{X}}_{12}, \cdots, \widehat{\tilde{X}}_{1k}),$$

$$\widehat{\tilde{X}}_2 = (\widehat{\tilde{X}}_{21}, \widehat{\tilde{X}}_{22}, \cdots, \widehat{\tilde{X}}_{2k}).$$

With $\widehat{\tilde{X}}_1$ and $\widehat{\tilde{X}}_2$, $\widehat{BF}$ can be calculated as

$$\widehat{BF} = 1 - \frac{n^{-1}\sum_{i=1}^{n}(\tilde{X}_i - \widehat{\tilde{X}}_{1i})^2 - \tilde{\sigma}_x^2}{n^{-1}\sum_{i=1}^{n}(\tilde{X}_i - \widehat{\tilde{X}}_{2i})^2 - \tilde{\sigma}_x^2}.$$

For post selection, we first obtain the selected variables from the methods Lasso, SCAD, or RF. Afterward, we estimate the coefficients by refitting a linear regression. To facilitate interpretation, we consider both $\boldsymbol{W}$ and $\boldsymbol{V}$ in variable selection for the remainder of this subsection. Consequently, we have:

$$\tilde{X} \sim (\boldsymbol{W}_s, \boldsymbol{V}_s)\boldsymbol{\beta}^{\boldsymbol{PS}}_{\boldsymbol{WV}}.$$

Subsequently, we can fit a low-dimensional model as below:

$$\tilde{X} \sim \boldsymbol{V}\boldsymbol{\beta}_V.$$

Here, $\boldsymbol{W}_s$ and $\boldsymbol{V}_s$ denote the selected $\boldsymbol{W}$ and $\boldsymbol{V}$ variables, while $\boldsymbol{\beta}^{\boldsymbol{PS}}_{\boldsymbol{WV}}$ and $\boldsymbol{\beta}_V$ represent the corresponding coefficients in the aforementioned equations. From there, BF can be estimated as:

$$\widehat{BF} = 1 - \frac{\widehat{Var}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s) - \tilde{\sigma}_x^2}{\widehat{Var}(\tilde{X}|\boldsymbol{V}) - \tilde{\sigma}_x^2},$$

where

$$\widehat{Var}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s) = n^{-1}\sum_{i=1}^{n_1}(\tilde{X}_i - (\boldsymbol{W_{si}}, \boldsymbol{V_{si}})\hat{\boldsymbol{\beta}}^{\boldsymbol{PS}}_{\boldsymbol{WV}})^2, \text{ and}$$

$$\widehat{Var}(\tilde{X}|\boldsymbol{V}) = n^{-1}\sum_{i=1}^{n_1}(\tilde{X}_i - \boldsymbol{V_i}\hat{\boldsymbol{\beta}}_V)^2.$$

As demonstrated above, obtaining a precise estimation of $\widehat{Var}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s)$ is crucial for a reliable estimation of BF. Chatterjee and Jafarov (39) revealed that the estimator $\widehat{Var}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s)$, as mentioned earlier, leads to a downward bias when using Lasso. Therefore, we decide to compute and compare three different types of $\widehat{Var}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s)$ in our study involving post selection. We will provide a description of each type below.

(i) K-fold cross validation

We fit penalized regression or RF with the cross-validated training dataset and get predicted $\tilde{X}$ with selected $(\boldsymbol{W}, \boldsymbol{V})$ for each fold. Denote the selected subset as S_k for each training set $\tilde{X}_{-k}$.

Denote $\boldsymbol{W}_{S_k}$ as selected $\boldsymbol{W}$ and $\boldsymbol{V}_{S_k}$ as selected $\boldsymbol{V}$ in the (K-1) training dataset for each fold, then we can fit a linear regression of $\tilde{X}_k$ on $\boldsymbol{W}_{S_k}$, $\boldsymbol{V}_{S_k}$ and the predicted value is denoted as $\widehat{\tilde{X}}_{1k}$. Also, we can fit a linear regression of $\tilde{X}_k$ on $\boldsymbol{V}_k$ and the predicted value is denoted as $\widehat{\tilde{X}}_{2k}$. After doing this for all K folds, we get the estimated values of $\tilde{X}$ for the whole sample 1, that is,

$$\widehat{\tilde{X}}_1 = (\widehat{\tilde{X}}_{11}, \widehat{\tilde{X}}_{12}, \cdots, \widehat{\tilde{X}}_{1k}),$$

$$\widehat{\tilde{X}}_2 = (\widehat{\tilde{X}}_{21}, \widehat{\tilde{X}}_{22}, \cdots, \widehat{\tilde{X}}_{2k}).$$

With $\widehat{\tilde{X}}_1$ and $\widehat{\tilde{X}}_2$, $\widehat{BF}$ can be calculated as

$$\widehat{BF} = 1 - \frac{n^{-1}\sum_{i=1}^{n}(\tilde{X}_i - \widehat{\tilde{X}}_{1i})^2 - \tilde{\sigma}_x^2}{n^{-1}\sum_{i=1}^{n}(\tilde{X}_i - \widehat{\tilde{X}}_{2i})^2 - \tilde{\sigma}_x^2}.$$

(ii) Modified variance estimator

When performing penalized regression for variable selection, the choice of the regularization parameter λ is crucial for obtaining an accurate finite sample estimator. The value of λ influences both the number of variables selected and the extent to which their estimated coefficients are shrunk toward zero. If λ is set too large, not all signal variables will be selected, resulting in rapidly degrading performance (mainly characterized by a significant upward bias) as the true β becomes less sparse with a larger signal per element. Conversely, if λ is set too small, many noise variables will be selected, which allows spurious correlations to decrease our variance estimate, leading to considerable downward bias. Based on the simulation result in (40), there is a balance to be maintained when selecting the appropriate λ:

$$\widehat{Var^*}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s) = (n - \hat{s}_\lambda)^{-1}\sum_i(\tilde{X}_i - (\boldsymbol{W}_{si}, \boldsymbol{V}_{si})\hat{\boldsymbol{\beta}}^{\boldsymbol{PS}}_{\boldsymbol{WV}})^2$$

where $\hat{s}_\lambda$ is the number of nonzero elements in $\hat{b}$ at the regulation parameter λ selected with K-fold (usually 5–10) cross-validation. Then we have:

$$\widehat{BF} = 1 - \frac{\widehat{Var^*}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s) - \tilde{\sigma}_x^2}{\widehat{Var}(\tilde{X}|\boldsymbol{V}) - \tilde{\sigma}_x^2}.$$

(iii) Refitted cross-validation estimator (RCV)

This estimator is derived from the RCV procedure proposed by (37). We first split the dataset into two roughly equal parts: $[\tilde{X}^{(1)}, \boldsymbol{W}^{(1)}, \boldsymbol{V}^{(1)}]$ and $[\tilde{X}^{(2)}, \boldsymbol{W}^{(2)}, \boldsymbol{V}^{(2)}]$. We then perform penalized regression and RF on the first part. For penalized regression, we fit Lasso or SCAD on $\boldsymbol{W}$ and $\boldsymbol{V}$ with cross-validated $\hat{\lambda}$ to obtain the non-zero estimated coefficients for $\boldsymbol{W}$ and $\boldsymbol{V}$. In the case of RF, we select the 10 most important variables based on the residual sum of squares (RSS). We then refit the model with the selected $\boldsymbol{W}$ and $\boldsymbol{V}$ to obtain the post selected estimators of their coefficients, denoted as $\hat{\boldsymbol{\beta}}^{\boldsymbol{PS(1)}}\boldsymbol{WV}$. Subsequently, using the selected $\boldsymbol{W}$ and $\boldsymbol{V}$ in $\boldsymbol{W}^{(2)}$ and $\boldsymbol{V}^{(2)}$, we can compute the following variance estimate on the second part.

