HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2023 November 15.

Published in final edited form as:

J Am Stat Assoc. 2022 ; 117(540): 2105–2119. doi:10.1080/01621459.2021.1904958.

Individual Data Protected Integrative Regression Analysis of High-Dimensional Heterogeneous Data

Tianxi Cai[a], **Molei Liu**[a], **Yin Xia**[b]

[a]Department of Biostatistics, Harvard School of Public Health, Harvard University, Boston, USA

[b]Department of Statistics, School of Management, Fudan University, Shanghai, China

Abstract

Evidence-based decision making often relies on meta-analyzing multiple studies, which enables more precise estimation and investigation of generalizability. Integrative analysis of multiple heterogeneous studies is, however, highly challenging in the ultra high-dimensional setting. The challenge is even more pronounced when the individual-level data cannot be shared across studies, known as DataSHIELD contraint. Under sparse regression models that are assumed to be similar yet not identical across studies, we propose in this paper a novel integrative estimation procedure for data-Shielding High-dimensional Integrative Regression (SHIR). SHIR protects individual data through summary-statistics-based integrating procedure, accommodates between-study heterogeneity in both the covariate distribution and model parameters, and attains consistent variable selection. Theoretically, SHIR is statistically more efficient than the existing distributed approaches that integrate debiased LASSO estimators from the local sites. Furthermore, the estimation error incurred by aggregating derived data is negligible compared to the statistical minimax rate and SHIR is shown to be asymptotically equivalent in estimation to the ideal estimator obtained by sharing all data. The finite-sample performance of our method is studied and compared with existing approaches via extensive simulation settings. We further illustrate the utility of SHIR to derive phenotyping algorithms for coronary artery disease using electronic health records data from multiple chronic disease cohorts.

Keywords

DataSHIELD; Distributed learning; High dimensionality; Model heterogeneity; Rate optimality; Sparsistency

CONTACT Yin Xia xiayin@fudan.edu.cn Department of Statistics, School of Management, Fudan University.
Authors are listed in alphabetical order.

Supplementary Material

In the Supplement, we provide some justifications for Conditions 1 and 6, present detailed proofs of Theorems 1–3, outline theoretical analyses of SHIR for various penalty functions, and present additional simulation results.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

1. Introduction

1.1. Background

Synthesizing information from multiple studies is crucial for evidence-based medicine and policy decision making. Meta-analyzing multiple studies allows for more precise estimates and enables investigation of generalizability. In the presence of heterogeneity across studies and high-dimensional predictors, such integrative analysis however is highly challenging. An example of such integrative analysis is to develop generalizable predictive models using electronic health records (EHR) data from different hospitals. In addition to high-dimensional features, EHR data analysis encounters privacy constraints in that individual patient data (IPD) typically cannot be shared across local hospital sites, which makes the challenge of integrative analysis even more pronounced. Breach of Privacy arising from data sharing is in fact a growing concern in general for scientific studies. Recently, Wolfson et al. (2010) proposed a generic individual-information protected integrative analysis framework, named DataSHIELD, that transfers only summary statistics[1] from each distributed local site to the central site for pooled analysis. Conceptually highly valued by research communities (see, e.g., Jones et al. 2012; Doiron et al. 2013), the DataSHIELD facilitates multi-study integrative analysis when IPD pooled meta-analysis is not feasible due to ethical and/or legal restrictions (Gaye et al. 2014). In the low-dimensional setting, a number of statistical methods have been developed for distributed analysis that satisfy the DataSHIELD constraint (see, e.g., Chen et al. 2006; Wu et al. 2012; Liu and Ihler 2014; Lu et al. 2015; Huang and Huo 2015; Han and Liu 2016; He et al. 2016; Zöller, Lenz, and Binder 2018; Duan et al. 2019, 2020). Distributed high-dimensional regression have largely focused on settings without between-study heterogeneity as detailed in Section 1.2. To the best of our knowledge, no existing distributed learning methods can effectively handle both high-dimensionality and the presence of model heterogeneity across the local sites.

1.2. Related Work

In the context of high-dimensional regression, several recently proposed distributed inference approaches can be potentially used for integrative analysis under the DataSHIELD constraint. Specifically, Tang, Zhou, and Song (2016), Lee et al. (2017), and Battey et al. (2018) proposed distributed inference procedures aggregating the local debiased LASSO estimators (Zhang and Zhang 2014; Van de Geer et al. 2014; Javanmard and Montanari 2014). By including debiasing procedure in their pipelines, the corresponding estimators can be used for inference directly. Lee et al. (2017) and Battey et al. (2018) proposed to further truncate the aggregated dense debiased estimators to achieve sparsity; see also Maity, Sun, and Banerjee (2019). Though this debiasing-based strategy can be extended to fit for our heterogeneous modeling assumption, it still loses statistical efficiency due to the failure to account for the heterogeneity of the information matrices across different sites. In addition, the use of debiasing procedure at local sites incurs additional error for estimation, as detailed in Section 4.4.

[1]Commonly used summary statistics include the locally fitted regression coefficient and itsHessian matrix in the low-dimensional parametric regression models (see, e.g., Duan et al. 2019, 2020).

Lu et al. (2015) and Li et al. (2016) proposed distributed approaches for ℓ_2-regularized logistic and Cox regression. However, their methods requires equential communications between local sites and the central machine, which may be time and resource consuming. Chen and Xie (2014) proposed to estimate high dimensional parameters by first adopting majority voting to select a positive set and then combining local estimation of the coefficients belonging to this set. Wang, Peng, and Dunson (2014) proposed to aggregate the local estimators through their median values rather than their mean, shown to be more robust to poor estimation performance of local sites with insufficient sample size (Minsker 2019). More recently, Wang et al. (2017) and Jordan, Lee, and Yang (2019) presented a communication-efficient surrogate likelihood framework for distributed statistical learning that only transfers the first-order summary statistics, that is, gradient between the local sites and the central site. Fan, Guo, and Wang (2019) extended their idea and proposed two iterative distributed optimization algorithms for the general penalized likelihood problems. However, their framework, as well as others summarized in this paragraph, is restricted to homogeneous scenarios and cannot be easily extended to the settings with heterogeneous models or covariates.

