Journal of Data Science 23 (3), 499–520
July 2025
DOI: 10.6339/25-JDS1166
Statistical Data Science

Variable Selection with FDR Control for Noisy Data – An Application to Screening Metabolites that Are Associated with Breast Cancer and Colorectal Cancer

Runqiu Wang[1], Ran Dai[1,*], Ying Huang[2], Marian L. Neuhouser[2], Johanna W. Lampe[2], Daniel Raftery[3], Fred K. Tabung[4], and Cheng Zheng[1,*]

[1]*Department of Biostatistics, University of Nebraska Medical Center, Omaha, Nebraska, U.S.A.*
[2]*Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, Washington, U.S.A.*
[3]*Department of Anesthesiology and Pain Medicine, University of Washington, Seattle, Washington, U.S.A.*
[4]*Department of Internal Medicine, College of Medicine and Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, U.S.A.*

Abstract

The rapidly expanding field of metabolomics presents an invaluable resource for understanding the associations between metabolites and various diseases. However, the high dimensionality, presence of missing values, and measurement errors associated with metabolomics data can present challenges in developing reliable and reproducible approaches for disease association studies. Therefore, there is a compelling need for robust statistical analyses that can navigate these complexities to achieve reliable and reproducible disease association studies. In this paper, we construct algorithms to perform variable selection for noisy data and control the False Discovery Rate when selecting mutual metabolomic predictors for multiple disease outcomes. We illustrate the versatility and performance of this procedure in a variety of scenarios, dealing with missing data and measurement errors. As a specific application of this novel methodology, we target two of the most prevalent cancers among US women: breast cancer and colorectal cancer. By applying our method to the Women's Health Initiative data, we successfully identify metabolites that are associated with either or both of these cancers, demonstrating the practical utility and potential of our method in identifying consistent risk factors and understanding shared mechanisms between diseases.

Keywords *cancer; FDR control; measurement error; metabolomics data; missing data; variable selection*

1 Introduction

Breast cancer (BC) and colorectal cancer (CRC) have a high incidence rate, ranking as the highest and third highest among women in the US, respectively (ACS, 2020). Both cancers share several diet and lifestyle risk factors (Kampman et al., 2018). According to the World Cancer Research Fund (WCRF)/American Institute for Cancer Research (AICR) Expert Panel, there is "convincing" evidence that adult weight gain and excess body fat increase the risk for post-menopausal BC and CRC, and that physical activity reduces the risk for both cancers.

*Corresponding author. Email: ran.dai@unmc.edu or cheng.zheng@unmc.edu.

Received June 29, 2024; Accepted January 7, 2025

Furthermore, alcoholic drinks have been found to increase the risk of post-menopausal BC and CRC. Higher intakes of red meat, animal fats, and refined carbohydrates have been associated with increased risks of both BC and CRC, whereas fruits, vegetables, whole grains, and dietary fiber tend to be linked with reduced risk (Yusof et al., 2012; Xiao et al., 2019; Putri et al., 2013). However, the WCRF/AICR Expert Panel's classification for the level of evidence supporting the associations for these dietary components remains "suggestive" or "probable" rather than "convincing", except for increased intakes of processed meat and the risk of CRC (Kampman et al., 2018). Given this context, there is a crucial need to identify "convincing" evidence for risk factors associated with BC and CRC. Additionally, it is essential to study the common risk factors for these two prevalent cancers to better understand their shared underlying mechanisms and develop effective prevention strategies.

Metabolomics, the extensive analysis of small molecules in organisms (Nannini et al., 2020), reflects both internal cellular processes and external exposures, making it a sensitive tool for tracing pathways associated with chronic diseases like cancer. Despite its use in early cancer detection (Cheung et al., 2019; Yang et al., 2020; His et al., 2019; Zhu et al., 2014), few studies have systematically explored metabolomics in relation to BC and CRC. Identifying metabolomic components could uncover new BC pathways. For example, studies using the European Prospective Investigation into Cancer (EPIC) and the Prospective Lung, Colorectal, and Ovarian Cancer (PLCO) cohorts found certain plasma components and pre-diagnostic diet-related metabolites significantly associated with BC risk (His et al., 2019; Playdon et al., 2017). Further uncovering BC and CRC-related features could illuminate disease-related biological pathways, enhancing prevention and treatment strategies.

Developing screening methods for metabolomics data presents several challenges due to their inherent characteristics. Metabolomics data are high-dimensional, often containing missing values and measurement errors. The high dimensionality of metabolomics data is a double-edged sword: while it encompasses all potential components associated with the disease, the majority of these components are unrelated, introducing a significant amount of noise. Missingness and measurement errors are inevitable when dealing with large-scale data collection. These factors pose considerable challenges in screening procedures for the metabolomics data that offer reproducibility guarantees. To address these issues, robust and innovative approaches must be designed, which can account for the complexities and limitations associated with high-dimensional, noisy data while still providing accurate and reliable results.

Measurement errors and missing data frequently arise in complex data analysis tasks, posing challenges that must be carefully addressed when designing variable selection procedures. Naive approaches often lead to problematic results. For example, using complete case analysis and removing all samples with any missing data leads to spurious results in variable selection. Measurement errors lead to inflated estimation errors for coefficients and inconsistency in the Lasso variable selection procedure (Sorensen et al., 2015). There has been a surge in the literature on variable selection in the presence of missing data and measurement errors. For missing data mechanisms like missing at random (MAR) and missing completely at random (MCAR), researchers have developed imputation-based methods and other techniques (Little and Rubin, 2002; Tsiatis, 2006; Rässler et al., 2013) tailored for variable selection purposes (Wolfson, 2011; Johnson, 2008; Garcia et al., 2010). In the context of variable selection with measurement errors, several methods have been proposed, including CocoLasso (Datta and Zou, 2017), corrected Lasso (Loh and Wainwright, 2012; Sorensen et al., 2015), generalized matrix uncertainty selector (Rosenbaum and Tsybakov, 2013), generalized matrix uncertainty Lasso (Sorensen et al., 2015), and generalized Dantzig selector (Antoniadis et al., 2010). These advancements demonstrate

the ongoing efforts to tackle the challenges posed by missing data and measurement errors in variable selection.

One critical challenge in variable selection is ensuring replicability guarantees. Over the past few decades, a novel measure for Type I error, the False Discovery Rate (FDR), or the expectation of the false discovery proportion (FDP), has been proposed to address this issue. The renowned Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) has sparked a new era in multiple hypothesis testing, leading to the rapid development of methods that control FDR. Among these techniques, the Knockoff-based methods (Barber and Candès, 2015, 2019; Candès et al., 2018) offer several advantages, making it particularly attractive for various applications. Some of the key benefits of the Knockoff methods include: mild assumptions about data structure, allowing for more flexibility in handling diverse datasets; compatibility with a wide array of models and variable selection procedures for both low and high dimensional data, powerful method with a finite sample FDR control guarantee, providing reliable results even with limited sample sizes.

These advantages make the Knockoff methods especially suitable for applications in metabolomic data, which typically exhibit complex correlation structures, high-dimensional features, and unknown signal strength. Furthermore, the Knockoff methods excel at handling arbitrary correlation structures and do not require prior knowledge of signal amplitudes or noise levels, making them powerful and versatile tools in the realm of variable selection.

1.1 Prior Work

Knockoff-Based Methods Advancements in multiple testing problems within a single experiment have resulted in the development of powerful knockoff-based methods that provide exact FDR control for selecting features with conditional associations with the response (Barber and Candès, 2015; Candès et al., 2018). The knockoff filter by Barber and Candès (2015) offers exact FDR control for linear models without needing detailed model information and has been developed for high-dimensional cases (Barber and Candès, 2019). The Model-X knockoff by Candès et al. (2018) extends this to nonlinear models with unknown response distributions but requires knowledge of the predictor $\boldsymbol{X}$'s distribution. Barber et al. (2020) demonstrated that the Model-X knockoff method is robust to errors in estimating the distribution of $\boldsymbol{X}$, while Huang and Janson (2020) relaxed its assumptions, allowing FDR control as long as the parametric form of the distribution of $\boldsymbol{X}$ is known. A number of publications have explored the construction of knockoffs with approximated distributions of $\boldsymbol{X}$. For instance, Romano et al. (2020) developed a Deep knockoff machine using deep generative models, Liu and Zheng (2019) created a Model-X generating method employing deep latent variable models, and more recently, Bates et al. (2021) proposed an efficient general metropolized knockoff sampler. Spector and Janson (2022) suggested constructing knockoffs by minimizing the reconstructability of features. Knockoff-based methods have also been extended to test the intersection of null hypotheses, leading to the development of the group and multitask knockoff methods (Dai and Barber, 2016) and prototype group knockoff methods (Chen et al., 2019).