$$\widehat{Var_1^{**}}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s) = (n - \hat{s}^{(1)})^{-1} \sum_i (X_i^{*(2)} - (\boldsymbol{W}_{si}^{(2)}, \boldsymbol{V}_{si}^{(2)}) \hat{\boldsymbol{\beta}}_{\boldsymbol{WV}}^{\boldsymbol{PS(1)}})^2,$$

where $\hat{s}^{(1)}$ is the number of selected variables in the first part. Repeating the mirror image procedure on $[\tilde{X}^{(2)}, \boldsymbol{W}_s^{(2)}, \boldsymbol{V}_s^{(2)}]$, we can obtain $\hat{\lambda}_2$, selected W obtained from Lasso in the second part and $\widehat{Var_2}(\tilde{X}|\boldsymbol{W}, \boldsymbol{V})$. Last, $\widehat{BF}$ can be derived as below:

$$\widehat{Var^{**}}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s) = \frac{1}{2}(\widehat{Var_1^{**}}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s) + \widehat{Var_2^{**}}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s))$$

$$\widehat{BF} = 1 - \frac{\widehat{Var^{**}}(\tilde{X}|\boldsymbol{W}_s, \boldsymbol{V}_s) - \tilde{\sigma}_x^2}{\widehat{Var}(\tilde{X}|\boldsymbol{V}) - \tilde{\sigma}_x^2}.$$

With $\widehat{BF}$, we have $\hat{X}_{2i} = \hat{X}_{1i}/\widehat{BF}$. We can run a regression of $\hat{X}_2$ on self-reported food frequency questionnaire data (Q) and $\boldsymbol{V}$ to build calibration equation using the n_2 calibration samples to estimate the parameter

$$\widehat{\gamma}_2 = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^\top)^\top (1, Q_i, \boldsymbol{V}_i^\top) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^\top)^\top \widehat{X}_{2i} \right\}.$$

Using the Stage 3 sample, we then estimate Z as $\widehat{Z}_{2i} = (1, Q_i, \boldsymbol{V}_i^\top)\widehat{\gamma}_2$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$. Finally, we estimate the association of Z with the time-to-event endpoint (T^*, Δ) by solving the score equation for Cox model:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^\tau \left[\begin{pmatrix} \widehat{Z}_{2i} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{(\widehat{Z}_{2j}, \boldsymbol{V}_j^\top)\theta\right\}}{\sum_k Y_k(t) \exp\left\{(\widehat{Z}_{2k}, \boldsymbol{V}_k^\top)\theta\right\}} \begin{pmatrix} \widehat{Z}_{2j} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t).$$

For method 2 to work, we can relax Equation (3) to that the conditional mean $\mathbb{E}[Z|\boldsymbol{W}, \boldsymbol{V}]$ from sample 2 is the same as the conditional mean $\mathbb{E}[X|\boldsymbol{W}, \boldsymbol{V}]$ from sample 1.

3.3. Method 3: three-step with self-reported data

If the self-reported data Q from the feeding study is accessible and we presume that the distributions of $(Q|Z, \boldsymbol{V})$ remain consistent between the controlled feeding study and the cohort, the bias in the naive estimator can be rectified by simply incorporating Q into the biomarker development equation. This is because the inclusion of Q ensures that $\mathbb{E}[\hat{Z}|Q, \boldsymbol{V}] = \mathbb{E}[\mathbb{E}[Z|Q, \boldsymbol{V}, \boldsymbol{W}]|Q, \boldsymbol{V}] = \mathbb{E}[Z|Q, \boldsymbol{V}]$.

The sequence of the first method remains unchanged, but in the first step of the regression model, the log-transformed self-reported food frequency questionnaire data (Q) is included. Specifically, for the first step, the predictors $\boldsymbol{W}$, $\boldsymbol{V}$, and Q are utilized to construct the biomarker. Following this, in the second step, we employ $\boldsymbol{W}$, $\boldsymbol{V}$, and Q to estimate Z. Lasso, SCAD, and RF, as previously described, are all applied in Method 3 for variable selection and effect estimation in high-dimensional statistical inference, considering both direct selection and post-selection.

With the estimated $\hat{\boldsymbol{\beta}}_3$ from the first step, $\hat{X}_{3i} = (1, \boldsymbol{W}_i, Q_i, \boldsymbol{V}_i^\top)\hat{\boldsymbol{\beta}}_3$, we can execute a regression of $\hat{X}_{3i}$ on the self-reported food frequency questionnaire data (Q) and $\boldsymbol{V}$ to construct a calibration equation using the n_2 calibration samples to estimate the parameter

$$\widehat{\gamma}_3 = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^\top)^\top (1, Q_i, \boldsymbol{V}_i^\top) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^\top)^\top \widehat{X}_{3i} \right\}.$$

Using the Stage 3 sample, we then estimate Z as $\widehat{Z}_{3i} = (1, Q_i, \boldsymbol{V}_i^\top)\widehat{\gamma}_3$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$. Finally, we estimate the association of Z with the time-to-event endpoint (T^*, Δ) by solving the score equation for Cox model:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^\tau \left[\begin{pmatrix} \widehat{Z}_{3i} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{(\widehat{Z}_{3j}, \boldsymbol{V}_j^\top)\theta\right\}}{\sum_k Y_k(t) \exp\left\{(\widehat{Z}_{3k}, \boldsymbol{V}_k^\top)\theta\right\}} \begin{pmatrix} \widehat{Z}_{3j} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t).$$

For method 3 to work, we can relax Equation (3) to that the conditional mean $\mathbb{E}[Z|\boldsymbol{W}, Q, \boldsymbol{V}]$ from sample 2 is the same as the conditional mean $\mathbb{E}[X|\boldsymbol{W}, Q, \boldsymbol{V}]$ from sample 1.

3.4. Method 4: direct estimation

We build the estimating equation by regressing $\tilde{X}$ on Q and $\boldsymbol{V}$ in the first step and directly apply it to the third step. Then we build the calibration equation using the feeding study by regressing $\tilde{X}$ on $\boldsymbol{V}$ and Q and use the calibration equation to predict Z and perform a Cox regression of Y on Z and $\boldsymbol{V}$ in the full cohort to estimate the association parameter. In other words, we have

$$\widehat{\gamma}_4 = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^\top)^\top (1, Q_i, \boldsymbol{V}_i^\top) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^\top)^\top \tilde{X}_i \right\},$$

TABLE 1 Simulation results with direct Lasso selection forcing personal characteristics in the model.