1.3. Our Contributions

In this article, we fill the methodological gap of high-dimensional distributed learning methods that can accommodate cross-study heterogeneity by proposing a novel data-**S**hielding **H**igh-dimensional **I**ntegrative **R**egression (SHIR) method under the DataSHIELD constraints. While SHIR can be viewed as analogous to the integrative analysis of debiased local LASSO estimators, it achieves debiasing *without* having to perform debiasing for the local estimators. SHIR solves LASSO problem only once in each local site *without* requiring the inverse Hessian matrices or the locally debiased estimators and only needs one turn in communication. Statistically, it serves as the tool for the integrative model estimation and variable selection, in the presence of high dimensionality and heterogeneity in model parameters across sites. In addition, under the ultra-high dimensional regime where p can grow exponentially with the total sample size N, we demonstrate that SHIR can achieve the same error rates asymptotically as the ideal estimator based on the IPD pooled analysis, denoted by IPDpool, and attain consistent variable selection. Such properties are not readily available in the existing literature and some novel technical tools are developed for the theoretical verification. We also show theoretically that SHIR is statistically more efficient than the approach based on integrating and thresholding locally debiased estimators (see, e.g., Lee et al. 2017; Battey et al. 2018). Results from our numerical studies confirm that SHIR performs similarly to the ideal IPDpool estimator outperforms the other methods.

1.4. Outline of the Paper

The rest of this article is organized as follows. We introduce the settings in Section 2 and describe SHIR, our proposed approach in Section 3. Theoretical properties of the SHIR estimator are studied in Section 4. We derive the upper bound for its prediction and estimation risks, compare it with the existing approach, and show that the errors incurred by aggregating derived data is negligible compared to the statistical minimax rate. When the true model is ultra-sparse, SHIR is shown to be asymptotically equivalent to the IPDpool estimator and achieves sparsistency. Section 5 compares the performance

of SHIR to existing methods through simulations. We apply SHIR to derive classification models for coronary artery disease (CAD) using EHR data from four different disease cohorts in Section 6. Section 7 concludes the paper with a discussion. Technical proofs of the theoretical results and additional numerical results are provided in the supplementary material.

2. Problem Statement

Throughout, for any integer d, $[d] = \{1, \ldots, d\}$. For any vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d$ and index set $\mathcal{S} = \{j_1, \ldots, j_k : j_1 < \cdots < j_k\} \subseteq [d]$, $\mathbf{x}_\mathcal{S} = (x_{j_1}, \ldots, x_{j_k})^\top$, $\mathbf{x}_{-1} = (x_2, \ldots, x_d)^\top$, $\|\mathbf{x}\|_q$ denotes the ℓ_q norm of $\mathbf{x}$ and $\|\mathbf{x}\|_\infty = \max_{j \in [d]} |x_j|$. Suppose there are M independent studies and n_m subjects in the mth study, for $m = 1, \ldots, M$. For the ith subject in the mth study, let $Y_i^{(\mathrm{m})}$ and $\mathbf{X}_i^{(\mathrm{m})}$, respectively, denote the response and the p-dimensional covariate vector, $\mathbf{D}_i^{(\mathrm{m})} = \left(Y_i^{(\mathrm{m})}, \mathbf{X}_i^{(\mathrm{m})\top}\right)^\top$, $\mathbf{Y}^{(\mathrm{m})} = \left(Y_1^{(\mathrm{m})}, \ldots, Y_{n_m}^{(\mathrm{m})}\right)^\top$, and $\mathbb{X}^{(\mathrm{m})} = \left(\mathbf{X}_1^{(\mathrm{m})}, \mathbf{X}_2^{(\mathrm{m})}, \ldots, \mathbf{X}_{n_m}^{(\mathrm{m})}\right)^\top$. We assume that the observations in study m, $\mathcal{D}^{(\mathrm{m})} = \left\{\mathbf{D}_i^{(\mathrm{m})}, i = 1, \ldots, n_m\right\}$, are independent and identically distributed. Without loss of generality, assume that $\mathbf{X}_i^{(\mathrm{m})}$ includes 1 as the first component and $\mathbf{X}_{i,-1}^{(\mathrm{m})}$ has mean $\mathbf{0}$. Define the population parameters of interests as

$$\boldsymbol{\beta}_0^{(\mathrm{m})} = \underset{\boldsymbol{\beta}^{(\mathrm{m})}}{\operatorname{argmin}}\, \mathscr{L}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right), \text{ where}$$

$$\mathscr{L}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right) = \mathrm{E}\left\{f\left(\boldsymbol{\beta}^{(\mathrm{m})\top}\mathbf{X}_i^{(\mathrm{m})}, Y_i^{(\mathrm{m})}\right)\right\},$$

$$\boldsymbol{\beta}^{(\mathrm{m})} = \left(\beta_1^{(\mathrm{m})}, \beta_2^{(\mathrm{m})}, \ldots, \beta_p^{(\mathrm{m})}\right)^\top$$

for some specified loss function f. Let $\boldsymbol{\beta}_j = \left(\beta_j^{(1)}, \ldots, \beta_j^{(\mathrm{M})}\right)^\top$, $\boldsymbol{\beta}^{(\bullet)} = \left(\boldsymbol{\beta}^{(1)\top}, \ldots, \boldsymbol{\beta}^{(\mathrm{M})\top}\right)^\top$, and $\boldsymbol{\beta}_{0j}$, $\boldsymbol{\beta}_0^{(\bullet)}$ denote the true values of $\boldsymbol{\beta}_j$, $\boldsymbol{\beta}^{(\bullet)}$. We consider the ultra-high dimensional setting, where the covariate dimension p could grow in an exponential rate of the sample size $N = \sum_{m=1}^{M} n_m$.

For each j, we follow the typical meta-analysis to decompose $\beta_j^{(\mathrm{m})}$ as $\beta_j^{(\mathrm{m})} = \mu_j + \alpha_j^{(\mathrm{m})}$ with $\boldsymbol{\alpha}_j = \left(\alpha_j^{(1)}, \ldots, \alpha_j^{(\mathrm{M})}\right)^\top$ and we set $\mathbf{1}_{M\times 1}^\top \boldsymbol{\alpha}_j = 0$ for identifiability. Here, μ_j represents average effect of the covariate X_j and $\boldsymbol{\alpha}_j$ captures the between-study heterogeneity of the effects. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$, $\boldsymbol{\alpha}^{(\bullet)} = \left(\boldsymbol{\alpha}^{(1)\top}, \ldots, \boldsymbol{\alpha}^{(\mathrm{M})\top}\right)^\top$, $\boldsymbol{\alpha}_{-1}^{(\bullet)} = \left(\boldsymbol{\alpha}_{-1}^{(1)\top}, \ldots, \boldsymbol{\alpha}_{-1}^{(\mathrm{M})\top}\right)^\top$, and $\boldsymbol{\mu}_0$ and $\boldsymbol{\alpha}_0^{(\bullet)}$ be the true values of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}^{(\bullet)}$, respectively. Consider the empirical global loss function

$$\widehat{\mathscr{L}}\left(\boldsymbol{\beta}^{(\bullet)}\right) = N^{-1}\sum_{m=1}^{M} n_m \widehat{\mathscr{L}}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right), \text{ where}$$

$$\widehat{\mathscr{L}}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right) = n_m^{-1}\sum_{i=1}^{n_m} f\left(\boldsymbol{\beta}^{(\mathrm{m})\top}\mathbf{X}_i^{(\mathrm{m})}, Y_i^{(\mathrm{m})}\right), m = 1, \ldots, M$$