Current Advance in FDR Control for Identifying Simultaneous Signals Simultaneous signal detection has been explored using BH procedure-based methods (Heller et al., 2014; Bogomolov and Heller, 2013, 2018), local FDR (Chi, 2008; Heller and Yekutieli, 2014), and nonparametric approaches (Zhao and Nguyen, 2020). These methods rely on the independence, or positive regression dependency property of the features, which do not hold for most metabolomics

studies. Recently, Li et al. (2021) and Dai and Zheng (2023) introduced the multi-environment knockoff and the simultaneous knockoff methods for feature selection and identifying consistent associations, potentially useful for detecting mutual BC and CRC risk factors. However, all these methods do not allow the existence of missing data or measurement errors, which presents an important and unavoidable issue in metabolomic data analyses.

1.2 Our Contributions

In this paper, we evaluated the performance of different knockoff extension methods to handle missing values and/or measurement errors in the context of FDR control for variable selection. In addition, we propose a method to select mutual metabolomics predictors for not only one, but multiple clinical outcomes. The main contributions of this paper are summarized below:

1. We construct a knockoff-based procedure for FDR-controlled multiple testing when there are measurement errors and/or missing data in predictors. This procedure can work on general conditional dependence models $Y|\boldsymbol{X}$ and data structures in $\boldsymbol{X}$. It can also identify mutual signals for multiple outcomes (e.g. BC and CRC).
2. We demonstrate the FDR control performance and the power of our method with extensive simulation settings. We also illustrate the application with the Women's Health Initiative (WHI) data examples.

The rest of the paper is organized as follows. In Section 2, we present notations and details of our proposed variable selection framework to control FDR when there are missing data and measurement errors in the predictors. The method can also identify mutual signals for multiple outcomes. In Section 3, we show the empirical performance of the proposed method under different model assumptions and data structures. Finally, in Section 4, we apply the proposed method to a nested case-control study of BC and CRC among WHI Bone Mineral Density (BMD) Subcohort data.

2 Methods

2.1 Notations

For any positive integer N, denote $[N] = \{1, \ldots, N\}$. For the n data samples without missingness and measurement error, we assume they are sampled from the underlying distribution of $(Y, \boldsymbol{X})$ with $Y \in \mathbb{R}$ being the response variable and $\boldsymbol{X} \in \mathbb{R}^p$ being the p-dimensional predictor. The samples $(Y_i, X_{i1}, \ldots, X_{ip}) \overset{\text{iid}}{\sim} \mathcal{D}$, for $i \in [n]$. We work on the multiple testing problem on the null hypotheses $H_{0j} := Y \perp\!\!\!\perp X_j|\boldsymbol{X}_{-j}$, where $\boldsymbol{X}_{-j} := \{X_k : k \in [p] \text{ and } k \neq j\}$. We aim at developing a selection procedure returning a selection set $\widehat{\mathcal{S}} \subseteq [p]$ with a controlled FDR:

$$\text{FDR}(\widehat{\mathcal{S}}) = \mathbb{E}\big[\text{FDP}(\widehat{\mathcal{S}})\big] = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{H}|}{|\widehat{\mathcal{S}}| \vee 1}\right], \tag{1}$$

where $\mathcal{H} = \{j \in [p] : H_{0j} \text{ is true}\}$ and $\vee$ means taking the maximum of the two elements.

Given the potential measurement error, we assume $\boldsymbol{X}$ (the true serum/urine metabolites' level) is not available and an error-prone version $\boldsymbol{W} \in \mathbb{R}^p$ (measured metabolites that subject to various sources of measurement error) is available, where $\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{\epsilon}_w$ with $\boldsymbol{\epsilon}_w \overset{\text{iid}}{\sim} \mathcal{F}_w$ where $\mathcal{F}_w$ can be estimated from external data sources. Typically, we can assume a multivariate normal distribution for $\boldsymbol{\epsilon}_w$, i.e., $\boldsymbol{\epsilon}_w \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$. Also, for the potential missing data, we further introduce

indicator variables $\Delta \in \{0,1\}^p$ to indicate the missing mechanism. We let $\Delta_j = 1$ to indicate that W_j is observable and we adopt the missing at random (MAR) assumption, i.e., $\mathbb{P}(\Delta_j = 1|\boldsymbol{W}, \boldsymbol{X}) = \mathbb{P}(\Delta_j = 1|\boldsymbol{W}_{-j})$ where $\boldsymbol{W}_{-j} := \{W_k : k \in [p] \text{ and } k \neq j\}$. Notice that this assumption is slightly stronger than the usual assumption of MAR ($\mathbb{P}(\Delta_j = 1|\boldsymbol{W}, \boldsymbol{X}) = \mathbb{P}(\Delta_j = 1|\boldsymbol{X}_{-j})$) based on true variables due to the potential measurement error. In metabolomic data, this assumption is reasonable, since the missingness is mostly related to the signal detected from the other similar metabolite peaks, rather than the true underlying concentration of other metabolites.

For the remaining of the paper, with n observations, our final observed data will be denoted as $(\mathbf{Y}, \boldsymbol{\Delta}, \boldsymbol{\Delta} \odot \mathbf{W})$, where $\odot$ represents the Hadamard product (i.e., elementwise product). Here $\mathbf{Y} \in \mathbb{R}^n$ is a vector of responses for the n individuals, $\mathbf{W} \in \mathbb{R}^{n \times p}$ denotes the matrix with elements W_{ij} for individual i and predictor j, $\boldsymbol{\Delta} \in \{0,1\}^{n \times p}$ denotes the matrix with elements Δ_{ij} the indicator variable for individual i and predictor j.

2.2 Imputation of Missing Data

In general, we generate K imputed datasets, denoted as $(\mathbf{Y}, \mathbf{W}^k)$, for $k \in [K]$. When $K = 1$, we consider simple imputation using the mean of the observed values or half of the minimum of the observed values depending on our assumption on whether the missing is random or is due to a detection limit. Also, we can consider multiple imputation methods with $K \geqslant 1$ where each dataset can be generated using a chained equation approach. Specifically, we first randomly impute the missing values based on the marginal distribution estimated from those individuals with the variable observed, i.e.,

$$W_{ij}^k = \Delta_{ij} W_{ij} + (1 - \Delta_{ij}) \frac{\sum_i^n \Delta_{ij} W_{ij}}{\sum_i^n \Delta_{ij}}.$$

Then we will update the imputed value iteratively over all $j \in [p]$ that $\sum_i \Delta_{ij} < n$. First, we will fit a regression model of W_{ij} on $W_{i,-j}^k$ and Y_i for those $\Delta_{ij} = 1$. Then we will update W_{ij}^k for those $\Delta_{ij} = 0$ using the predicted value from the model and current $W_{i,-j}^k$, Y_i. Here the regression models can be in the form of generalized linear models (*default*), classification and regression trees (*cart*), or random forest (*rf*). Alternatively, when a large unlabeled subsample exists, another option is to perform multiple imputations excluding the outcome Y from the above steps. We explore both options numerically in Section 3.

2.3 Knockoff Construction

For each imputed dataset $\mathbf{W}^k$, we construct the knockoff $\widetilde{\mathbf{W}}^k$ using second-order Model-X knockoff by sampling $\widetilde{\mathbf{W}}^k$ from $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$, where $\widetilde{\boldsymbol{\mu}} = \mathbf{W}^k - \mathbf{W}^k \boldsymbol{\Sigma}^{-1} \operatorname{diag}\{\mathbf{s}\}$, $\widetilde{\boldsymbol{\Sigma}} = 2\operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\}$ and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\mathbf{W}$, such that

$$Cov\left(\left[\mathbf{W}^k, \widetilde{\mathbf{W}^k}\right]\right) = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\} \\ \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\} & \boldsymbol{\Sigma} \end{pmatrix}.$$

where $\mathbf{s}$ satisfies $\widetilde{\boldsymbol{\Sigma}} = 2\operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\}$ is semi-positive definite and $\mathbf{s}$ can be solved using the approximate semidefinite program (ASDP) algorithm as given in Candès et al. (2018). We use the R function *create.second* within the R package *knockoff* to implement this construction method. As a remark, the second order knockoff method works primarily for normally distributed data, and the multivariate Gaussian approximation is reasonable after the log-transformation of the metabolite data; the method is also robust against mild model mis-specifications (Candès et al., 2018).

2.4 Test Statistics

One advantage of the Knockoff procedure we adopt for FDR control is that we do not need to know the null distribution of our test statistics, as long as they are compatible with the Knockoff method. We consider a variety of test statistics:

- **Lasso:** We assume a working model in a generalized linear model (GLM) framework

$$f_Y(y;\theta,\phi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right\},$$

where $\theta=\mathbf{X}^\top\boldsymbol{\beta}$. Here $\boldsymbol{\beta}$ is the parameter of interest and ϕ is the dispersion parameter while $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are prespecified functions. The expected response is given by the mean function $\mu(\theta)=b'(\theta)=g^{-1}(\theta)$, where $g^{-1}(\cdot)$ is the inverse of a canonical link function $g(\cdot)$. We choose the Lasso variable selection procedure and construct the statistics as

$$\widehat{\boldsymbol{\beta}}(\lambda)=\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{2p}}\sum_{i=1}^{n}\frac{(Y_i-\mu([\mathbf{W}_i\widetilde{\mathbf{W}}_i]^\top\boldsymbol{\beta}))^2}{V_i}+\lambda\|\boldsymbol{\beta}\|_1,$$

where $V_i=V(g^{-1}([\mathbf{W}_i\widetilde{\mathbf{W}}_i]^\top\boldsymbol{\beta}))$ and $V(\cdot)$ is the variance function specified for the GLM for Y. Then we use the absolute value $|\widehat{\beta}_j(\lambda)|$ as defined above with a specific λ value or λ selected from cross-validation as test statistics (i.e., $Z_j=|\widehat{\beta}_j(\lambda)|$ and $\widetilde{Z}_j=|\widehat{\beta}_{p+j}(\lambda)|$ for $j\in[p]$).