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.69	0.844	0.867	0.94	0.53	0.469	0.461	0.85	0.28	0.453	0.462	0.94
		2	0.03	0.298	0.362	0.96	0.02	0.187	0.208	0.97	0.02	0.276	0.288	0.96
		3	−0.02	0.249	0.287	0.96	−0.02	0.155	0.182	0.97	−0.02	0.245	0.276	0.97
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	5	1	0.30	5.454	1.294	0.95	0.66	0.761	0.596	0.87	0.34	0.613	0.528	0.95
		2	−0.06	0.375	0.402	0.95	−0.01	0.209	0.215	0.96	0.00	0.287	0.297	0.95
		3	−0.03	0.259	0.300	0.97	−0.02	0.167	0.186	0.98	−0.03	0.264	0.283	0.96
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	10	1	78.75	776.634	107.201	0.94	0.89	1.328	0.819	0.88	0.38	0.662	0.625	0.94
		2	−3.72	36.126	2.578	0.95	−0.07	0.175	0.229	0.99	−0.04	0.236	0.307	0.95
		3	−0.03	0.259	0.312	0.95	−0.03	0.160	0.192	0.97	−0.03	0.257	0.292	0.94
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
Random	2	1	0.63	0.770	0.760	0.90	0.52	0.420	0.459	0.88	0.23	0.399	0.440	0.93
		2	0.04	0.369	0.324	0.96	0.04	0.212	0.217	0.97	0.02	0.268	0.289	0.94
		3	0.00	0.356	0.291	0.96	−0.01	0.178	0.182	0.97	0.04	0.746	0.278	0.96
		4	−0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	−0.01	0.258	0.324	0.97
	5	1	0.69	0.708	0.850	0.88	0.57	0.459	0.506	0.9	0.27	0.381	0.474	0.97
		2	0.03	0.258	0.350	1.00	0.04	0.184	0.225	0.97	0.03	0.242	0.304	0.99
		3	−0.01	0.227	0.303	0.97	0.00	0.149	0.194	0.96	−0.01	0.219	0.290	0.97
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.92	1.022	0.955	0.87	0.61	0.503	0.513	0.78	0.37	0.520	0.486	0.89
		2	0.07	0.322	0.349	0.93	0.03	0.206	0.214	0.94	0.07	0.304	0.294	0.91
		3	0.04	0.287	0.296	0.95	−0.01	0.172	0.183	0.94	0.03	0.278	0.281	0.95
		4	0.05	0.309	0.330	0.95	−0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95

$\hat{Z}_{4i} = (1, Q_i, \boldsymbol{V}_i)\hat{\boldsymbol{\gamma}}_{\boldsymbol{4}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{4}}$ by solving estimating equations

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^\tau \left[\begin{pmatrix} \widehat{Z}_{4i} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{(\widehat{Z}_{4j}, \boldsymbol{V}_j^\top)\theta\right\}}{\sum_k Y_k(t) \exp\left\{(\widehat{Z}_{4k}, \boldsymbol{V}_k^\top)\theta\right\}} \begin{pmatrix} \widehat{Z}_{4j} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t).$$

For method 4 to work, we can relax Equation (3) to that the conditional mean $\mathbb{E}[Z|Q, \boldsymbol{V}]$ from sample 3 is the same as the conditional mean $\mathbb{E}[X|Q, \boldsymbol{V}]$ from sample 1.

4. Simulation

We simulate data with varying levels of sparsity, effect size, and shape within the context of high-dimensional statistical inference. Our goal is to investigate how the sparsity, effect size, and shape among different measurements influence the bias and variance of various estimators. We compare the bias, empirical standard deviation (SD), estimated standard error (SE), and coverage rate for a nominal 95% confidence interval (CR) across different sample sizes, effect shapes, effect sizes, and correlation structures. Here the CR is computed from the asymptotic SE formula as shown in the Theorem 1 of (20) with the term $\hat{\Sigma}_{\gamma k}$ estimated from 100 Bootstrap samples using data from the first two samples given that there is no closed-form variance formula for $\Sigma_{\gamma k}$ when W is of high-dimension. We examine both scenarios, with and without penalties applied to the personal characteristics V during the first stage of penalized regression. Time-to-event outcomes are generated using the Cox model.

$$\begin{aligned}
&(Z, V) \sim \mathcal{N}\left(0, \begin{pmatrix} 1-\sigma_x^2 & \rho \\ \rho & 1 \end{pmatrix}\right), \\
&\boldsymbol{W} = b_0 + b_1 X + b_2 V + \epsilon_w, \\
&X = Z + \epsilon_x, \\
&\tilde{X} = X + \epsilon_{\tilde{X}}, \\
&Q = a_0 + a_1 Z + a_2 V + \epsilon_q, \\
&\lambda(t|Z, X, V, \boldsymbol{W}, Q) = \lambda(t|Z, V) = \lambda_0(t)\exp(\theta_z Z + \theta_v V),
\end{aligned}$$

where Z, V, X, and $Q \in \mathbb{R}$, while $\boldsymbol{W} \in \mathbb{R}^p$ is high-dimensional. In this study, ϵ_x and ϵ_q are independently sampled from normal distributions with mean zero and standard deviations σ_x and σ_q.

TABLE 2 Simulation results with post Lasso selection forcing personal characteristics in the model.

Pattern	Sparsity	Method	Setting 1				Setting 2				Setting 3			
			Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.46	0.813	0.658	0.93	0.36	0.421	0.375	0.91	0.16	0.376	0.387	0.94
		2.1	−0.20	0.226	0.217	0.79	−0.15	0.154	0.160	0.76	−0.103	0.208	0.216	0.90
		2.2	0.18	0.480	0.436	0.93	0.10	0.262	0.248	0.93	0.07	0.308	0.322	0.95
		2.3	0.02	0.374	0.329	0.94	0.00	0.219	0.203	0.94	−0.01	0.266	0.273	0.94
		3	−0.01	0.260	0.404	0.97	−0.02	0.167	0.190	0.97	0.00	0.274	0.336	0.97
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	5	1	0.39	0.858	1.286	0.95	0.41	0.592	0.434	0.92	0.18	0.448	0.417	0.95
		2.1	−0.24	0.181	0.296	0.80	−0.18	0.131	0.169	0.81	−0.115	0.206	0.217	0.89
		2.2	0.09	0.470	0.792	0.96	0.11	0.319	0.267	0.91	0.06	0.333	0.335	0.94
		2.3	−0.09	0.277	0.524	0.95	−0.04	0.222	0.205	0.97	−0.04	0.259	0.269	0.93
		3	−0.02	0.280	0.348	0.96	−0.02	0.187	0.190	0.98	−0.01	0.303	0.303	0.97
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	10	1	0.49	1.302	1.341	0.95	0.62	1.157	0.563	0.89	0.26	0.542	0.501	0.93
		2.1	−0.32	0.406	0.316	0.73	−0.27	0.199	0.191	0.72	−0.170	0.174	0.222	0.89
		2.2	0.12	0.609	0.771	0.95	0.16	0.528	0.317	0.93	0.08	0.355	0.374	0.93
		2.3	−0.14	0.285	0.407	0.94	−0.08	0.321	0.216	0.93	−0.06	0.269	0.275	0.92
		3	−0.03	0.269	0.334	0.95	−0.03	0.181	0.196	0.98	−0.03	0.256	0.304	0.96
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
Random	2	1	0.46	0.740	0.653	0.93	0.36	0.373	0.381	0.87	0.16	0.384	0.379	0.94
		2.1	−0.18	0.263	0.221	0.75	−0.15	0.155	0.162	0.79	−0.102	0.214	0.211	0.92
		2.2	0.19	0.554	0.438	0.94	0.11	0.255	0.257	0.94	0.08	0.336	0.321	0.95
		2.3	0.04	0.432	0.325	0.95	0.00	0.208	0.209	0.95	0.00	0.293	0.272	0.94
		3	−0.05	0.390	0.335	0.97	0.00	0.210	0.190	0.96	−0.04	0.311	0.301	0.96
		4	−0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	−0.01	0.258	0.324	0.97
	5	1	0.42	0.532	0.702	0.94	0.37	0.353	0.413	0.90	0.16	0.323	0.410	0.97
		2.1	−0.20	0.184	0.251	0.85	−0.15	0.142	0.173	0.81	−0.102	0.184	0.232	0.93
		2.2	0.14	0.335	0.479	0.98	0.11	0.223	0.276	0.97	0.07	0.266	0.345	0.98
		2.3	0.00	0.268	0.349	0.95	0.00	0.184	0.221	0.97	0.00	0.232	0.290	0.98
		3	−0.01	0.241	0.480	0.96	0.00	0.152	0.201	0.96	−0.01	0.232	0.317	0.95
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.57	0.792	0.675	0.89	0.37	0.383	0.388	0.86	0.23	0.414	0.401	0.91
		2.1	−0.20	0.269	0.244	0.77	−0.16	0.158	0.161	0.75	−0.077	0.221	0.220	0.88
		2.2	0.23	0.492	0.442	0.94	0.10	0.249	0.253	0.92	0.12	0.336	0.329	0.91
		2.3	0.01	0.297	0.319	0.96	−0.02	0.188	0.201	0.94	0.03	0.277	0.277	0.94
		3	0.04	0.285	0.325	0.95	0.00	0.181	0.190	0.96	0.04	0.282	0.300	0.94
		4	0.05	0.309	0.330	0.95	−0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95