Minimizing $\widehat{\mathscr{L}}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ is obviously equivalent to estimating $\boldsymbol{\beta}^{(\mathrm{m})}$ using $\mathscr{D}^{(\mathrm{m})}$ only. To improve the estimation of $\boldsymbol{\beta}_0^{(\bullet)}$ by synthesizing information from $\mathscr{D}^{(\bullet)}$ and overcome the high dimensionality, we employ penalized loss functions, $\widehat{\mathscr{L}}\left(\boldsymbol{\beta}^{(\bullet)}\right) + \lambda\rho\left(\boldsymbol{\beta}^{(\bullet)}\right)$, with the penalty function $\rho(\cdot)$ designed to leverage prior structure information on $\boldsymbol{\beta}_0^{(\bullet)}$. Under the prior assumption that $\boldsymbol{\mu}_0$ is sparse and $\boldsymbol{\alpha}_{0,-1}^{(1)}, \ldots, \boldsymbol{\alpha}_{0,-1}^{(M)}$ are sparse and share the same support, we impose a mixture of LASSO and group LASSO penalty: $\rho\left(\boldsymbol{\beta}^{(\bullet)}\right) = \sum_{j=2}^{p}|\mu_j| + \lambda_g\sum_{j=2}^{p}\|\boldsymbol{\alpha}_j\|_2$, where $\lambda_g \geq 0$ is a tuning parameter. Similar penalty has been used in Cheng, Lu, and Liu (2015). Our construction differs slightly from that of Cheng, Lu, and Liu (2015) where $\|\boldsymbol{\alpha}_{j,-1}\|_2$ was used instead of $\|\boldsymbol{\alpha}_j\|_2$. This modified penalty leads to two main advantages: (i) the estimator is invariant to the permutation of the indices of the M studies; and (ii) it yields better theoretical estimation error bounds for the heterogeneous effects. Then an idealized IPDpool estimator for $\boldsymbol{\beta}_0^{(\bullet)}$ can be obtained as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathrm{IPDpool}}^{(\bullet)} &= \underset{\boldsymbol{\beta}^{(\bullet)}}{\operatorname{argmin}}\,\widehat{\mathscr{Q}}\left(\boldsymbol{\beta}^{(\bullet)}\right), \text{ where}\\ \widehat{\mathscr{Q}}\left(\boldsymbol{\beta}^{(\bullet)}\right) &= \widehat{\mathscr{L}}\left(\boldsymbol{\beta}^{(\bullet)}\right) + \lambda\rho\left(\boldsymbol{\beta}^{(\bullet)}\right),\end{aligned} \tag{1}$$

for some tuning parameter $\lambda \geq 0$. However, the IPDpool estimator is not feasible under the DataSHIELD constraint. Our goal is to construct an alternative estimator that attains the same efficiency as $\widehat{\boldsymbol{\beta}}_{\mathrm{IPDpool}}^{(\bullet)}$ asymptotically but only requires sharing summary data. When p is small, the sparse meta analysis (SMA) approach by He et al. (2016) achieves this goal via estimating $\boldsymbol{\beta}^{(\bullet)}$ as $\widehat{\boldsymbol{\beta}}_{\mathrm{SMA}}^{(\bullet)} = \operatorname{argmin}_{\boldsymbol{\beta}^{(\bullet)}}\widehat{Q}_{\mathrm{SMA}}\left(\boldsymbol{\beta}^{(\bullet)}\right)$, where $\widehat{Q}_{\mathrm{SMA}}\left(\boldsymbol{\beta}^{(\bullet)}\right) = N^{-1}\sum_{m=1}^{M}\left(\boldsymbol{\beta}^{(\mathrm{m})} - \check{\boldsymbol{\beta}}^{(\mathrm{m})}\right)^{\top}\check{\mathbb{V}}_m^{-1}\left(\boldsymbol{\beta}^{(\mathrm{m})} - \check{\boldsymbol{\beta}}^{(\mathrm{m})}\right) + \lambda\rho\left(\boldsymbol{\beta}^{(\bullet)}\right)$, $\check{\boldsymbol{\beta}}^{(\mathrm{m})} = \operatorname{argmin}_{\boldsymbol{\beta}^{(\mathrm{m})}}\widehat{\mathscr{L}}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right)$ and $\check{\mathbb{V}}_m = \left\{n_m^{-1}\nabla^2\widehat{\mathscr{L}}_m\left(\check{\boldsymbol{\beta}}^{(\mathrm{m})}\right)\right\}^{-1}$. The SMA method is DataSHIELD since only derived statistics $\check{\boldsymbol{\beta}}^{(\mathrm{m})}$ and $\check{\mathbb{V}}_m$ are shared in the integrative regression. The SMA estimator attains oracle property when p is relatively small but fails for large p due to the failure of $\check{\boldsymbol{\beta}}^{(\mathrm{m})}$.

3. Data-SHIR

3.1. SHIR Method

In the high-dimensional setting, one may overcome the limitation of the SMA approach by replacing $\check{\boldsymbol{\beta}}^{(\mathrm{m})}$ with the regularized LASSO estimator,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}^{(\mathrm{m})} = \underset{\boldsymbol{\beta}^{(\mathrm{m})}}{\operatorname{argmin}}\widehat{\mathscr{L}}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right) + \lambda_m\left\|\boldsymbol{\beta}_{-1}^{(\mathrm{m})}\right\|_1 \tag{2}$$

However, aggregating $\left\{\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}^{(\mathrm{m})}, m \in [M]\right\}$ is problematic with large p due to their inherent biases. To overcome the bias issue, we build the SHIR method motivated by SMA and the debiasing approach for LASSO (see, e.g., Van de Geer et al. 2014) yet achieve debiasing *without* having to perform debiasing for M local estimators. Specifically, we propose the SHIR estimator for $\boldsymbol{\beta}_0^{(\bullet)}$ as $\widehat{\boldsymbol{\beta}}_{\mathrm{SHIR}}^{(\bullet)} = \mathrm{argmin}_{\boldsymbol{\beta}^{(\bullet)}} \widehat{Q}_{\mathrm{SHIR}}\left(\boldsymbol{\beta}^{(\bullet)}\right)$, where

$$\widehat{Q}_{\mathrm{SHIR}}\left(\boldsymbol{\beta}^{(\bullet)}\right) = N^{-1} \sum_{m=1}^{M} n_m \left\{\boldsymbol{\beta}^{(\mathrm{m})\top} \widehat{\mathbb{H}}_m \boldsymbol{\beta}^{(\mathrm{m})} - 2\boldsymbol{\beta}^{(\mathrm{m})\top} \widehat{\mathbf{g}}_m\right\} + \lambda\rho\left(\boldsymbol{\beta}^{(\bullet)}\right), \tag{3}$$