- **Lasso Order:** We assume the same GLM model and run over a range of λ values decreasing from $+\infty$ (a fully sparse model) to 0 (a fully dense model) and define Z_j ($\widetilde{Z}_j$) as the maximum λ such that $\widehat{\beta}_j(\lambda)\neq 0$ ($\widehat{\beta}_{p+j}(\lambda)\neq 0$). If there is no λ such that $\widehat{\beta}_j(\lambda)\neq 0$ ($\widehat{\beta}_{p+j}(\lambda)\neq 0$), then we will simply define Z_j ($\widetilde{Z}_j$) as 0.
- **Random Forest (RF):** We use the variable importance factors from the random forest fitting of $\mathbf{Y}$ on $[\mathbf{W}\ \widetilde{\mathbf{W}}]$ with either fixed tuning parameters or tuning parameters selected from cross-validation as $[\ \mathbf{Z}\ \ \widetilde{\mathbf{Z}}\ ]$.
- **Generalized Dantzig Selector (GDS):** We choose the GDS variable selection procedure and construct the statistics as $\widehat{\boldsymbol{\beta}}_{\mathrm{DS}}(\lambda)$, where

$$\widehat{\boldsymbol{\beta}}_{\mathrm{DS}}(\lambda)=\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{2p}}\Bigg[\|\boldsymbol{\beta}\|_1:\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}W_{ij}\big\{Y_i-\mu\big(\big[\mathbf{W}_i\widetilde{\mathbf{W}}_i\big]^\top\boldsymbol{\beta}\big)\big\}\right|\leqslant\lambda\text{ and }\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{W}}_{ij}\big\{Y_i-\mu\big(\big[\mathbf{W}_i\widetilde{\mathbf{W}}_i\big]^\top\boldsymbol{\beta}\big)\big\}\right|\leqslant\lambda\Bigg].$$

- **Generalized Matrix Uncertainty Selector (GMUS):** The test statistics is a feasible solution of

$$\widehat{\boldsymbol{\beta}}_{\mathrm{MU}}(\lambda)=\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{2p}}\left\{\|\boldsymbol{\beta}\|_1:\frac{1}{n}\left\|\sum_{i=1}^{n}\big[\mathbf{W}_i\widetilde{\mathbf{W}}_i\big]^\top\big\{Y_i-\mu\big(\big[\mathbf{W}_i\widetilde{\mathbf{W}}_i\big]^\top\boldsymbol{\beta}\big)\big\}\right\|_\infty\leqslant\lambda+\delta\|\boldsymbol{\beta}\|_1\right\}.$$

- **Corrected Lasso:** The test statistics can be defined as minimizing the loss

$$\widehat{\boldsymbol{\beta}}_{\mathrm{RCL}}(d)=\operatorname*{arg\,min}_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_1\leqslant d}\left\{\sum_{i=1}^{n}\frac{(Y_i-\mu([\mathbf{W}_i\widetilde{\mathbf{W}}_i]^\top\boldsymbol{\beta}))^2}{V_i}-\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_\epsilon\boldsymbol{\beta}\right\},$$

where $\boldsymbol{\Sigma}_\epsilon$ is the variance-covariance matrix for the measurement error in $\mathbf{W}$, and d can be a pre-fixed tuning parameter or selected from cross-validation.

As a remark, when measurement errors exist, the Corrected Lasso and GMUS methods are known to handle measurement errors appropriately for coefficient estimation and thus the knockoff approach can be applied to control the FDR. The performance of other test statistics is expected to be affected by the measurement errors. The impact of the scales and correlations of measurement errors will be evaluated numerically (see Section 3).

2.5 Variable Selection

We first compute the p-value for each feature based on the formula $p_j = \frac{1+\sum_{k=1}^{K} \mathbb{1}\{Z_j^k \leqslant \widetilde{Z}_j^k\}}{1+K}$ that combine the information from different imputed datasets. Here $\mathbb{1}\{\cdot\}$ is the indicator function which takes value 1 if the event holds and 0 otherwise. Then we reorder the feature indices $j \in [p]$ with the decreasing order of $\max_{k=1}^{K} \max\{|Z_j^k|, |\widetilde{Z}_j^k|\}$ and we denote the new index of the original feature j as $\zeta(j)$. We apply the Selective SeqStep and Selective SeqStep+ procedures (Barber and Candès, 2015) with p-value threshold 0.5 to find the selection threshold $\widehat{k}_c$ such that

$$\widehat{k}_c = \max\left\{ j \in [p] : \frac{c + \sum_{k=1}^{j} \mathbb{1}\{p_{\zeta^{-1}(k)} > 1/2\}}{\sum_{k=1}^{j} \mathbb{1}\{p_{\zeta^{-1}(k)} \leqslant 1/2\} \vee 1} \leqslant q \right\},$$

for $c = 0$ (SeqStep), 1 (SeqStep+), where q is the FDR level to be controlled at. Then the final selection sets will be $\widehat{S}_c = \{j : p_j < 1/2\} \cap \{j : \zeta(j) \in [\widehat{k}_c]\}$.

Notice that when the imputation model is correct, and there is no measurement error, i.e., $\mathbf{W} = \mathbf{X}$, then we have the imputed datasets $(\mathbf{Y}, \mathbf{W}^k) \sim \mathcal{D}$ for each k, thus for our knockoff construction and test statistics, it has been shown that $\mathbb{P}(\mathbf{Z}_j^k < \widetilde{\mathbf{Z}}_j^k) = 0.5$ for $j \in \mathcal{H}$ (Barber and Candès, 2015; Candès et al., 2018). Since p_j is symmetrically distributed and $\mathbb{E}[p_j] = \frac{1+K/2}{1+K} = \frac{1}{2} + \frac{1}{2(1+K)} > \frac{1}{2}$ for $j \in \mathcal{H}$, therefore the estimated FDP based on the threshold $\hat{k}_c$, i.e., $\widehat{\mathrm{FDP}} = \frac{c+\sum_{k=1}^{\hat{k}_c} \mathbb{1}\{p_{\zeta^{-1}(k)}>1/2\}}{\sum_{k=1}^{\hat{k}_c} \mathbb{1}\{p_{\zeta^{-1}(k)}\leqslant 1/2\}\vee 1}$ is a conservative estimate for the FDP and thus the Selective SeqStep+ procedure controls FDR (Barber and Candès, 2015). Although the increasing number of imputed dataset K could lead to better p_j estimations, in practice, we don't need very large K since we only need to know whether p_j is below 0.5 or not and the accurate p_j is not necessary when it is far from 0.5. In practice, due to the potential model misspecification and finite sample performance, the imputed distribution might be slightly different from the true distribution, nonetheless, the robustness of the knockoff approach (Barber et al., 2020) ensures that the FDR inflation is small.

2.6 Joint Selection for Multiple Outcomes

Now we consider the mutual signal identification problem. Assume we have data from M independent experiments and denote $[M] = \{1, \ldots, M\}$. Within the m-th experiment, the underlying complete data without measurement errors are $(Y_i^m, X_{i1}^m, \ldots, X_{ip}^m) \overset{\text{iid}}{\sim} \mathcal{D}_m$, $i = 1, \ldots, n_m$. In our setting, the outcome variables Y^1, Y^2 represent the two different cancer outcomes BC and CRC. Define H_{0j}^m as the null hypothesis indicating the j-th feature not being a signal in the m-th experiment (i.e. $X_j^m \perp\!\!\!\perp Y^m | \mathbf{X}_{-j}^m$ where $\mathbf{X}_{-j}^m := \{X_k^m : k \in [p] \text{ and } k \neq j\}$), and denote $\mathcal{H}^m = \{j \in [p] : H_{0j}^m \text{ is true}\}$, where $[p] := \{1, \ldots, p\}$. Instead of testing the H_{0j}^m's, we are interested in testing the union null hypotheses $H_{0j} = \bigcup_{m=1}^{M} H_{0j}^m$, for $j \in [p]$.