The censoring time is sampled from a mixture of a uniform distribution Unif(0, 10) and a point mass at 10, with equal probability. Three settings are considered: (i) baseline setting, (ii) weak biomarker effect and strong self-reported data effect, and (iii) strong biomarker effect. We experiment with three sparsity levels of $\boldsymbol{W}$ (2, 5, and 10), and consider two different patterns of the effect size for $\boldsymbol{W}$: equivalent and random. More details on the parameter settings can be found in Supplementary material (Section 1.1).

The bias, mean estimated standard error (SE), empirical standard deviation (SD), and coverage rate (CR) of 95% nominal confidence interval for all four methods from 100 simulations are listed in Tables 1, 2 with Lasso penalized

TABLE 3 Simulation results for direct SCAD selection forcing personal characteristics in the model.

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.53	1.219	0.715	0.94	0.39	0.463	0.384	0.89	0.17	0.359	0.398	0.94
		2	−0.02	0.356	0.277	0.95	−0.04	0.179	0.180	0.95	−0.04	0.230	0.252	0.94
		3	−0.02	0.248	0.306	0.97	−0.01	0.176	0.185	0.96	−0.02	0.266	0.290	0.96
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	5	1	1.00	2.047	2.359	0.91	1.48	8.942	0.734	0.91	0.37	1.263	0.538	0.95
		2	−0.12	0.255	0.387	0.93	−0.07	0.257	0.201	0.93	−0.06	0.326	0.262	0.91
		3	−0.02	0.259	0.317	0.97	−0.03	0.156	0.193	0.98	−0.01	0.331	0.319	0.96
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	10	1	3.15	5.880	8.839	0.90	1.56	2.116	2.040	0.82	0.55	0.927	0.971	0.92
		2	−8.89	62.117	1.686	0.88	−0.12	0.239	0.357	0.88	−0.12	0.230	0.297	0.93
		3	−0.02	0.272	0.338	0.95	−0.02	0.165	0.197	0.95	−0.02	0.272	0.312	0.95
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
Random	2	1	0.42	0.572	0.623	0.93	0.39	0.371	0.380	0.86	0.16	0.357	0.383	0.93
		2	−0.03	0.281	0.281	0.95	−0.02	0.179	0.190	0.96	−0.03	0.242	0.257	0.94
		3	−0.25	2.306	0.299	0.97	0.00	0.190	0.185	0.95	−0.18	1.627	0.284	0.96
		4	−0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	−0.01	0.258	0.324	0.97
	5	1	0.52	0.618	0.742	0.89	0.43	0.356	0.424	0.91	0.20	0.339	0.421	0.97
		2	−0.03	0.231	0.300	0.97	−0.02	0.162	0.196	0.98	−0.02	0.221	0.265	0.98
		3	−0.01	0.226	0.311	0.97	0.00	0.149	0.197	0.96	−0.01	0.221	0.299	0.96
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.81	1.084	0.923	0.87	0.51	0.499	0.471	0.81	0.29	0.459	0.446	0.89
		2	−0.01	0.291	0.310	0.94	−0.06	0.162	0.185	0.95	0.01	0.265	0.259	0.92
		3	0.04	0.328	0.309	0.95	−0.01	0.171	0.185	0.93	0.03	0.292	0.294	0.96
		4	0.05	0.309	0.330	0.95	−0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95

regression. The performance of our proposed Method 2 varies under different scenarios. Post selection and direct selection are both performed for variable selection. The results are for the ones forcing personal characteristics (V) in the model while the results for not forcing V in the model can be found in Supplementary Tables 2, 3.

In general, the post selection methods perform slightly better than the direct-selection methods with lower SDs and SEs. For a few settings, the direct selection approach does not perform stably in terms of bias and SD. The direct selection not forcing the inclusion of personal characteristics showed more stable results compared with direct selection forcing the inclusion of personal characteristics for Method 2 but the variance is larger in general for all other methods. For the post selection approach, the performances of the three variance estimation methods, are shown as 2.1 (K-fold cross-validation), 2.2 (modified variance estimator), and 2.3 (RCV) in Tables 2, 4, 6. Method 2.3 (RCV estimation under post selection within Method 2) performs the best among all three approaches across different settings and patterns. Some key advantages of Method 2.3 include lower bias and smaller standard deviations (SD) and standard errors (SE), along with good coverage rates (CR).

Tables 1, 2 shows the results using Lasso penalized regression when forcing personal characteristics in the model. As the sparsity level increases, the performance of most methods seems to degrade, with higher biases and lower coverage rates. Methods 3 and 4 demonstrate good performances in most of the settings. However, when the strength of the biomarker is strong and the strength of FFQ is relatively weak (i,e., Setting 3), we can see Method 2 generally generated the most efficient result compared with Methods 3 and 4. When we have strong biomarker effects (Setting 3), Method 2.3 outperforms the other methods.

Tables 3, 4 present the results for SCAD penalized regression when personal characteristics are forced into the model. Corresponding results without forcing personal characteristics can be found in Supplementary Tables 4, 5. When comparing SCAD with Lasso, we can observe a similar trend in terms of bias control and standard deviation (SD) across various effect size patterns and settings. In addition, when comparing the three approaches for variance estimation in constructing the bias factor (BF) using Method 2 with post selection, the RCV approach continues to

TABLE 4 Simulation results for post SCAD selection forcing personal characteristics in the model.