$\widehat{\mathbb{H}}_m = \nabla^2 \widehat{\mathscr{L}}_m\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}^{(\mathrm{m})}\right)$ is an estimate of the Hessian matrix and $\widehat{\mathbf{g}}_m = \widehat{\mathbb{H}}_m \widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}^{(\mathrm{m})} - \nabla \widehat{\mathscr{L}}_m\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}^{(\mathrm{m})}\right)$. Our SHIR estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{SHIR}}^{(\bullet)}$ satisfy the DataSHIELD constraint as $\widehat{Q}_{\mathrm{SHIR}}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ depends on $\mathscr{D}^{(\mathrm{m})}$ only through summary statistics $\widehat{\mathscr{D}}_m = \left\{n_m, \widehat{\mathbb{H}}_m, \widehat{\mathbf{g}}_m\right\}$, which can be obtained within the m^{th} study, and requires only one round of data transfer from local sites to the central node.

With $\left\{\widehat{\mathbb{H}}_m, \widehat{\mathbf{g}}_m, m = 1, \ldots, M\right\}$, we may implement the SHIR procedure using coordinate descent algorithms (Friedman, Hastie, and Tibshirani 2010) along with reparameterization. Let

$$\widehat{Q}_{\mathrm{SHIR}}\left(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}\right) = \widehat{\mathscr{L}}_{\mathrm{SHIR}}\left(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}\right) + \lambda\rho\left(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}; \lambda_g\right),$$

where $\rho\left(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}; \lambda_g\right) = \|\boldsymbol{\mu}_{-1}\|_1 + \lambda_g \|\boldsymbol{\alpha}_{-1}^{(\bullet)}\|_{2,1}$, $\|\boldsymbol{\alpha}_{-1}^{(\bullet)}\|_{2,1} = \sum_{j=2}^{p} \|\boldsymbol{\alpha}_j\|_2$ and

$$\widehat{\mathscr{L}}_{\mathrm{SHIR}}\left(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}\right) = N^{-1} \sum_{m=1}^{M} n_m \left\{\left(\boldsymbol{\mu}^{\top} + \boldsymbol{\alpha}^{(\mathrm{m})\mathrm{T}}\right) \widehat{\mathbb{H}}_m \left(\boldsymbol{\mu} + \boldsymbol{\alpha}^{(\mathrm{m})}\right) - 2\widehat{\mathbf{g}}_m^{\top}\left(\boldsymbol{\mu} + \boldsymbol{\alpha}^{(\mathrm{m})}\right)\right\}.$$

Then the optimization problem in Equation (3) can be reparameterized and represented as:

$$\left(\widehat{\boldsymbol{\mu}}_{\mathrm{SHIR}}, \widehat{\boldsymbol{\alpha}}_{\mathrm{SHIR}}^{(\bullet)}\right) = \mathrm{argmin}_{(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)})} \widehat{Q}_{\mathrm{SHIR}}\left(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}\right), \quad \text{s.t. } \mathbf{1}_{M\times 1}^{\top} \boldsymbol{\alpha}_j = 0, j \in [p],$$

and $\widehat{\boldsymbol{\beta}}_{\mathrm{SHIR}}$ is obtained with the transformation: $\beta_j^{(\mathrm{m})} = \mu_j + \alpha_j^{(\mathrm{m})}$ for every $j \in [p]$. The above procedure is presented in Algorithm A1 in Section A.5 of the supplementary material.

Remark 1.—The first term in $\widehat{Q}_{\mathrm{SHIR}}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ is essentially the second-order Taylor expansion of $\widehat{\mathscr{L}}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ at the local LASSO estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{LASSO}}^{(\bullet)}$. The SHIR method can also be viewed as approximately aggregating local debiased LASSO estimators without actually carrying out the standard debiasing process. To see this, let $\widehat{Q}_{\mathrm{dLASSO}}\left(\boldsymbol{\beta}^{(\bullet)}\right) = N^{-1} \sum_{m=1}^{M} n_m \left(\boldsymbol{\beta}^{(\mathrm{m})} - \widehat{\boldsymbol{\beta}}_{\mathrm{dLASSO}}^{(\mathrm{m})}\right)^{\top} \widehat{\mathbb{H}}_m \left(\boldsymbol{\beta}^{(\mathrm{m})} - \widehat{\boldsymbol{\beta}}_{\mathrm{dLASSO}}^{(\mathrm{m})}\right) + \lambda\rho\left(\boldsymbol{\beta}^{(\bullet)}\right)$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{dLASSO}}^{(\mathrm{m})}$ is the debiased LASSO estimator for the m^{th} study with

$$\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{dLASSO}} = \widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}} - \widehat{\boldsymbol{\Theta}}_m \nabla \widehat{\mathscr{L}}_m\left(\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}}\right), \quad \text{for } m = 1, \ldots, M, \tag{4}$$

and $\widehat{\boldsymbol{\Theta}}_m$ is a regularized inverse of $\widehat{\mathbb{H}}_m$. We may write

$$\begin{aligned}
&\widehat{Q}_{\mathrm{dLASSO}}\left(\boldsymbol{\beta}^{(\bullet)}\right) \\
&= N^{-1} \sum_{m=1}^{M} \left\{ n_m \left[\boldsymbol{\beta}^{(\mathrm{m})\top} \widehat{\mathbb{H}}_m \boldsymbol{\beta}^{(\mathrm{m})} - 2\boldsymbol{\beta}^{(\mathrm{m})\top} \widehat{\mathbb{H}}_m \widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{dLASSO}} \right] \right. \\
&\quad \left. + C_m \right\} + \lambda \rho\left(\boldsymbol{\beta}^{(\bullet)}\right) \\
&\approx N^{-1} \sum_{m=1}^{M} \left\{ n_m \left[\boldsymbol{\beta}^{(\mathrm{m})\top} \widehat{\mathbb{H}}_m \boldsymbol{\beta}^{(\mathrm{m})} - 2\boldsymbol{\beta}^{(\mathrm{m})\top} \widehat{\mathbf{g}}_m \right] + C_m \right\} \\
&\quad + \lambda \rho\left(\boldsymbol{\beta}^{(\bullet)}\right) \\
&= \widehat{Q}_{\mathrm{SHIR}}\left(\boldsymbol{\beta}^{(\bullet)}\right) + N^{-1} \sum_{m=1}^{M} C_m
\end{aligned}$$