We define

$$\mathcal{S} = \{j \in [p] : H_{0j} \text{ is false}\} \quad \text{and} \quad \mathcal{H} = \mathcal{S}^c = \bigcup_{m=1}^{M} \mathcal{H}^m = \{j \in [p] : H_{0j} \text{ is true}\}.$$

We aim at developing a selection procedure returning a selection set $\widehat{\mathcal{S}} \subseteq [p]$ with a controlled FDR, as defined in (1). For this task, we can get test statistics Z_j^{km} and $\widetilde{Z}_j^{km}$ for each m separately first as the above sections and then compute the p-values as

$$p_j = \frac{1 + \sum_{k=1}^{K} \mathbb{1}\{\prod_{m=1}^{M}(Z_j^{km} - \widetilde{Z}_j^{km}) \leqslant 0\}}{1 + K}.$$

Then we can use this new p value to get $\widehat{k}_c$ and $\widehat{S}_c$ reordering the feature index by the decreasing order of $\max_{k=1}^{K} \prod_{m=1}^{M} |Z_j^{km} - \widetilde{Z}_j^{km}|$ or $\sum_{k=1}^{K} \prod_{m=1}^{M} |Z_j^{km} - \widetilde{Z}_j^{km}|$. This approach is valid under a similar argument as in Dai and Zheng (2023).

3 Simulation

In this section, we perform extensive numerical experiments to understand the finite sample performance of the proposed methods in Section Methods 2.

3.1 Simulation for Gaussian Distributions

We first show the performance of proposed methods in Section Methods 2 when predictors are sampled from multivariate Gaussian distributions.

3.1.1 Data Generation and Settings

We generate data with various sample sizes n and dimension p. We sample the underlying feature matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with various variation and correlation settings; and the outcome $\mathbf{Y} \in \mathbb{R}^n$ from a sparse logistic regression with varied effect sizes. Then we sample the measurement errors $\boldsymbol{\epsilon}_w$ and construct features with measurement errors $\mathbf{W} = \mathbf{X} + \boldsymbol{\epsilon}_w$ with different measurement error scales and correlations. Next, we sample the missing data indicators $\boldsymbol{\Delta}$ under the missing at random (MAR) mechanism with varied missing proportion p_{mis}. Under MAR assumption, we consider both missing probabilities dependent on error-prone variables $\mathbf{W}$ (Type W) and the error-free variables $\mathbf{X}$ (Type X).

We run 200 simulations under each of the following three different settings:

1. **Data with only missing data but not measurement errors.** Under this setting, we compare the performance of the following methods: Lasso, Lasso Order, and RF with multiple imputations ($K = 5$ imputed datasets). We considered either including or excluding the outcome $\mathbf{Y}$ (Imp Y) when performing the imputations and considered three different imputation methods (Imp M): R package MICE with default method (*default*), classification and regression tree (*cart*), and random forest (*rf*).
2. **Data with only measurement errors but not missing data.** We compare the performance of our method with the following statistics as described in Section Test statistics 2.4: Lasso, Lasso Order, RF, GDS, GMUS, and Corrected Lasso.
3. **Data with both missing data and measurement errors.** We compare the performance of our method with the following statistics as described in Section Test statistics 2.4: Lasso, Lasso Order, RF, GDS, GMUS, and Corrected Lasso with multiple imputations.

More details on the data generation and simulation settings are postponed in Web Appendix A.1.

3.1.2 Results

For Setting 1 with only missing data, we compare the performance of three variable selection methods: Lasso, Lasso Order, and RF (Table 1). Across the variety of settings we experimented with, all three variable selection methods effectively control the FDR empirically. The Lasso variable selection method demonstrates the highest power among the tested methods, followed by Lasso Order, and finally, RF. The RF method tends to be conservative in selecting variables, resulting in slightly lower power, which may be attributed to the Lasso method possessing the correctly specified model. The prediction performance of the RF method is satisfactory (Table S1). As the sample size (n) decreases and the number of variables (p) increases, all methods maintain satisfactory FDR control, albeit with a slight reduction in power, as anticipated. The proportion of missing data does not significantly impact the performance of these methods in terms of FDR and power. When examining the three imputation methods, their performances are found to be relatively similar. For smaller sample sizes ($n = 400$), the *default* method offers marginally better power, while the *rf* method slightly outperforms the others for larger sample sizes. The decision to include the dependent variable (Y) in the imputation model does not substantially alter the performance of these methods.

In Table 2, we present the FDP and power for our experiments conducted on data with measurement errors. We compare various variable selection methods, including Lasso, Lasso Order, RF, GDS, Corrected Lasso, and GMUS. Notably, the Corrected Lasso and GMUS methods are specifically designed for measurement error correction. The results are organized according to the number of variables p, effect size, and scale. For FDP, Lasso, Lasso Order, RF, and GDS methods exhibit similar values across all scenarios. When p is small and the effect size is large, the FDP is marginally higher than the nominal FDR. Corrected Lasso consistently displays lower FDP values than the nominal FDR, but it tends to be slightly over-conservative. The GMUS method generally yields FDP values comparable to those of Lasso and Lasso Order. Regarding power, the Lasso method persistently achieves higher values compared to other methods across all settings. The GDS and Lasso Order methods also demonstrate satisfactory power, while the RF method consistently exhibits lower power. The GMUS method's power performance is comparable to that of Lasso Order and GDS, while Corrected Lasso consistently shows the lowest power values among all methods. Overall, as the number of variables increases, FDP experiences a slight decrease and power undergoes a more noticeable reduction. The effect size has a more pronounced impact on power, with larger effect sizes resulting in higher power values. The scale also affects power, with smaller scales generally leading to increased power values. However, the FDP values remain relatively stable irrespective of changes in effect size or scale.

Table 3 summarizes the FDR and power of our proposed methods under simulation settings with both measurement errors and missing data. The variable selection methods we included are the same as in Setting 2. We set the sample size $n = 1000$ and consider two settings for the dimension p: a low-dimensional setting with $p = 60$ and a high-dimension setting with $p = 210$. For the measurement errors, we consider two levels of Scales: $\sigma_\epsilon^2 = 0.1$ or 0.6. For imputation methods, we consider the default, cart, and rf. We compare the performance based on whether to include outcome Y in the imputation model. The variable selection methods Lasso, Lasso Order, RF, and GDS do not consider measurement errors, while Corrected Lasso and GMUS correct measurement errors. Here, we present the result in Table 3 where the missing probabilities

Table 1: Simulation results (FDP and Power) for Setting 1 (missing data) varying n, p and p_{mis}; with different imputation methods (Imp M) and choice of whether to include $\mathbf{Y}$ in the imputation (Imp Y).

					FDP			Power		
n	p	p_{mis}	Imp M	Imp Y	Lasso	Lasso Order	RF	Lasso	Lasso Order	RF
400	60	0.05	cart	yes	0.15	0.14	0.14	0.99	0.65	0.46
400	60	0.05	cart	no	0.16	0.15	0.14	0.98	0.67	0.50
400	60	0.05	default	yes	0.18	0.13	0.15	0.99	0.64	0.50
400	60	0.05	default	no	0.18	0.13	0.16	0.99	0.66	0.52
400	60	0.05	rf	yes	0.17	0.14	0.14	0.99	0.66	0.50
400	60	0.05	rf	no	0.16	0.15	0.13	0.99	0.65	0.49
400	60	0.15	cart	yes	0.17	0.14	0.17	0.98	0.64	0.54
400	60	0.15	cart	no	0.15	0.14	0.13	0.98	0.67	0.47
400	60	0.15	default	yes	0.17	0.15	0.13	0.99	0.68	0.50
400	60	0.15	default	no	0.17	0.17	0.14	0.99	0.70	0.49
400	60	0.15	rf	yes	0.17	0.15	0.15	0.98	0.67	0.50
400	60	0.15	rf	no	0.18	0.17	0.16	0.98	0.70	0.52
1000	60	0.05	cart	yes	0.17	0.18	0.18	1.00	0.94	0.76
1000	60	0.05	cart	no	0.17	0.19	0.15	1.00	0.95	0.72
1000	60	0.05	default	yes	0.20	0.17	0.16	1.00	0.93	0.70
1000	60	0.05	default	no	0.18	0.19	0.16	1.00	0.94	0.73
1000	60	0.05	rf	yes	0.19	0.17	0.17	1.00	0.92	0.74
1000	60	0.05	rf	no	0.18	0.17	0.17	1.00	0.93	0.73
1000	60	0.15	cart	yes	0.18	0.20	0.18	1.00	0.94	0.73
1000	60	0.15	cart	no	0.19	0.20	0.18	1.00	0.94	0.74
1000	60	0.15	default	yes	0.20	0.18	0.16	1.00	0.94	0.73
1000	60	0.15	default	no	0.19	0.18	0.18	1.00	0.93	0.73
1000	60	0.15	rf	yes	0.22	0.22	0.18	1.00	0.94	0.75
1000	60	0.15	rf	no	0.21	0.21	0.18	1.00	0.93	0.74
1000	120	0.05	cart	yes	0.18	0.17	0.17	1.00	0.85	0.58
1000	120	0.05	cart	no	0.18	0.16	0.16	1.00	0.84	0.59
1000	120	0.05	default	yes	0.19	0.15	0.16	1.00	0.84	0.57
1000	120	0.05	default	no	0.19	0.16	0.15	1.00	0.84	0.56
1000	120	0.05	rf	yes	0.18	0.16	0.16	1.00	0.85	0.58
1000	120	0.05	rf	no	0.18	0.17	0.16	1.00	0.86	0.58
1000	120	0.15	cart	yes	0.19	0.18	0.17	1.00	0.85	0.59
1000	120	0.15	cart	no	0.17	0.18	0.17	1.00	0.85	0.58
1000	120	0.15	default	yes	0.20	0.19	0.16	1.00	0.87	0.58
1000	120	0.15	default	no	0.18	0.18	0.18	1.00	0.84	0.60
1000	120	0.15	rf	yes	0.20	0.18	0.19	1.00	0.85	0.62
1000	120	0.15	rf	no	0.19	0.19	0.18	1.00	0.85	0.59