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.49	0.940	1.289	0.92	0.46	0.591	0.514	0.86	0.18	0.420	0.443	0.95
		2.1	−0.23	0.180	0.227	0.78	−0.18	0.154	0.166	0.74	−0.12	0.201	0.208	0.87
		2.2	0.15	0.393	0.470	0.94	0.11	0.293	0.251	0.93	0.07	0.308	0.321	0.95
		2.3	0.08	0.500	2.590	0.93	0.01	0.268	0.670	0.93	0.02	0.329	1.304	0.94
		3	−0.01	0.259	0.331	0.97	−0.01	0.171	0.194	0.96	0.00	0.262	0.315	0.97
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	5	1	0.46	2.725	11.428	0.93	0.42	2.520	3.700	0.93	0.25	0.732	0.667	0.91
		2.1	−0.29	0.312	0.309	0.71	−0.18	0.175	0.201	0.79	−0.13	0.237	0.224	0.86
		2.2	0.46	3.503	0.759	0.95	0.17	0.884	0.406	0.93	0.10	0.464	0.361	0.96
		2.3	0.41	4.576	2.323	0.91	−0.21	1.818	1.132	0.87	−0.01	0.315	0.650	0.94
		3	−0.01	0.322	0.352	0.96	−0.01	0.229	0.195	0.98	0.01	0.450	0.462	0.97
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	10	1	1.12	2.994	8.838	0.94	1.08	3.364	3.596	0.88	0.44	0.690	1.498	0.90
		2.1	−0.43	0.773	0.858	0.83	−0.27	0.319	0.307	0.78	−0.19	0.208	0.268	0.86
		2.2	0.16	0.628	1.148	0.96	0.18	0.534	0.357	0.95	0.09	0.334	0.436	0.95
		2.3	0.05	0.629	2.045	0.95	−0.01	0.893	2.983	0.93	0.10	1.633	1.124	0.95
		3	−0.02	0.276	0.412	0.94	−0.02	0.174	0.200	0.95	−0.02	0.269	0.357	0.95
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
Random	2	1	0.45	0.641	2.098	0.97	0.44	0.717	0.674	0.87	0.20	0.614	0.432	0.96
		2.1	−0.25	0.191	0.254	0.72	−0.18	0.147	0.161	0.77	−0.13	0.198	0.227	0.86
		2.2	0.06	0.742	0.431	0.96	0.11	0.263	0.258	0.91	0.07	0.341	0.323	0.91
		2.3	0.06	0.429	0.839	0.94	0.03	0.243	0.321	0.95	0.01	0.286	0.362	0.94
		3	−0.03	0.288	0.348	0.96	0.02	0.294	0.191	0.96	−0.03	0.313	0.309	0.95
		4	−0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	−0.01	0.258	0.324	0.97
	5	1	0.64	1.151	3.383	0.92	0.46	0.497	0.657	0.90	0.21	0.381	0.476	0.97
		2.1	−0.24	0.183	0.433	0.77	−0.18	0.150	0.181	0.75	−0.12	0.181	0.222	0.91
		2.2	0.15	0.342	0.504	0.97	0.11	0.214	0.278	0.97	0.07	0.268	0.351	0.99
		2.3	0.06	0.409	0.811	0.97	0.02	0.231	0.327	0.98	0.02	0.229	0.364	0.99
		3	−0.01	0.230	0.476	0.97	0.00	0.162	0.205	0.96	−0.01	0.231	0.324	0.95
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.96	1.940	4.993	0.92	0.84	2.196	1.202	0.86	0.35	0.762	0.525	0.95
		2.1	−0.24	0.247	0.300	0.73	−0.19	0.171	0.175	0.78	−0.10	0.216	0.217	0.84
		2.2	0.25	0.480	0.465	0.88	0.11	0.256	0.262	0.94	0.12	0.339	0.337	0.93
		2.3	0.24	1.034	1.989	0.93	0.04	0.348	0.374	0.92	0.08	0.442	0.338	0.92
		3	0.05	0.313	0.346	0.94	0.00	0.192	0.192	0.94	0.04	0.289	0.318	0.95
		4	0.05	0.309	0.330	0.95	−0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95

outperform the others in controlling bias and providing the most efficient results. With direct selection using SCAD in Method 2, the bias is generally well-controlled, and the confidence rate (CR) is promising when personal characteristics are not fixed for variable filtering. These results are comparable to those obtained with Lasso. However, when variables are post selected, SCAD's performance is not as strong as Lasso's. This is particularly noticeable in scenarios with large sparsity, where SCAD struggles to control bias effectively. In summary, Lasso demonstrates superior performance in variance estimation and bias control when compared to SCAD.

TABLE 5 Simulation results for direct RF selection not forcing personal characteristics in the model.

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	1.08	20.056	42.699	0.91	2.29	1.361	2.066	0.90	6.28	17.557	22.674	0.91
		2	−0.07	7.706	4.795	0.93	0.35	0.528	0.645	0.93	3.56	17.387	15.925	0.91
		3	1.62	1.506	2.256	0.91	1.03	0.724	0.785	0.84	1.94	2.153	3.978	0.9
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	5	1	2.77	3.167	5.799	0.90	1.90	1.236	1.366	0.76	1.67	1.840	2.236	0.88
		2	0.17	0.753	1.028	0.93	0.00	0.242	0.324	0.98	0.62	1.384	0.919	0.87
		3	1.16	1.282	1.335	0.89	0.69	0.520	0.587	0.84	0.87	0.876	1.067	0.93
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	10	1	2.34	2.588	4.669	0.88	2.40	1.475	2.011	0.86	1.18	1.374	1.195	0.88
		2	−0.18	3.913	0.618	0.95	−0.13	0.254	0.393	0.93	0.07	0.370	0.463	0.95
		3	0.99	1.141	1.293	0.87	0.60	0.486	0.641	0.90	0.67	0.765	0.835	0.89
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
Random	2	1	−2.68	31.746	61.244	0.92	1.29	25.138	109.942	0.86	−80.83	641.046	101.489	0.94
		2	−1.85	9.158	14.467	0.95	−0.89	5.915	19.942	0.88	19.35	184.338	46.897	0.9
		3	3.53	6.489	25.208	0.98	2.56	23.515	24.412	0.92	−0.63	42.632	40.913	0.94
		4	−0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	−0.01	0.258	0.324	0.97
	5	1	32.47	339.650	88.785	0.91	−0.50	66.632	120.634	0.88	−16.21	142.496	168.673	0.92
		2	8.23	83.762	26.403	0.92	−0.06	8.969	53.692	0.91	−0.43	22.138	67.965	0.94
		3	−1.26	20.206	24.392	0.91	1.39	23.664	33.013	0.82	1.70	39.330	52.827	0.87
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	4.92	81.304	82.072	0.89	6.89	58.844	71.207	0.89	−5.83	125.518	375.281	0.88
		2	0.01	7.096	21.133	0.98	48.46	496.413	595.578	0.90	2.63	31.842	43.618	0.91
		3	−14.44	127.206	13.957	0.89	2.56	4.880	172.054	0.84	4.08	17.941	31.275	0.89
		4	0.05	0.309	0.330	0.95	−0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95

RF offers an alternative approach for constructing a biomarker prediction model in the second stage. Tables 5, 6 display the results obtained using RF. When the 10 most important variables are directly selected with RF, the estimated bias is considerably large in most scenarios. However, when post selection is applied to variables using RF, results with variance estimation approaches 2.2 and 2.3 both exhibit small bias and promising confidence rates (CR). These outcomes are comparable to those achieved with Lasso for Method 2 using the RCV estimation (2.3). In summary, while RF does not provide accurate estimations of associated parameters when using direct selection, its performance is similar to Lasso when post selection is employed.

Overall, with the linear settings, Lasso provides a consistent estimator in most cases and largely attenuates the bias compared with SCAD and RF. For more general model settings, RF has potential advantages when the linear model does not hold. The post selection option with RCV variance estimation for BF construction provides consistent estimation on associated parameters with stable CR and is recommended especially when we have sparse high-dimensional data structure.