where we use $\widehat{\boldsymbol{\Theta}}_m \widehat{\mathbb{H}}_m \approx \mathbb{I}$ in the above approximation and the term

$$C_m = n_m \left\{ \widehat{\mathbb{H}}_m \widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}} - \widehat{\mathbb{H}}_m \widehat{\boldsymbol{\Theta}}_m \nabla \widehat{\mathscr{L}}_m \left(\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}} \right) \right\}^{\top} \left\{ \widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}} - \widehat{\boldsymbol{\Theta}}_m \nabla \widehat{\mathscr{L}}_m \left(\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}} \right) \right\}$$

does not depend on $\boldsymbol{\beta}^{(\bullet)}$. We only use $\widehat{\boldsymbol{\Theta}}_m \widehat{\mathbb{H}}_m \approx \mathbb{I}$ heuristically above to show a connection between our SHIR estimator and the debiased LASSO, but the validity and asymptotic properties of the SHIR estimator do not require obtaining any $\widehat{\boldsymbol{\Theta}}_m$ or establishing a theoretical guarantee for $\widehat{\boldsymbol{\Theta}}_m \widehat{\mathbb{H}}_m$ being sufficiently close to $\mathbb{I}$.

Remark 2.—Compared with existing debiasing-based methods (Lee et al. 2017; Battey et al. 2018), the SHIR approach is both computationally and statistically efficient. It does not rely on the debiased statistics (4) and achieves debiasing without calculating $\widehat{\boldsymbol{\Theta}}_m$, which can only be estimated well under strong conditions (Van de Geer et al. 2014; Janková and Van De Geer 2016).

3.2. Tuning Parameter Selection

The implementation of SHIR requires selection of three sets of tuning parameters, $\{\lambda_m, m \in [M]\}$, λ and λ_g. We select $\{\lambda_m, m \in [M]\}$ for the LASSO problem locally via the standard K-fold cross-validation (CV). Selecting λ and λ_g needs to balance the tradeoff between the model's degrees of freedom, denoted by $\mathrm{DF}(\lambda, \lambda_g)$, and the quadratic loss in $\widehat{Q}_{\mathrm{SHIR}}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ It is not feasible to tune λ and λ_g via the CV since individual-level data are not available in the central site. We propose to select λ and λ_g as the minimizer of the generalized information criterion (GIC) (Wang and Leng 2007; Zhang, Li, and Tsai 2010), defined as

$$\mathrm{GIC}(\lambda, \lambda_g) = \mathrm{Deviance}(\lambda, \lambda_g) + \gamma_N \mathrm{DF}(\lambda, \lambda_g),$$

where γ_N is some prespecified scaling parameter and

$$\text{Deviance}(\lambda, \lambda_g) = N^{-1} \sum_{m=1}^{M} n_m \left\{ \widehat{\boldsymbol{\beta}}_{\text{SHIR}}^{(\text{m})\top}(\lambda, \lambda_g) \widehat{\mathbb{H}}_m \widehat{\boldsymbol{\beta}}_{\text{SHIR}}^{(\text{m})}(\lambda, \lambda_g) - 2\widehat{\mathbf{g}}_m^{\top} \widehat{\boldsymbol{\beta}}_{\text{SHIR}}^{(\text{m})}(\lambda, \lambda_g) \right\}.$$

Following Zhang, Li, and Tsai (2010) and Vaiter et al. (2012), we define $\text{DF}(\lambda, \lambda_g)$ as the trace of

$$\left[\partial^2_{\widehat{\mathcal{S}}_\mu, \widehat{\mathcal{S}}_\alpha} \widehat{Q}_{\text{SHIR}}\left(\widehat{\boldsymbol{\mu}}_{\text{SHIR}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\text{SHIR}}\right)\right]^{-1} \left[\partial^2_{\widehat{\mathcal{S}}_\mu, \widehat{\mathcal{S}}_\alpha} \widehat{\mathscr{L}}_{\text{SHIR}}\left(\widehat{\boldsymbol{\mu}}_{\text{SHIR}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\text{SHIR}}\right)\right],$$

where $\widehat{\mathcal{S}}_\mu = \{j : \widehat{\mu}_{\text{SHIR}, j}(\lambda, \lambda_g) \neq 0\}$, $\widehat{\mathcal{S}}_\alpha = \{j : \|\widehat{\boldsymbol{\alpha}}_{\text{SHIR}, j}(\lambda, \lambda_g)\|_2 \neq 0\}$, the operator $\partial^2_{\widehat{\mathcal{S}}_\mu, \widehat{\mathcal{S}}_\alpha}$ is defined as the second order partial derivative with respect to $\left(\boldsymbol{\mu}^{\top}_{\widehat{\mathcal{S}}_\mu}, \boldsymbol{\alpha}^{(2)\top}_{\widehat{\mathcal{S}}_\alpha}, \ldots \boldsymbol{\alpha}^{(\text{M})\top}_{\widehat{\mathcal{S}}_\alpha}\right)^{\top}$, after plugging $\boldsymbol{\alpha}^{(1)} = -\sum_{m=2}^{M} \boldsymbol{\alpha}^{(\text{m})}$ into $\widehat{Q}_{\text{SHIR}}\left(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}\right)$ or $\widehat{\mathscr{L}}_{\text{SHIR}}\left(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(\bullet)}\right)$.

Remark 3.—As discussed in Kim, Kwon, and Choi (2012), γ_N can be chosen depending on the goal with commonly choices including $\gamma_N = 2/N$ for AIC (Akaike 1974), $\gamma_N = \log N/N$ for BIC (Bhat and Kumar 2010), $\gamma_N = \log\log p \log N/N$ for modified BIC (Wang, Li, and Leng 2009) and $\gamma_N = 2\log p/N$ for RIC (Foster and George 1994). We used the BIC with $\gamma_N = \log N/N$ in our numerical studies.