Table 2: Simulation results (FDP and Power) for Setting 2 (data with measurement errors) for $n = 1000$, varying p, scales of the effect (A_β) and errors (σ_ϵ^2).

p	A_β	σ_ϵ^2	Lasso	Lasso Order	RF	GDS	Corrected Lasso	GMUS
FDP								
60	0.5	0.6	0.23	0.21	0.22	0.21	0.13	0.21
60	0.5	1	0.23	0.22	0.20	0.21	0.19	0.22
60	1.5	0.6	0.25	0.25	0.24	0.22	0.15	0.22
60	1.5	1	0.25	0.25	0.28	0.23	0.25	0.25
120	0.5	0.6	0.21	0.21	0.17	0.20	0.05	0.20
120	0.5	1	0.18	0.19	0.16	0.18	0.14	0.17
120	1.5	0.6	0.20	0.20	0.17	0.18	0.07	0.17
120	1.5	1	0.20	0.20	0.20	0.20	0.10	0.19
Power								
60	0.5	0.6	0.91	0.84	0.71	0.91	0.44	0.89
60	0.5	1	0.88	0.79	0.61	0.87	0.49	0.83
60	1.5	0.6	0.94	0.88	0.75	0.94	0.44	0.92
60	1.5	1	0.93	0.87	0.75	0.92	0.63	0.90
120	0.5	0.6	0.80	0.72	0.49	0.78	0.17	0.76
120	0.5	1	0.70	0.59	0.40	0.67	0.25	0.64
120	1.5	0.6	0.83	0.77	0.54	0.81	0.15	0.80
120	1.5	1	0.76	0.65	0.49	0.74	0.25	0.72

depend on **W** rather than **X** with smaller scales. The additional simulation results can be found in Web Appendix A.2. Tables S2 and S3.

With the small scale of measurement errors (Scale = 0.1), the FDRs for most methods are controlled close to or below the nominal value of 0.2, with Corrected Lasso consistently having the lowest FDR values across all conditions. The results are observed for both low and high dimensions, Type W and Type X, and for different Imp M and Imp Y conditions. With bigger scale measurement errors (Scale = 0.6), for Type W, the FDR is under control for most of the methods except Lasso Order and RF, while for Type X, only the RF and corrected Lasso method controls FDR while the other methods fail. In terms of power, all methods show a decrease as the dimension increases. When the measurement error is small, Lasso, GDS, and GMUS maintain a high power of 0.9, whereas Lasso Order and RF experience a significant decline in power. Corrected Lasso has relatively low power across all settings. When the measurement error is larger, all methods show a substantial power reduction. Overall, Lasso, GDS and GMUS achieve the best performance in terms of power, followed by Lasso Order and RF, while Corrected Lasso consistently exhibits poor power. To demonstrate the performance gain with our proposed methods to work with missing data and measurement errors for Setting 3, we further compared our methods with directly applying Lasso with knockoff (targeted $q = 0.2$) to the dataset, and using minimal value imputation to treat the missing data (Oracle). The Oracle method has satisfactory power but fails in controlling the FDR. With $\sigma_\epsilon^2 = 0.1$, this Oracle method has power = 0.95 and FDP = 0.33; with $\sigma_\epsilon^2 = 0.6$, this Oracle method has power = 0.87 and FDP = 0.34. More details on this comparison can be found in Web Appendix A.2.

Table 3: Simulation results (FDP and Power) for Setting 3 (data with both measurement errors and missing data) for $n = 1000$, $\sigma_\epsilon^2 = 0.1$, $p_{\text{mis}} = 0.15$, and $A_{\boldsymbol{\beta}} = 1$, varying p, Imp M and Imp Y when the missing probability depends on the error-prone variables $\mathbf{W}$.

p	Imp M	Imp Y	Lasso	Lasso Order	RF	GDS	Corrected Lasso	GMUS
FDP								
60	default	yes	0.18	0.19	0.15	0.20	0.02	0.15
60	default	no	0.16	0.20	0.18	0.19	0.02	0.13
60	cart	yes	0.16	0.20	0.18	0.19	0.02	0.13
60	cart	no	0.18	0.18	0.17	0.17	0.03	0.12
60	rf	yes	0.17	0.19	0.17	0.19	0.02	0.14
60	rf	no	0.17	0.20	0.15	0.19	0.02	0.14
210	default	yes	0.20	0.15	0.15	0.17	0.00	0.16
210	default	no	0.18	0.16	0.15	0.17	0.00	0.19
210	cart	yes	0.18	0.17	0.16	0.17	0.00	0.17
210	cart	no	0.18	0.17	0.16	0.18	0.00	0.19
210	rf	yes	0.19	0.18	0.17	0.18	0.00	0.18
210	rf	no	0.19	0.18	0.17	0.17	0.00	0.18
Power								
60	default	yes	1.00	0.92	0.80	1.00	0.46	0.99
60	default	no	1.00	0.92	0.79	1.00	0.42	0.98
60	cart	yes	1.00	0.92	0.80	1.00	0.44	0.98
60	cart	no	1.00	0.91	0.80	1.00	0.43	0.98
60	rf	yes	1.00	0.91	0.80	1.00	0.42	0.98
60	rf	no	1.00	0.92	0.78	1.00	0.42	0.98
210	default	yes	0.91	0.63	0.44	0.88	0.000	0.89
210	default	no	0.90	0.66	0.46	0.87	0.000	0.89
210	cart	yes	0.90	0.65	0.45	0.87	0.000	0.88
210	cart	no	0.90	0.65	0.45	0.87	0.001	0.88
210	rf	yes	0.91	0.68	0.48	0.89	0.002	0.88
210	rf	no	0.90	0.67	0.47	0.88	0.000	0.88

In Table S5 of Web Appendix A.3, we show the performance of proposed methods in Settings 2 and 3 for detecting mutual signals from 2 datasets. The simultaneous knockoff method with our procedure to handle missing data and measurement errors (GMUS) still controls the FDR and achieves comparable power.

3.2 Simulation from Empirical Data Distributions

In this section, we perform numerical experiments based on the real metabolomics data from the LC-MS platform (RelQuant).

3.2.1 Data generation and settings

We generate data with the same sample size $n = 1331$ and dimension $p = 148$ as the real data in the LC-MS platform (RelQuant). The data contains both missing values and measurement

errors. We approximate the distribution of **W** from the empirical distribution of real data (details can be found in Web Appendix A.4).

The outcome Y is a binary outcome from a logistic regression. The missing data indicators R are also sampled under the missing at random (MAR) mechanism with proportion $p_{\text{mis}} = 0.15$. The same with the Gaussian data simulation (setting for data with both missing data and measurement errors), we compare the performance of the following methods: Lasso, Lasso Order, RF, GDS, GMUS, and Corrected Lasso with multiple imputations.

We run 200 simulations under each of the settings. More details on the data generation and simulation settings are postponed in Web Appendix A.4.

3.2.2 Results

Table 4 illustrates the FDR and power under the empirical real data conditions, including missing values and measurement errors, for scenarios where missing probabilities depend on **W** (Type W) rather than **X** (Type X). Additional simulation results, focusing on missing probabilities dependent on **X**, are available in Web Appendix A.4. We also examine two levels of measurement error scales (σ_ϵ^2), and consider the three imputation methods (Imp M): *default*, *cart*, and *rf*, as well as evaluate the impact of imputing Y (Imp Y) on performance.

In terms of FDR, with small measurement errors (scale = 0.1), Lasso, Lasso Order, RF, Corrected Lasso, and GMUS effectively controlled the FDR, whereas GDS do not. At a larger error scale (scale = 0.5), only RF and Corrected Lasso strictly maintain the FDR below 0.2. GMUS managed to keep the FDR around 0.2, but Lasso, Lasso Order and GDS were unsuccessful in controlling the FDR. Regarding power, Lasso, GDS, and GMUS exhibit superior performance, followed by RF. However, Corrected Lasso demonstrates the limited power. These results were consistent under various imputation conditions (Imp M and Imp Y).