5. Data analysis

We exemplify our methodologies utilizing data from the WHI NPAAS feeding study ($n = 153$), NPAAS biomarker study ($n = 450$), and the comprehensive WHI cohort data [comprising the WHI Observational Study (OS) and the Dietary Modification Trial Control Arm (DM-C), $n = 122,970$]. A log-transformed self-reported ratio of sodium to potassium intake from FFQ serves as Q. Covariates such as age, BMI, race/ethnicity, education level, self-reported physical activity, and smoking status are considered as $\boldsymbol{V}$. The high-dimensional 24 h urine measurements, acquired via nuclear magnetic resonance (NMR) and gas chromatography-mass spectrometry (GC-MS) platforms, are denoted as $\boldsymbol{W}$. The disease outcome under consideration is total cardiovascular disease (CVD). The prevalence of CVD events is <10% (41), suggesting that the rare disease assumption is not substantially violated. Follow-up times commence at the moment of FFQ measurement (year-1 visit in DM-C and at enrollment in OS) and persist until the earliest of the specific CVD outcomes under consideration, death, loss to follow-up, or September 30, 2010, whichever occurs first.

TABLE 6 Simulation results for post RF selection not forcing personal characteristics in the model.

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.47	0.606	0.655	0.88	0.41	0.383	0.394	0.83	0.17	0.373	0.391	0.93
		2.1	−0.09	0.217	0.248	0.90	−0.09	0.153	0.167	0.91	−0.07	0.218	0.238	0.94
		2.2	0.00	0.275	0.306	0.97	0.04	0.199	0.217	0.96	−0.02	0.246	0.264	0.95
		2.3	0.00	0.279	0.324	0.97	−0.01	0.189	0.205	0.96	0.00	0.259	0.275	0.95
		3	−0.02	0.260	0.308	0.96	−0.01	0.170	0.193	0.97	−0.01	0.262	0.291	0.95
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	5	1	0.72	1.298	0.974	0.90	0.61	0.569	0.564	0.85	0.25	0.483	0.463	0.92
		2.1	−0.10	0.266	0.297	0.91	−0.14	0.161	0.187	0.84	−0.06	0.236	0.253	0.94
		2.2	0.02	0.354	0.384	0.97	0.06	0.229	0.263	0.94	−0.01	0.262	0.282	0.94
		2.3	0.17	1.820	0.366	0.96	−0.07	0.180	0.209	0.96	0.01	0.285	0.299	0.96
		3	0.00	0.293	0.345	0.98	0.00	0.183	0.201	0.97	0.03	0.365	0.337	0.93
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
	10	1	1.02	1.184	2.059	0.91	1.20	1.336	1.205	0.86	0.59	0.736	0.818	0.90
		2.1	−0.17	0.229	0.412	0.92	−0.24	0.299	0.244	0.73	−0.05	0.269	0.317	0.96
		2.2	0.04	0.338	0.646	0.97	0.17	0.416	0.444	0.92	0.04	0.309	0.379	0.95
		2.3	−0.01	0.335	0.517	0.95	−0.07	0.268	0.275	0.90	0.09	0.372	0.431	0.91
		3	0.06	0.524	0.395	0.94	0.01	0.188	0.213	0.97	0.25	1.677	13.383	0.93
		4	−0.01	0.280	0.355	0.94	−0.02	0.158	0.194	0.95	−0.01	0.280	0.355	0.94
Random	2	1	0.43	0.570	0.651	0.91	0.39	0.367	0.395	0.88	0.15	0.341	0.381	0.93
		2.1	−0.09	0.230	0.248	0.95	−0.07	0.164	0.177	0.92	−0.06	0.222	0.238	0.97
		2.2	0.03	0.313	0.328	0.96	0.04	0.203	0.224	0.97	0.00	0.254	0.275	0.95
		2.3	−0.02	0.282	0.308	0.97	0.01	0.199	0.210	0.96	−0.01	0.251	0.269	0.97
		3	0.05	0.740	0.316	0.97	0.01	0.188	0.194	0.96	−0.25	2.308	0.293	0.96
		4	−0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	−0.01	0.258	0.324	0.97
	5	1	0.68	1.701	0.743	0.93	0.41	0.350	0.421	0.87	0.18	0.325	0.408	0.98
		2.1	−0.08	0.380	0.257	0.91	−0.07	0.146	0.177	0.96	−0.05	0.202	0.249	0.98
		2.2	0.08	0.459	0.351	0.97	0.05	0.189	0.228	0.96	0.01	0.232	0.290	1.00
		2.3	−0.04	0.238	0.461	0.95	0.00	0.176	0.210	0.99	−0.01	0.223	0.280	1.00
		3	0.01	0.236	0.334	0.97	0.02	0.159	0.204	0.97	0.00	0.225	0.305	0.96
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.70	2.558	1.365	0.87	0.45	0.443	0.425	0.84	0.26	0.431	0.421	0.92
		2.1	−0.05	0.377	0.310	0.87	−0.11	0.167	0.174	0.85	−0.02	0.247	0.240	0.91
		2.2	0.13	0.507	0.444	0.94	0.04	0.213	0.221	0.94	0.05	0.291	0.284	0.91
		2.3	0.06	0.797	0.579	0.87	−0.03	0.206	0.195	0.93	0.02	0.271	0.263	0.91
		3	0.06	0.298	0.319	0.96	0.00	0.172	0.195	0.94	0.05	0.283	0.296	0.95
		4	0.05	0.309	0.330	0.95	−0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95

In our analytical process, hazard rates are modeled as implicitly conditioned on the continued survival of the study subject. This implies that death is not viewed as a source of censoring in our formulation. Rather, death merely constrains the follow-up period during which hazard rate information is collected for the subject. This differs from considering death as censoring non-fatal outcomes, which would be the case in a competing risk formulation.

We scrutinized the normality of the log-transformed self-reported intake (Q), the log-transformed metabolites from 24-hour urine measurements ($\boldsymbol{W}$), and the log-transformed evaluated sodium/potassium ratio ($\tilde{X}$) utilizing the NPAAS-FS

TABLE 7 Association between 20% increase in sodium-to-potassium ratio with total CVD in high-dimensional space under Lasso, SCAD, and RF approaches.

Approach	Method	Lasso		SCAD		Random forest	
		HR	95 CI%	HR	95 CI%	HR	95 CI%
Direct selection with V not fixed	1	1.30	(1.16, 1.46)	1.18	(1.04, 1.34)	1.39	(0.61, 3.18)
	2	1.09	(1.04, 1.15)	1.05	(0.99, 1.11)	1.10	(0.71, 1.70)
	3	1.08	(1.00, 1.17)	1.08	(0.97, 1.21)	1.28	(1.00, 1.65)
	4	1.08	(1.03, 1.13)	1.08	(1.03, 1.13)	1.08	(1.03, 1.13)
Direct-selection with V fixed	1	1.21	(1.04, 1.40)	1.18	(0.80, 1.73)	–	–
	2	1.06	(1.00, 1.14)	1.05	(1.00, 1.11)	–	–
	3	1.08	(1.00, 1.18)	1.08	(0.88, 1.33)	–	–
	4	1.08	(1.03, 1.13)	1.08	(1.03, 1.13)	–	–
Post selection with V not fixed	1	1.20	(1.08, 1.33)	1.17	(0.81, 1.69)	1.19	(1.06, 1.34)
	2.3	1.08	(1.01, 1.15)	1.07	(0.87, 1.31)	1.08	(1.03, 1.13)
	3	1.09	(0.98, 1.21)	1.10	(0.97, 1.25)	1.11	(1.05, 1.16)
	4	1.08	(1.03, 1.13)	1.08	(1.03, 1.13)	1.08	(1.03, 1.13)
Post selection with V fixed	1	1.16	(1.03, 1.31)	1.15	(0.87,1.52)	–	–
	2.3	1.06	(0.99, 1.14)	1.05	(0.90, 1.22)	–	–
	3	1.08	(1.00, 1.17)	1.08	(0.97, 1.20)	–	–
	4	1.08	(1.03, 1.13)	1.08	(1.02, 1.14)	–	–

dataset. No evidence indicated a violation of any normality assumptions (B-H adjusted p value > 0.1) (42).