Remark 4.—For linear models, it has been shown that the proper choice of γ_N guarantees GIC's model selection consistency under various divergence rates of the dimension p (Kim, Kwon, and Choi 2012). For example, for fixed p, GIC is consistent if $N\gamma_N \to \infty$ and $\gamma_N \to 0$. When p diverges in polynomial rate N^{ξ}, then GIC is consistent provided that $\gamma_N = \log N/N$ (BIC) if $0 < \xi < 1/2$; $\gamma_N = \log\log p \log N/N$ (modified BIC) if $0 < \xi < 1$. When p diverges in exponential rate $O\left(\exp\left(\kappa N^{\xi}\right)\right)$ with $0 < \upsilon < \xi$, GIC is consistent as $\gamma_N = N^{\upsilon - 1}$. These results can be naturally extended to more general log-likelihood functions.

4. Theoretical Results

In this section, we present theoretical properties of $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{SHIR}}$ for $\rho\left(\boldsymbol{\beta}^{(\bullet)}\right) = \rho\left(\boldsymbol{\beta}^{(\bullet)}\right)$ but discuss how our theoretical results can be extended to other sparse structures in Section 7. In Sections 4.2 and 4.3, we derive theoretical consistency and equivalence for the prediction and estimation risks of the SHIR, under high dimensional sparse model and smooth loss function f. In Section 4.4, we compare the risk bounds for SHIR with an estimator derived based on those of the debiasing-based aggregation approaches (Lee et al. 2017; Battey et al. 2018). In addition, Section 4.5 shows that the SHIR achieves sparsistency, that is, variable selection consistency, for the nonzero sets of $\boldsymbol{\mu}_0$ and $\boldsymbol{\alpha}_0^{(\bullet)}$. We begin with some notation and definitions that will be used throughout the article.

4.1. Notation and Definitions

Let $o\{\alpha(n)\}$, $O\{\alpha(n)\}$, $\omega\{\alpha(n)\}$, $\Omega\{\alpha(n)\}$ and $\Theta\{\boldsymbol{\alpha}(n)\}$ respectively represent the sequences that grow in a smaller, equal/ smaller, larger, equal/larger and equal rate of the sequence $\alpha(n)$. Similarly, let o_{P}, O_{P}, ω_{P}, Ω_{P} and Θ_{P} represent each of the corresponding rates with probability approaching 1 as $n \to \infty$.

For any vector $\boldsymbol{v}_0 \in \mathbb{R}^d$, denote the ℓ_2-ball around $\boldsymbol{v}_0$ with radius $r > 0$ as $\mathscr{B}_r(\boldsymbol{v}_0) = \left\{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v} - \boldsymbol{v}_0\|_2 \le r\right\}$. Following Vershynin (2018), we define the sub-Gaussian norm of a random variable X as $\|X\|_{\psi_2} := \sup_{q \ge 1} q^{-1/2}\left(\mathrm{E}|X|^q\right)^{1/q}$ and for any random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$, its sub-Gaussian norm defined as $\|\boldsymbol{X}\|_{\psi_2} = \sup_{v \in \mathscr{B}_1(\mathbf{0})} \|\boldsymbol{v}^\top \boldsymbol{X}\|_{\psi_2}$. For any symmetric matrix $\mathbb{X}$, let $\Lambda_{\min}(\mathbb{X})$ and $\Lambda_{\max}(\mathbb{X})$ denote its minimum and maximum eigenvalue, respectively. For $a \in \mathbb{R}$, denote by $\mathrm{sign}(a)$ the sign of a, and for event $\mathscr{E}$, denote by $\mathbf{I}(\mathscr{E})$ the indicator for $\mathscr{E}$. Denote by $\mathcal{S}_\mu = \{j : \mu_{0j} \ne 0\}$, $\mathcal{S}_\alpha = \left\{j : \|\boldsymbol{\alpha}_{0j}\|_2 \ne 0\right\}$, $\mathcal{S}_0 = \mathcal{S}_\mu \cup \mathcal{S}_\alpha$, $\mathcal{S}_{\mathrm{full}} = \{\mathcal{S}_\mu, \mathcal{S}_\alpha\}$, $s_\mu = |\mathcal{S}_\mu|$, $s_\alpha = |\mathcal{S}_\alpha|$ and $s_0 = |\mathcal{S}_0|$. Let $f_1'(a, y) = \partial f(a, y)/\partial a$ and $f_1''(a, y) = \partial^2 f(a, y)/\partial a^2$. Also, let $\mathbb{H}\left(\boldsymbol{\beta}^{(\bullet)}\right) = N^{-1}\mathrm{bdiag}\left\{n_1\mathbb{H}_1\left(\boldsymbol{\beta}^{(1)}\right), n_2\mathbb{H}_2\left(\boldsymbol{\beta}^{(2)}\right), \ldots, n_M\mathbb{H}_M\left(\boldsymbol{\beta}^{(\mathrm{M})}\right)\right\}$, $\widehat{\mathbb{H}} = \mathbb{H}\left(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathrm{LASSO}}\right)$, $\overline{\mathbb{H}}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right) = \mathrm{E}\left[\mathbb{H}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right)\right]$, and $\overline{\mathbb{H}}_m = \overline{\mathbb{H}}_m\left(\boldsymbol{\beta}_0^{(\mathrm{m})}\right)$. Finally, we introduce the compatibility condition ($\mathscr{C}_{\mathrm{comp}}$) as below.

Definition 1 (Compatibility Condition $\left(\mathscr{C}_{\mathrm{comp}}\right)$.—The Hessian matrix $\mathbb{H}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ and the index set $\mathcal{S}$ satisfy the Compatibility Condition, if for all $\left(\boldsymbol{\mu}_\Delta^\top, \boldsymbol{\alpha}_\Delta^{(\bullet)\top}\right)^\top = \left(\boldsymbol{\mu}_\Delta^\top, \boldsymbol{\alpha}_\Delta^{(1)\top}, \ldots, \boldsymbol{\alpha}_\Delta^{(\mathrm{M})\top}\right)^\top \in \mathscr{C}(t, \mathcal{S})$ with any constant $t > 0$, there exists a constant $\phi_0\left\{t, \mathcal{S}, \mathbb{H}\left(\boldsymbol{\beta}^{(\bullet)}\right)\right\}$ such that,

$$\left(\|\boldsymbol{\mu}_\Delta\|_1 + \lambda_g\|\boldsymbol{\alpha}_\Delta^{(\bullet)\top}\|_{2,1}\right)^2 \le N^{-1}\sum_{m=1}^{M} n_m\left|\mathcal{S}\right|\left\|\mathbb{H}_m^{1/2}\left(\boldsymbol{\beta}^{(\mathrm{m})}\right)\left(\boldsymbol{\mu}_\Delta + \boldsymbol{\alpha}_\Delta^{(\mathrm{m})}\right)\right\|_2^2 / \phi_0\left\{t, \mathcal{S}, \mathbb{H}\left(\boldsymbol{\beta}^{(\bullet)}\right)\right\},$$