A comparative analysis of the performance of six variable selection methods between Gaussian distributed data and empirical data distribution from the LC-MS platform reveals inconsistencies in the Lasso method performance. Specifically, Lasso effectively controls FDR in some conditions (Type W) for Gaussian data but fails to do so for empirical data distributions. Different imputation methods (Imp M) and whether considering outcome variable Y (Imp Y) during the imputation does not impact the performance significantly. Overall, when missing values but no measurement error exist in the data, the Lasso and Lasso order methods perform the best; when measurement errors exist in the data, the GMUS method performs the best with the highest power, and the controlled FDR under the assumption that the missing probability depends on the error-prone variables **W**. In our real data analysis, this missing mechanism assumption is likely to hold, so, GMUS is considered as one of the primary methods in addition to Lasso and Lasso Order for the real data analysis.

4 Real Data Analysis

We analyzed serum and urine specimens from the WHI BMD data (181 CRC cases; 577 BC cases; 758 matched controls) using several metabolomics platforms. The details on how the matched samples were selected can be found in Web Appendix B.1. We applied global metabolomics platforms (gas chromatography–mass spectrometry (GC-MS) and nuclear magnetic resonance (NMR)) for profiling urine metabolites and targeted platforms for profiling serum metabolites. In serum, using liquid chromatography with tandem mass spectrometry (LC-MS/MS), we targeted

Table 4: Simulation results (FDP and Power) for Empirical Data Distribution from LC-MS platform (RelQuant) for $n = 1331$, $p = 148$, $p_{\text{mis}} = 0.15$, and $A_{\boldsymbol{\beta}} = 1$, varying σ_ϵ^2, Imp M and Imp Y when the missing probability depends on the error-prone variables $\mathbf{W}$.

σ_ϵ^2	Imp M	Imp Y	Lasso	Lasso Order	RF	GDS	Corrected Lasso	GMUS
FDP								
0.1	default	yes	0.20	0.19	0.13	0.21	0.04	0.19
0.1	default	no	0.18	0.19	0.14	0.20	0.04	0.18
0.1	cart	yes	0.18	0.18	0.14	0.20	0.04	0.19
0.1	cart	no	0.18	0.18	0.15	0.20	0.05	0.19
0.1	rf	yes	0.19	0.19	0.14	0.21	0.04	0.20
0.1	rf	no	0.19	0.20	0.13	0.22	0.03	0.20
0.5	default	yes	0.24	0.23	0.14	0.23	0.07	0.19
0.5	default	no	0.23	0.22	0.14	0.22	0.06	0.19
0.5	cart	yes	0.21	0.20	0.14	0.23	0.07	0.20
0.5	cart	no	0.22	0.22	0.14	0.22	0.07	0.20
0.5	rf	yes	0.23	0.24	0.16	0.22	0.07	0.21
0.5	rf	no	0.22	0.24	0.15	0.23	0.07	0.21
Power								
0.1	default	yes	1.00	0.94	0.71	1.00	0.16	1.00
0.1	default	no	1.00	0.93	0.72	1.00	0.14	1.00
0.1	cart	yes	1.00	0.93	0.71	1.00	0.14	1.00
0.1	cart	no	1.00	0.93	0.71	1.00	0.15	1.00
0.1	rf	yes	1.00	0.93	0.71	1.00	0.14	1.00
0.1	rf	no	1.00	0.93	0.71	1.00	0.16	1.00
0.5	default	yes	1.00	0.94	0.71	1.00	0.29	0.99
0.5	default	no	1.00	0.93	0.69	0.99	0.29	0.99
0.5	cart	yes	0.99	0.92	0.70	0.99	0.32	0.99
0.5	cart	no	1.00	0.93	0.70	0.99	0.29	0.99
0.5	rf	yes	0.99	0.94	0.71	0.99	0.33	0.99
0.5	rf	no	0.99	0.94	0.71	0.99	0.31	0.99

water-soluble metabolites covering over 50 major metabolic pathways, and using the recently developed Lipidyzer platform, we detected about 900 lipids from 13 different classes: Cholesterol ester (CE), Ceramides (CER), Diacylglycerol (DAG), Dihydroceramides (DCER), Free fatty acids (FFA), Hexosylceramides (HCER), Lactosylceramide (LCER), Lysophosphatidylcholine (LPC), Lysophosphatidylethanolamine (LPE), Phosphatid-ylcholine (PC), Phosphatidylethanolamine (PE), Sphingomyelin (SM), Triacylglycerol (TAG). Over 1500 metabolites were obtained from urine and serum samples using these four complementary analytical platforms. The proportion of missing data as well as the signal noise ratio for measurement errors are summarized in Table S7 in Web Appendix B where we can see that the GC-MS and LC-MS suffer most from the measurement error with some SNR less than one and GC-MS has the most missing data. We applied the knockoff method with missing and measurement errors to find the metabolite factors associated with BC, CRC, and shared factors for both cancers using the proposed methods.

4.1 Data Preprocessing

We first preprocessed the data from the four different platforms (NMR, GC-MS, LC-MS, and lipidomic) to remove outliers, batch effects, and variables with excessive missing values (details see Web Appendix B.2). For LC-MS and lipidomic data, we considered both concentration and composition data which led to a total of 6 groups of metabolites and we analyze each group separately. The binary variable case/control of certain cancers served as our response variable. Here the preprocessed metabolites in non-quality control (QC) samples form the predictor matrix $\mathbf{W} \in \mathbb{R}^{n \times p}$, where n is the number of patients and p is the number of metabolites.

To achieve the goals of finding the risk metabolite factors for BC and CRC, we analyzed the data including the BC cases vs all controls, and the CRC cases vs all controls respectively. Then the summary statistics were combined using the method described in Section Joint selection for multiple outcomes to identify the shared factors associated with both breast and CRC.

To build the model, first, for each platform dataset, we imputed the data by different imputation methods including half-min imputation, and multiple imputations with predictive mean matching method. For the multiple imputations, we performed MICE with ($K = 5$) using different analytic data for each outcome (i.e., BC cases + all controls for BC analysis and CRC + all controls for CRC analysis) were used to impute the missing values, and outcomes were included in the multiple imputation procedure. The variance-covariance matrix for the measurement errors was estimated using QC samples (details can be found in Web Appendix B.3). Then, we applied three preferred methods (Lasso, Lasso Order, GMUS) and three alternative methods (RF, GDS, Corrected Lasso) as described in Section Methods to generate test statistics. Then we applied our knockoff and simultaneous knockoff variable selection procedures with a target FDR level of 0.1. To make our results more robust, we performed stability selection by running different methods 100 times and recording the percentage of the replications each variable was selected.

For sensitivity analysis, we performed the same analysis as above but using different analytical data for each outcome (i.e., BC cases + matched BC control for BC analysis and CRC cases + matched CRC control for CRC analysis).

4.2 Results

We presented the metabolites that were selected to be associated with BC, CRC, and both of these two cancers with $\geqslant$ 50% of the replications in Tables 5, 6 and 7, respectively when using the three preferred variable selection methods (i.e., Lasso, Lasso Order, and GMUS). The directions of the marginal associations are also indicated (+: positive, −: negative). The list of the metabolites that were selected to be associated with BC, and CRC with $\geqslant$ 10% of the replications using all methods can be found in Web Appendix C.

Comparing the three variable selection methods, Lasso Order gives the most selections, followed by Lasso and GMUS. Across different methods, there are metabolites that are mutually selected by the different methods, for example, TAG 48:5(FA 18:3) and DAG 14:1/18:1; on the other hand, each method also selects some unique metabolites. Comparing the two imputation options, they produce relatively consistent results. Since we only select variables with high replications ($\geqslant$ 50%, with FDR controlled for every replication), we include selections from all the proposed variable selection methods as identified signals. Sensitivity analysis using single cancer-type cases and their own matched controls is performed. More details on the analysis and the selected metabolites are presented in Web Appendix B. The variables selected are largely the same, although fewer variables are selected due to reduced sample sizes.

Table 5: Metabolites that are robustly (selected among ⩾50% of replications) associated with BC risks and the direction of their marginal association to the BC risks.

Method	Platform	Half Min Imputation	Multiple Imputation
Lasso	GC-MS	Alpha-ketoglutarate (76%)(−)	
Lasso	Lipidyzer (composition)		TAG 48:5(FA18:3) (80%)(−) DAG 14:1/18:1 (78%)(+)
Lasso	Lipidyzer (concentration)	DAG 14:1/18:1 (97%)(+)	DAG 14:1/18:1 (97%)(+) TAG 48:5(FA18:3) (64%)(−)
Lasso Order	NMR	N-methylnicotinic acid (59%)(+)	N-methylnicotinic acid (63%)(+)
Lasso Order	GC-MS	Alpha-ketoglutarate (54%)(−)	
Lasso Order	LC-MS (AbsQuant)	3HBA (67%)(+) Cystine (66%)(−)	
Lasso Order	Lipidyzer (composition)	TAG 47:0(FA15:0) (68%)(−)	TAG 48:5(FA18:3) (80%)(−) DAG 14:1/18:1 (76%)(+)
GMUS	Lipidyzer (composition)	TAG 52:2(FA18:2) (66%)(−)	PC 16:0/18:2 (53%)(+)
GMUS	Lipidyzer (concentration)	DAG 14:1/18:1 (63%)(+)	DAG 14:1/18:1 (93%)(+)
Corrected Lasso	GC-MS	Alpha-ketoglutarate (55%)(−)	Alpha-ketoglutarate (73%)(−)

5 Discussion

In conclusion, our extensive empirical studies show that appropriately handling missing data and measurement errors using the knockoff approach can control FDR at the targeted rate and gain power in terms of finding metabolites associated with BC and CRC risks. When the general simultaneous knockoff methods (Dai and Zheng, 2023) are used for two outcomes (see Web Appendix A.3 for details), we find that appropriately handling missing data with multiple imputation and measurement error using GMUS will control FDR for multiple outcomes at the targeted rate.