The estimated HR and corresponding 95% confidence interval according to a 20% increase in the sodium-potassium ratio are shown in Table 7 for the methods of Lasso, SCAD, and RF.

We observe that the estimated HR is >1 in all cases, indicating a higher risk of CVD with increased sodium-to-potassium ratios, regardless of the different high-dimensional approaches used. These findings are consistent with those reported in previously published studies (20). The most conservative estimate for $\hat{\tilde{\sigma}}_x^2$, 0, is used to construct the BF in Method 2. Moreover, RCV variance estimation is employed to construct the BF for the post selection approach with Method 2. The estimation of the associated parameter derived from Method 2 is smaller in scale compared to Method 1 and is similar to Methods 3 and 4. We note that the 95% CI does not include an HR of 1 with Lasso and RF in most cases, indicating a significant association between calibrated dietary intake and the risk of CVD. Conversely, the 95% CI with SCAD exhibits less efficient results with larger variance, indicating a non-significant association between calibrated dietary intake and the risk of CVD in several instances.

6. Discussion

We investigated the prerequisites for a valid biomarker in high-dimensional space for regression calibration purposes. Various methods to handle high-dimensional data (i.e., Lasso, SCAD, and RF) and approaches to variable selection (i.e., direct and post selection) were applied and compared across different scenarios, such as sparsity level and pattern of effect size. This paper offers researchers a comprehensive understanding of how to handle high-dimensional data in calibrated regression studies. Building linear regression models in high-dimensional space presents challenges, such as overfitting to samples and multicollinearity, which can lead to inadequate estimations.

In order to identify the most effective measurements associated with consumed dietary intakes in the feeding study, Lasso, SCAD, and RF were applied for variable selection within the high-dimensional dataset. Overall, Lasso demonstrated more stable results for variable selection compared to the other two approaches. Method 2, with the BF constructed using RCV estimation under the Lasso post selection approach, consistently provided good estimations in most cases.

It is worth noting that various factors, such as filtering conditions and methods for obtaining tuning parameters, can influence the accuracy of the biomarker prediction model when using penalized regression methods and RF. Depending on these choices, the accuracy of the estimated association parameters can vary significantly. Consequently, researchers should carefully consider these factors to achieve the most accurate and reliable results when working with high-dimensional data in calibrated regression studies.

Identifying effective measurements associated with consumed dietary intakes is crucial for biomarker construction. Statistical inference presents challenges with penalized estimators. In this paper, a bootstrapping approach was employed for variance estimation in high-dimensional data for penalized regression and RF. However, there are alternative approaches for variance estimation in high-dimensional data with penalized regression that could be considered in future analyses.

One issue with the estimated covariance matrix relates to zero components. Specifically, when coefficients are zero, the approximate covariance matrix results in zero for estimated variance. Although the estimation of non-zero components is robust, the signs of zero components can be either negative or positive. This issue is also present in the sandwich formula of the covariance matrix developed by (31). Wasserman and Roeder (43) proposed a two-stage procedure for valid inference. Their method involves randomly dividing the data into training and testing datasets. Penalized linear regression is used in the training data to select informative variables in the first stage, while ordinary least squares (OLS) are applied in the testing data to compute standard errors. A drawback of the single-split method is that results may depend on how the data is split. To address this, Meinshausen et al. (44) suggested a multi-split method, which repeats the single-split multiple times. Lockhart et al. (45) introduced the covariance test statistic to test the significance of predictor variables that enter the current Lasso model. For ultra-high-dimensional cases where the sample size is equal to or smaller than the variable dimension, the sure independent screening (SIS) technique proposed by (31) can be considered for variable screening in future work.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the data can only be accessed through the collaborative mode as described on the Women's Health Initiative website. Requests to access these datasets should be directed to www.whi.org.

Author contributions

RD and CZ designed the research and had primary responsibility for the final content. YZ, RD, and CZ derived the methods. YZ performed simulation studies and run the data analysis. RD drafted the paper. YH and RP revised the draft and provide critical feedback to the paper. All authors read and approved the final manuscript.

Funding

This work was supported in part by grants R01 CA119171 and R01 CA277133 from the U.S. National Cancer Institute and U54 GM115458 from the National Institute of General Medical Sciences. The WHI programs are funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts, HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C.

Acknowledgments

The authors acknowledge the following investigators in the Women's Health Initiative (WHI) Program: Program Office: Jacques E. Rossouw, Shari Ludlam, Dale Burwen, Joan McGowan, Leslie Ford, and Nancy Geller, National Heart, Lung, and Blood Institute, Bethesda, Maryland; Clinical Coordinating Center, Women's Health Initiative Clinical Coordinating Center: Garnet L. Anderson, Ross L. Prentice, Andrea Z. LaCroix, and Charles L. Kooperberg, Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington; Investigators and Academic Centers: JoAnn E. Manson, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; Barbara V. Howard, MedStar Health Research Institute/Howard University, Washington, DC; Marcia L. Stefanick, Stanford Prevention Research Center, Stanford, California; Rebecca Jackson, The Ohio State University, Columbus, Ohio; Cynthia A. Thomson, University of Arizona, Tucson/Phoenix, Arizona; Jean Wactawski-Wende, University at Buffalo, Buffalo, New York; Marian C. Limacher, University of Florida, Gainesville/Jacksonville, Florida; Robert M. Wallace, University of Iowa, Iowa City/Davenport, Iowa; Lewis H. Kuller, University of Pittsburgh, Pittsburgh, Pennsylvania; and Sally A. Shumaker, Wake Forest University School of Medicine, Winston-Salem, North Carolina; Women's Health Initiative Memory Study: Sally A. Shumaker, Wake Forest University School of Medicine, Winston-Salem, North Carolina. For a list of all the investigators who have contributed to WHI science, please visit: https://www.whi.org/researchers/SitePages/WHI%20Investigators.aspx. Decisions concerning study design, data collection and analysis, interpretation of the results, the preparation of the manuscript, and the decision to submit the manuscript for publication resided with committees that comprised WHI investigators and included National Heart, Lung, and Blood Institute representatives.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The contents of the paper are solely the responsibility of the authors.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnut.2023.1215768/full#supplementary-material

References

1. Adams KF, Schatzkin A, Harris TB, Kipnis V, Morris T, Ballard-Barbash R. Overweight, obesity and mortality in a large prospective cohort of persons 50 to 71 years old. *N Engl J Med.* (2006) 355:763–78. doi: 10.1056/NEJMoa055643

2. World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR). *Food, Nutrition and the Prevention of Cancer: A Global Perspective.* Washington, DC: American Institute for Cancer Research (2007).