where

$$\mathscr{C}(t, \mathcal{S}) = \left\{\left(\boldsymbol{u}^\top, \boldsymbol{v}^{(\bullet)\top}\right)^\top = \left(\boldsymbol{u}^\top, \boldsymbol{v}^{(1)\top}, \ldots, \boldsymbol{v}^{(\mathrm{M})\top}\right)^\top : \boldsymbol{v}^{(1)} + \cdots + \boldsymbol{v}^{(\mathrm{M})} = \mathbf{0},\ \|\boldsymbol{u}_{\mathcal{S}^c}\|_1 + \lambda_g\|\boldsymbol{v}^{(\bullet)}_{\mathcal{S}^c}\|_{2,1}, \le t\left(\|\boldsymbol{u}_{\mathcal{S}}\|_1 + \lambda_g\|\boldsymbol{v}^{(\bullet)}_{\mathcal{S}}\|_{2,1}\right)\right\}$$

and $\phi_0\left\{t, \mathcal{S}, \mathbb{H}\left(\boldsymbol{\beta}^{(\bullet)}\right)\right\}$ represents the compatibility constant of $\mathbb{H}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ on the set $\mathcal{S}$.

4.2. Prediction and Estimation Consistency

To establish theoretical properties of the SHIR estimators in terms of estimation and prediction risks, we first introduce some sufficient conditions. Throughout the following analysis, we assume that $n_m = \Theta(N/M)$ for $m \in [M]$ and $\lambda_g = \Theta\left(M^{-1/2}\right)$

Condition 1.—There exists an absolute constant $\phi_0 > 0$ such that for all $\delta_1 = \Theta\left\{(s_0 M \log p/N)^{1/2}\right\}$, $\boldsymbol{\beta}^{(\bullet)} = \left(\boldsymbol{\beta}^{(1)\top}, \ldots, \boldsymbol{\beta}^{(\mathrm{M})\top}\right)^\top$ satisfying $\boldsymbol{\beta}^{(\mathrm{m})} \in \mathscr{B}_{\delta_1}\left(\boldsymbol{\beta}_0^{(\mathrm{m})}\right)$, the Hessian

matrices $\mathbb{H}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ and the index set $\mathcal{S}_0$ satisfy $\mathscr{C}_{\text{comp}}$ (Definition 1) with compatibility constant $\phi_0\left\{t, \mathcal{S}_0, \mathbb{H}\left(\boldsymbol{\beta}^{(\bullet)}\right)\right\} \geq \phi_0$.

Condition 2.—For all $m \in [M]$, $X_{ij}^{(\mathrm{m})} f_1^{'}\left(\boldsymbol{\beta}_0^{(\mathrm{m})\mathrm{T}} \mathbf{X}_i^{(\mathrm{m})}, Y_i^{(\mathrm{m})}\right)$ is sub-Gaussian, that is, there exists some positive constant $\kappa = \Theta(1)$ such that $\left\| X_{ij}^{(\mathrm{m})} f_1^{'}\left(\boldsymbol{\beta}_0^{(\mathrm{m})\top} \mathbf{X}_i^{(\mathrm{m})}, Y_i^{(\mathrm{m})}\right)\right\|_{\psi 2} < \kappa$. In addition, there exists $B > 0$ such that $\max_{m \in [M], i \in [n_m]} \left\|\mathbf{X}_i^{(\mathrm{m})}\right\|_\infty \leq B$.

Condition 3.—There exists positive $C_L = \Theta(1)$ such that $\left|f_1^{''}(a, y) - f_1^{''}(b, y)\right| \leq C_L |a - b|$ for all $a, b \in \mathbb{R}$.

Remark 5.—Condition 1 is in a similar spirit as the restricted eigenvalue or restricted strong convexity condition introduced by Negahban et al. (2012). The first part of Condition 2 controls the tail behavior of $X_{ij}^{(\mathrm{m})} f_1^{'}(a, y)$ so that the random error $\nabla \widehat{\mathscr{L}}_m\left(\boldsymbol{\beta}_0^{(\mathrm{m})}\right)$ can be bounded properly and the method could be benefited from the group sparsity of $\boldsymbol{\alpha}^{(\bullet)}$ (Huang and Zhang 2010). This condition can be easily verified for sub-Gaussian design and an extensive class of models, for example, the logistic model. In addition, the condition $\max_{m \in [M], i \in [n_m]} \left\|\mathbf{X}_i^{(\mathrm{m})}\right\|_\infty \leq B$ holds for bounded design with $B = \Theta(1)$ and for subGaussian design with $B = \Theta\left[\{\log(pN)\}^{1/2}\right]$ Condition 3 assumes a smooth function f to guarantee that the empirical Hessian matrix $\nabla^2 \widehat{\mathscr{L}}_m\left(\widehat{\boldsymbol{\beta}}_{\text{LASSO}}^{(\mathrm{m})}\right)$ is close enough to $\nabla^2 \widehat{\mathscr{L}}_m\left(\boldsymbol{\beta}_0^{(\mathrm{m})}\right)$, and the term $\widehat{\mathbf{g}}_m = \left[\widehat{\mathbb{H}}_m \widehat{\boldsymbol{\beta}}_{\text{LASSO}}^{(\mathrm{m})} - \nabla \widehat{\mathscr{L}}_m\left(\widehat{\boldsymbol{\beta}}_{\text{LASSO}}^{(\mathrm{m})}\right)\right]$ is close enough to $\left[\nabla^2 \widehat{\mathscr{L}}_m\left(\boldsymbol{\beta}_0^{(\mathrm{m})}\right) \boldsymbol{\beta}_0^{(\mathrm{m})} - \nabla \widehat{\mathscr{L}}_m\left(\boldsymbol{\beta}_0^{(\mathrm{m})}\right)\right]$.

The following Proposition 1 illustrates that for sub-Gaussian weighted design with regular Hessian matrix, Condition 1 (Compatibility Condition) holds with probability approaching 1. This can be viewed as an extension of the existing results, that is, sub-Gaussian design and linear model with lasso penalty (Rivasplata 2012), to our case with nonlinear model and the mixture penalty. We present the proof of Proposition 1 in Section A.1 of the supplementary material.