We identified a group of metabolites that are associated with either BC, CRC, or both cancers. The biomarker findings are largely consistent with the existing literature. For example, pentanedioic acid derivatives have been proposed as a potential agent for the treatment of BC (Zhang et al., 2022). N-methyl nicotinic acid level in LC-ESI-MS has been reported to be positively associated with BC (Valko-Rokytovská et al., 2021). The increase of 3-hydroxybutyric acid (3HBA) level has been found as an indication of the increased fatty acid oxidation, a hallmark for cancer aggressiveness (Cappelletti et al., 2017). For CRC, serum 2,3-dihydroxybutanoic acid has been reported as a biomarker (Loktionov, 2020). Glucose (Vulcan et al., 2017), glycerate (Ni et al., 2014), adenosine (Hata et al., 2023), N-methyl nicotinic acid, cystine (Miller et al., 2013), malate (Neitzel et al., 2020), histidine (Rothwell et al., 2023) and CER (16:0) (Machala et al., 2019) have also been discovered to be associated with CRC in other independent studies. Choline has been reported to be positively associated with risks for both BC (Bae et al., 2014) and CRC (Xu et al., 2008). The results confirm some findings of previous literature and also discover a few new potential metabolite biomarkers for future validation. The matching method (see details

Table 6: Metabolites that are robustly (selected among ⩾50% of replications) associated with CRC risks and the direction of their marginal association to the CRC risks.

Method	Platform	Half Min Imputation	Multiple Imputation
Lasso	GC-MS		2,3-Dihydroxybutanoic acid (93%)(+)
Lasso	LC-MS (AbsQuant)	Glucose (84%)(+) Cystine (52%)(−)	Glucose (62%)(+)
Lasso	LC-MS (RelQuant)	Glycerate (69%)(+) Adenosine (66%)(−)	Adenosine (69%)(−) Glycerate (67%)(+)
Lasso	Lipidyzer (composition)	TAG 48:5(FA18:2) (59%)(+)	
Lasso Order	NMR	N-methylnicotinic acid (76%)(−)	N-methylnicotinic acid (54%)(−)
Lasso Order	GC-MS		2,3-Dihydroxybutanoic acid (63%)(+)
Lasso Order	LC-MS (AbsQuant)	Cystine (100%)(−) 3HBA (75%)(+)	Cystine (94%)(−)
Lasso Order	LC-MS (RelQuant)	Malate (57%)(+)	
Lasso Order	Lipidyzer (composition)	TAG 48:5(FA18:2) (60%)(+)	
Lasso Order	Lipidyzer (concentration)	TAG 47:2(FA14:0) (61%)(−)	
GMUS	LC-MS (AbsQuant)		Histidine (69%)(−)
GMUS	Lipidyzer (concentration)	CER 16:0 (52%)(+)	CER 16:0 (66%)(+)

Table 7: Metabolites that are robustly (selected among ⩾50% of replications) associated with both BC and CRC risks and the direction of their marginal association to these two cancer risks.

Method	Platform	Half Min Imputation	Multiple Imputation
Lasso Order	NMR		N-methylnicotinic acid (57%)(B:+)(C:-)
Lasso Order	LC-MS (AbsQuant)	Cystine (89%)(B:-)(C:-) 3HBA (68%)(B:+)(C:+)	Cystine (99%)(B:-)(C:-) 3HBA (83%)(B:+)(C:+) Glutamic acid (78%)(B:+)(C:+)
GMUS	LC-MS (AbsQuant)	Choline (56%)(B:+)(C:+)	Choline (63%)(B:+)(C:+)

in Web Appendix B) to construct the analytical dataset ensures non-overlapping independent samples for the BC and CRC outcomes to satisfy the assumption of the simultaneous knockoff method Dai and Zheng (2023). One small caveat of the case-control study is the potential X distribution shift. However, this impact is very minor in our analysis given the distribution of X is estimated using the case-control study rather than larger population data.

One limitation of the current study is our cohort only contains post menopausal women. A limitation of the current study is the small sample size for QCs which leads to large variations

in the estimation of the variance-covariance matrix of measurement error. This could make the measurement error correction method vulnerable to potential misspecification of the measurement error distribution and be sensitive to the result of an outlier in 1 or 2 QC pairs. In the current application, we use second-order Model-X knockoff construction. When the variables X and measurement errors both follow the multivariate Gaussian distribution, the second-order condition is sufficient to guarantee the exchangeability of the whole distribution. However, when the variable is non-Gaussian distributed, the higher order moment mismatching could lead to the difference in the distribution of Z and $\widetilde{Z}$ for null variables, which will affect the FDP from the knockoff, especially the simultaneous knockoff procedure. When the measurement error is non-Gaussian, Corrected Lasso will not be suitable, and estimating the optimal error bound for GMUS will be challenging and require a larger sample size for QCs. Further measurement error correction methods for both the estimation of the effect and the variable selection will be worth future research. Another issue is that some metabolites are highly correlated to each other, which will make the knockoff feature very close to the original feature and thus lead to low power. Further method development for group variable selection (by treating metabolites from the same pathway as a group or treating highly correlated metabolites as a group) is worth further exploration but is beyond the scope of this paper. In addition, the current analysis is based on considering each platform's data separately. In the future, methods need to be developed to handle multiple platform data together by solving the challenge of very different measurement scales and potential screening methods to reduce the number of features to allow powerful knockoff construction.

Data Availability Statement

The data that support the findings in this paper is not publicly available but could be requested through WHI in a collaborative mode as described on the Women's Health Initiative website (www.whi.org).

Supplementary Material

We provide an additional pdf file that includes additional simulation results and real data analysis. The R codes for the analysis of this paper are available at https://github.com/RunqiuWang22/Variable_Selection_FDR_noisy.

Acknowledgement

The authors acknowledge the following investigators in the 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 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 Ari-

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

Funding

This research is partly supported by the National Cancer Institute under grants R01 CA119171, CA277133, and P30 CA015704 and by the National Institute of General Medical Sciences under grant U54 GM115458. 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.

References

ACS (2020). *Cancer Facts and Figures 2020*. American Cancer Society, Atlanta, GA.

Antoniadis A, Fryzlewicz P, Letué F, Sapatinas T (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics*, 37(4): 531–552. https://doi.org/10.1111/j.1467-9469.2009.00685.x

Bae S, Ulrich CM, Neuhouser ML, Malysheva O, Bailey LB, Xiao L, et al. (2014). Plasma choline metabolites and colorectal cancer risk in the women's health initiative observational study. *Cancer Research*, 74(24): 7442–7452. https://doi.org/10.1158/0008-5472.CAN-14-1835

Barber RF, Candès EJ (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5): 2055–2085. https://doi.org/10.1214/15-AOS1337

Barber RF, Candès EJ (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5): 2504–2537. https://doi.org/10.1214/18-AOS1765

Barber RF, Candès EJ, Samworth RJ (2020). Robust inference with knockoffs. *The Annals of Statistics*, 48(3): 1409–1431.

Bates S, Candès E, Janson L, Wang W (2021). Metropolized knockoff sampling. *Journal of the American Statistical Association*, 116(535): 1413–1427. https://doi.org/10.1080/01621459.2020.1729163

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological*, 57(1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bogomolov M, Heller R (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association*, 108(504):

1480–1492. https://doi.org/10.1080/01621459.2013.829002

Bogomolov M, Heller R (2018). Assessing replicability of findings across two studies of multiple features. *Biometrika*, 105(3): 505–516. https://doi.org/10.1093/biomet/asy029

Candès E, Fan Y, Janson L, Lv J (2018). Panning for gold: 'Model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 80(3): 551–577. https://doi.org/10.1111/rssb.12265

Cappelletti V, Iorio E, Miodini P, Silvestri M, Dugo M, Daidone MG (2017). Metabolic footprints and molecular subtypes in breast cancer. *Disease Markers*, 2017(1): 7687851.

Chen J, Hou A, Hou TY (2019). A prototype knockoff filter for group selection with FDR control. *Information and Inference*, 9(2): 271–288. https://doi.org/10.1093/imaiai/iaz012

Cheung PK, Ma MH, Tse HF, Yeung KY, Tsang HC, Chu MK, et al. (2019). The applications of metabolomics in the molecular diagnostics of cancer. *Expert Review of Molecular Diagnostics*, 19(9): 785–793. https://doi.org/10.1080/14737159.2019.1656530

Chi Z (2008). False discovery rate control with multivariate p-values. *Electronic Journal of Statistics*, 2: 368–411.