3. Paeratakul S, Popkin BM, Kohlmeier L, Hertz-Picciotto I, Guo X, Edwards LJ. Measurement error in dietary data: implications for the epidemiologic study of the diet-disease relationship. *Eur J Clin Nutr.* (1998) 52:722–7. doi: 10.1038/sj.ejcn.1600633

4. Prentice RL, Mossavar-Rahmani Y, Huang Y, Horn LV, Beresford SAA, Caan B, et al. Evaluation and comparison of food records, recalls, and frequencies for energy and protein assessment by using recovery biomarkers. *Am J Epidemiol.* (2011) 174:591–603. doi: 10.1093/aje/kwr140

5. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective.* Boca Raton, FL: CRC Press (2006). doi: 10.1201/9781420010138

6. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. *J Natl Cancer Instit.* (2011) 103:1086–92. doi: 10.1093/jnci/djr189

7. Huang Y, Wang CY. Cox regression with accurate covariate unascertainable: a nonparametric-correction approach. *J Am Stat Assoc.* (2000) 45:1209–19. doi: 10.1080/01621459.2000.10474321

8. Kipnis V, Midthune D, Freedman L, Bingham S, Schatzkin A, Subar A, et al. Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutr.* (2002) 5:915–23. doi: 10.1079/PHN2002383

9. Song X, Huang X. On corrected score approach for proportional hazards model with covariatemeasurement error. *Biometrics.* (2005) 61:702–14. doi: 10.1111/j.1541-0420.2005.00349.x

10. Carroll RJ, Midthune D, Subar AF, Shumakovich M, Freedman LS, Thompson FE, et al. Taking advantage of the strengths of 2 different dietary assessment instruments to improve intake estimates for nutritional epidemiology. *Am J Epidemiol.* (2012) 175:340–7. doi: 10.1093/aje/kwr317

11. Yan Y, Yi GY. A corrected profile likelihood method for survival data with covariate measurement error under the Cox model. *Can J Stat.* (2015) 43:454–80. doi: 10.1002/cjs.11258

12. Li Y, Ryan L. Inference on survival data with covariate measurement error-An imputationapproach. *Scand J Stat.* (2006) 33:169–90. doi: 10.1111/j.1467-9469.2006.00460.x

13. Prentice RL, Huang Y, Tinker LF, Beresford SA, Neuhouser ML, Pettinger M, et al. Regression calibration in nutritional epidemiology: example of fat density and total energy in relationship to postmenopausal breast cancer. *Am J Epidemiol.* (2013) 178:1663–72. doi: 10.1093/aje/kwt198

14. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Stat Med.* (2014) 33:2137–55. doi: 10.1002/sim.6095

15. Bartlett JW, Keogh RH. Bayesian correction for covariate measurement error: a frequentist evaluation and comparison with regression calibration. *Stat Methods Med Res.* (2018) 27:1695–708. doi: 10.1177/0962280216667764

16. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol.* (1990) 132:734–45. doi: 10.1093/oxfordjournals.aje.a115715

17. Shaw PA, Prentice RL. Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics.* (2012) 68:397–407. doi: 10.1111/j.1541-0420.2011.01690.x

18. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika.* (1982) 69:331–42. doi: 10.1093/biomet/69.2.331

19. Zheng C, Beresford SAA, Horn LV, Tinker LF, Thomson CA, Neuhouser ML, et al. Simultaneous association of total energy consumption and activity-related energy expenditure with cardiovascular disease, cancer, and diabetes risk among postmenopausal women. *Am J Epidemiol.* (2014) 180:526–35. doi: 10.1093/aje/kwu152

20. Zheng C, Zhang Y, Huang Y, Prentice R. Using controlled feeding study for biomarker development in regression calibration for disease association estimation. *Stat Biosci.* (2023) 15:57–113. doi: 10.1007/s12561-022-09349-3

21. Lampe JW, Huang Y, Neuhouser ML, Tinker LF, Song X, Schoeller DA, et al. Dietary biomarker evaluation in a controlled feeding study in women from the Women's Health Initiative cohort. *Am J Clin Nutr.* (2017) 105:466–75. doi: 10.3945/ajcn.116.144840

22. Prentice R, Pettinger M, Neuhouser M, Raftery D, Zheng C, Gowda N, et al. Biomarker-calibrated macronutrient intake and chronic disease risk among postmenopausal women. *J Nutr.* (2021) 151:2330–41. doi: 10.1093/jn/nxab091

23. Zhang Y, Dai R, Huang Y, Prentice R, Zheng C. Using simultaneous regression calibration to study the effect of multiple error-prone exposures on disease risk utilizing biomarkers developed from a controlled feeding study. *Ann Appl Stat.* (2023). [Epub ahead of print].

24. Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics.* (1993) 35:109–35. doi: 10.1080/00401706.1993.10485033

25. Breiman L. Better subset regression using the nonnegative garrote. *Technometrics.* (1995) 37:373–84. doi: 10.1080/00401706.1995.10484371

26. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B.* (1996) 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x

27. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc.* (2001) 96:1348–60. doi: 10.1198/016214501753382273

28. Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann Stat.* (2004) 32:928–61. doi: 10.1214/009053604000000256

29. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat.* (2004) 32:407–99. doi: 10.1214/009053604000000067

30. Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann Stat.* (2008) 36:1509–33. doi: 10.1214/009053607000000802

31. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B.* (2008) 70:849–911. doi: 10.1111/j.1467-9868.2008.00674.x

32. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

33. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY: Springer Science & Business Media (2009). doi: 10.1007/978-0-387-84858-7

34. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* Boca Raton, FL: Chapman & Hall; CRC Press (1993). doi: 10.1007/978-1-4899-4541-9

35. Zou H, Hastie T, Tibshirani R. On the "degrees of freedom" of the lasso. *Ann Stat.* (2007) 35:2173–92. doi: 10.1214/009053607000000127

36. Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regression. *Stat Sin.* (2014) 24:35–67.

37. Fan J, Guo S, Hao N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J R Stat Soc Ser B.* (2012) 74:37–65. doi: 10.1111/j.1467-9868.2011.01005.x

38. Carroll RJ, Ruppert D, Stefanski LA. Measurement *Error in Nonlinear Models Chapman and Hall London.* New York, NY: Springer-Verlag (1995). doi: 10.1007/978-1-4899-4477-1

39. Chatterjee S, Jafarov J. Prediction error of cross-validated lasso. *arXiv preprint arXiv:150206291.* (2015). doi: 10.48550/arXiv.1502.06291

40. Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regression. *Stat Sin.* (2016) 26:35–67. doi: 10.5705/ss.2014.042

41. Prentice RL, Huang Y, Neuhouser ML, Manson JE, Mossavar-Rahmani Y, Thomas F, et al. Associations of biomarker-calibrated sodium and potassium intakes with cardiovascular disease risk among postmenopausal women. *Am J Epidemiol.* (2017) 186:1035–43. doi: 10.1093/aje/kwx238

42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

43. Wasserman L, Roeder K. High dimensional variable selection. *Ann Stat.* (2009) 37:2178–201. doi: 10.1214/08-AOS646

44. Meinshausen N, Meier L, Bühlmann P. P-values for high-dimensional regression. *J Am Stat Assoc.* (2009) 104:1671–81. doi: 10.1198/jasa.2009.tm08647

45. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. *Ann Stat.* (2014) 42:413–68. doi: 10.1214/13-AOS1175