Dai R, Barber R (2016). The knockoff filter for FDR control in group-sparse and multitask regression. In: *Proceedings of The 33rd International Conference on Machine Learning* (MF Balcan, KQ Weinberger, eds.), volume 48 of *Proceedings of Machine Learning Research*, 1851–1859. PMLR, New York, New York, USA.

Dai R, Zheng C (2023). False discovery rate-controlled multiple testing for union null hypotheses: A knockoff-based approach. *Biometrics*, 79(4): 3497–3509. https://doi.org/10.1111/biom.13848

Datta A, Zou H (2017). Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45: 2400–2426. https://doi.org/10.1214/16-AOS1527

Garcia RI, Ibrahim JG, Zhu H (2010). Variable selection in the Cox regression model with covariates missing at random. *Biometrics*, 66(1): 97–104. https://doi.org/10.1111/j.1541-0420.2009.01274.x

Hata N, Shigeyasu K, Umeda Y, Yano S, Takeda S, Yoshida K, et al. (2023). ADAR1 is a promising risk stratification biomarker of remnant liver recurrence after hepatic metastasectomy for colorectal cancer. *Scientific Reports*, 13(1): 2078. https://doi.org/10.1038/s41598-023-29397-z

Heller R, Bogomolov M, Benjamini Y (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences*, 111(46): 16262–16267. https://doi.org/10.1073/pnas.1314814111

Heller R, Yekutieli D (2014). Replicability analysis for genome-wide association studies. *Annals of Applied Statistics*, 8(1): 481–498. https://doi.org/10.1214/13-AOAS697

His M, Viallon V, Dossus L, Gicquiau A, Achaintre D, Scalbert A, et al. (2019). Prospective analysis of circulating metabolites and breast cancer in epic. *BMC Medicine*, 17(1): 178. https://doi.org/10.1186/s12916-019-1408-4

Huang D, Janson L (2020). Relaxing the assumptions of knockoffs by conditioning. *The Annals of Statistics*, 48(5): 3021–3042.

Johnson BA (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 70(2): 351–370. https://doi.org/10.1111/j.1467-9868.2008.00639.x

Kampman E, Thompson R, Wiseman M, Mitrou G, Allen K (2018). PO-087 the WCRF/AICR third expert report on diet, nutrition, physical activity and cancer: Updated recommendations. *ESMO Open*, 3: A260. https://doi.org/10.1136/esmoopen-2018-EACR25.615

Li S, Sesia M, Romano Y, Candès E, Sabatti C (2021). Searching for robust associations with a multi-environment knockoff filter. *Biometrika*, 109(3): 611–629. https://doi.org/10.1093/biomet/asab055

Little RJ, Rubin DB (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Liu Y, Zheng C (2019). Deep latent variable models for generating knockoffs. *Stat*, 8(1): e260. https://doi.org/10.1002/sta4.260

Loh PL, Wainwright MJ (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3): 1637–1664. https://doi.org/10.1214/12-AOS1018

Loktionov A (2020). Biomarkers for detecting colorectal cancer non-invasively: DNA, RNA or proteins? *World Journal of Gastrointestinal Oncology*, 12(2): 124. https://doi.org/10.4251/wjgo.v12.i2.124

Machala M, Procházková J, Hofmanová J, Králiková L, Slavík J, Tylichová Z, et al. (2019). Colon cancer and perturbations of the sphingolipid metabolism. *International Journal of Molecular Sciences*, 20(23): 6051. https://doi.org/10.3390/ijms20236051

Miller JW, Beresford SA, Neuhouser ML, Cheng TYD, Song X, Brown EC, et al. (2013). Homocysteine, cysteine, and risk of incident colorectal cancer in the women's health initiative observational cohort. *The American Journal of Clinical Nutrition*, 97(4): 827–834. https://doi.org/10.3945/ajcn.112.049932

Nannini G, Meoni G, Amedei A, Tenori L (2020). Metabolomics profile in gastrointestinal cancers: Update and future perspectives. *World Journal of Gastroenterology*, 26(20): 2514–2532. https://doi.org/10.3748/wjg.v26.i20.2514

Neitzel C, Demuth P, Wittmann S, Fahrer J (2020). Targeting altered energy metabolism in colorectal cancer: Oncogenic reprogramming, the central role of the tca cycle and therapeutic opportunities. *Cancers*, 12(7): 1731. https://doi.org/10.3390/cancers12071731

Ni Y, Xie G, Jia W (2014). Metabonomics of human colorectal cancer: New approaches for early diagnosis and biomarker discovery. *Journal of Proteome Research*, 13(9): 3857–3870. https://doi.org/10.1021/pr500443c

Playdon MC, Ziegler RG, Sampson JN, Stolzenberg-Solomon R, Thompson HJ, Irwin ML, et al. (2017). Nutritional metabolomics and breast cancer risk in a prospective study. *The American Journal of Clinical Nutrition*, 106(2): 637–649. https://doi.org/10.3945/ajcn.116.150912

Putri SP, Nakayama Y, Matsuda F, et al. (2013). Current metabolomics: Practical applications. *Journal of Bioscience and Bioengineering*, 115(6): 579–589. https://doi.org/10.1016/j.jbiosc.2012.12.007

Rässler S, Rubin DB, Zell ER (2013). Imputation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5: 20. https://doi.org/10.1002/wics.1240

Romano Y, Sesia M, Candès E (2020). Deep knockoffs. *Journal of the American Statistical Association*, 115(532): 1861–1872. https://doi.org/10.1080/01621459.2019.1660174

Rosenbaum M, Tsybakov AB (2013). Improved matrix uncertainty selector. In: *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, 276–290. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Rothwell JA, Bešević J, Dimou N, Breeur M, Murphy N, Jenab M, et al. (2023). Circulating amino acid levels and colorectal cancer risk in the European prospective investigation into cancer and nutrition and UK biobank cohorts. *BMC Medicine*, 21(1): 80. https://doi.org/10.1186/s12916-023-02739-4

Sorensen O, Frigessi A, Thoresen M, Glad IK (2015). Measurement error in lasso: Impact and likelihood bias correction. *Statistica Sinica*, 25(2): 809–829.

Spector A, Janson L (2022). Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, 50(1): 252–276. https://doi.org/10.1214/21-AOS2104

Tsiatis AA (2006). *Semiparametric Theory and Missing Data*. Springer.

Valko-Rokytovská M, Očenáš P, Salayová A, Kostecká Z (2021). Breast cancer: Targeting of steroid hormones in cancerogenesis and diagnostics. *International Journal of Molecular Sciences*, 22(11): 5878. https://doi.org/10.3390/ijms22115878

Vulcan A, Manjer J, Ohlsson B (2017). High blood glucose levels are associated with higher risk of colon cancer in men: A cohort study. *BMC Cancer*, 17(1): 1–8. https://doi.org/10.1186/s12885-016-3022-6

Wolfson J (2011). EEBoost: A general method for prediction and variable selection based on estimating equations. *Journal of the American Statistical Association*, 106(493): 296–305. https://doi.org/10.1198/jasa.2011.tm10098

Xiao Y, Xia J, Li L, et al. (2019). Associations between dietary patterns and the risk of breast cancer: A systematic review and meta-analysis of observational studies. *Breast Cancer Research*, 21(1): 16. https://doi.org/10.1186/s13058-019-1096-1

Xu X, Gammon MD, Zeisel SH, Lee YL, Wetmur JG, Teitelbaum SL, et al. (2008). Choline metabolism and risk of breast cancer in a population-based study. *The FASEB Journal*, 22(6): 2045. https://doi.org/10.1096/fj.07-101279

Yang L, Wang Y, Cai H, Wang S, Shen Y, Ke C (2020). Application of metabolomics in the diagnosis of breast cancer: A systematic review. *Journal of Cancer*, 11(9): 2540–2551. https://doi.org/10.7150/jca.37604

Yusof AS, Isa ZM, Shah SA (2012). Dietary patterns and risk of colorectal cancer: A systematic review of cohort studies (2000–2011). *Asian Pacific Journal of Cancer Prevention*, 13(9): 4713–4717. https://doi.org/10.7314/APJCP.2012.13.9.4713

Zhang C, Quinones A, Le A (2022). Metabolic reservoir cycles in cancer. In: *Seminars in Cancer Biology*, volume 86, 180–188. Elsevier.

Zhao SD, Nguyen YT (2020). Nonparametric false discovery rate control for identifying simultaneous signals. *Electronic Journal of Statistics*, 14(1): 110–142. https://doi.org/10.1214/20-EJS1726

Zhu J, Djukovic D, Deng L, Gu H, Himmati F, Chiorean EG, et al. (2014). Colorectal cancer detection using targeted serum metabolic profiling. *Journal of Proteome Research*, 13(9): 4120–4130. https://doi.org/10.1021/pr500494u