Proposition 1.—Assume that $s_0 = o\{N/(M \log p)\}$ and Condition 3 holds. Assume in addition that there exists absolute constants κ_x, $C_x > 0$, such that for all $m \in [M]$, $C_x^{-1} \leq \Lambda_{\min}\left(\overline{\mathbb{H}}_m\right) \leq \Lambda_{\max}\left(\overline{\mathbb{H}}_m\right) \leq C_x$, $\max_{\mathbf{x} \in \mathscr{B}_1(\mathbf{0})} \mathrm{E}\left[\mathbf{x}^\top \mathbf{X}_i^{(\mathrm{m})}\right]^4 \leq C_x$ and for any $\delta_1 = \Theta\left\{(s_0 M \log p / N)^{1/2}\right\}$ and $\boldsymbol{\beta}^{(\mathrm{m})} \in \mathscr{B}_{\delta_1}\left(\boldsymbol{\beta}_0^{(\mathrm{m})}\right)$, $\left\|\mathbf{X}_i^{(\mathrm{m})}\left\{f_1^{'}\left(\boldsymbol{\beta}^{(\mathrm{m})\top} \mathbf{X}_i, Y_i^{(\mathrm{m})}\right)\right\}^{1/2}\right\|_{\psi 2} \leq \kappa_x$. Then we have that, Condition 1 is satisfied with probability approaching 1.

Remark 6.—As an important example in practice, it is not hard to verify that, for logistic models with $f(a, y) = ya - \log(1 + e^a)$, and sub-Gaussian covariates $\mathbf{X}_i^{(\mathrm{m})}$, the key assumption on the weighted design required in Proposition 1, $\left\|\mathbf{X}_i^{(\mathrm{m})}\left\{f_1^{'}\left(\boldsymbol{\beta}^{(\mathrm{m})\mathrm{T}} \mathbf{X}_i, Y_i^{(\mathrm{m})}\right)\right\}^{1/2}\right\|_{\psi 2} \leq \kappa_x$, is satisfied.

We further assume in Condition 4 that the local LASSO estimators achieve the minimax optimal error rates to a logarithmic scale (Raskutti, Wainwright, and Yu 2011; Negahban et al. 2012).

Condition 4.—The local estimators satisfy that $\max_{m\in[M]}\left\|\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}}-\boldsymbol{\beta}^{(\mathrm{m})}_{0}\right\|_1=O_{\mathrm{P}}\left\{s_0(\log p/n_m)^{1/2}\right\}$, and $\max_{m\in[M]}\left\|\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}}-\boldsymbol{\beta}^{(\mathrm{m})}_{0}\right\|_2\asymp\max_{m\in[M]}\left\|\mathbb{X}^{(\mathrm{m})}\left(\widehat{\boldsymbol{\beta}}^{(\mathrm{m})}_{\mathrm{LASSO}}-\boldsymbol{\beta}^{(\mathrm{m})}_{0}\right)\right\|_2=O_{\mathrm{P}}\left\{(s_0\log p/n_m)^{1/2}\right\}$.

Remark 7.—Extensive literatures, such as Van de Geer et al. (2008), Bühlmann and Van De Geer (2011), and Negahban et al. (2012), have established a complete theoretical framework regarding to this property. See, for example, Negahban et al. (2012), in which Condition 4 can be proved for strongly convex loss function f.

Next, we present the risk bounds for the SHIR including the prediction risk $\left\|\widehat{\mathbb{H}}^{1/2}\left(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathrm{SHIR}}-\boldsymbol{\beta}^{(\bullet)}_{0}\right)\right\|_2$ and estimation risk $\|\widehat{\boldsymbol{\mu}}_{\mathrm{SHIR}}-\boldsymbol{\mu}_0\|_1+\lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathrm{SHIR}}-\boldsymbol{\alpha}^{(\bullet)}_{0}\|_{2,1}$.

Theorem 1 (Risk bounds for the SHIR).—Under Conditions 1–4, there exists $\lambda=\Theta\left(\{(\log p+M)/N\}^{1/2}+Bs_0M\log p/N\right)$ and $\lambda_g=\Theta\left(M^{-1/2}\right)$ such that

$$\left\|\widehat{\mathbb{H}}^{1/2}\left(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathrm{SHIR}}-\boldsymbol{\beta}^{(\bullet)}_{0}\right)\right\|_2=O_{\mathrm{P}}\left(\{s_0(\log p+M)/N\}^{1/2}+Bs_0^{3/2}M\log p/N\right);$$

$$\|\widehat{\boldsymbol{\mu}}_{\mathrm{SHIR}}-\boldsymbol{\mu}_0\|_1+\lambda_g\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathrm{SHIR}}-\boldsymbol{\alpha}^{(\bullet)}_{0}\|_{2,1}=O_{\mathrm{P}}\left(s_0\{(\log p+M)/N\}^{1/2}+Bs_0^2M\log p/N\right).$$

Note that, in Theorem 1, the rate of the penalty coefficient for $\sum_{j=2}^{p}\|\boldsymbol{\alpha}_j\|_2$ is $\lambda\lambda_g=\Theta\left[\{(\log p+M)/NM\}^{1/2}+BM^{1/2}s_0\log p/N\right]$. The second term in each of the upper bounds of Theorem 1 is the error incurred by aggregation noise of derived data instead of raw data. These terms are asymptotically negligible under sparsity as $s_0=o\left(\{N(\log p+M)\}^{1/2}/[BM\log p]\right)$. Then $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathrm{SHIR}}$ achieves the same error rate as the ideal estimator $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathrm{IPDpool}}$ obtained by combining raw data as shown in the following section, and is nearly rate optimal.

4.3. Asymptotic Equivalence in Prediction and Estimation

Under specific sparsity assumptions, we show the asymptotic equivalence, with respect to prediction and estimation risks, of the SHIR and the ideal IPDpool estimator $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\mathrm{IPDpool}}$ or alternatively defined as

$$\left(\widehat{\boldsymbol{\mu}}_{\mathrm{IPDpool}},\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathrm{IPDpool}}\right)=\underset{\left(\boldsymbol{\mu},\boldsymbol{\alpha}^{(\bullet)}\right)}{\operatorname{argmin}}\ \widehat{\mathscr{L}}\left(\boldsymbol{\mu},\boldsymbol{\alpha}^{(\bullet)}\right)+\tilde{\lambda}\rho\left(\boldsymbol{\mu},\boldsymbol{\alpha}^{(\bullet)};\lambda_g\right),\quad \text{s.t. } \mathbf{1}^{\mathrm{T}}_{M\times 1}\boldsymbol{\alpha}_j=0,\ j\in[p],$$

where $\tilde{\lambda}$ is a tuning parameter.

Theorem 2.—(*Asymptotic Equivalence*) Under assumptions in Theorem 1 and assume $s_0 = o\left(\{N(\log p + M)\}^{1/2}/[BM\log p]\right)$, there exists $\tilde{\lambda} = \Theta\{(\log p + M)/N\}^{1/2}$ and $\lambda_g = \Theta\left(M^{-1/2}\right)$ such that the IPDpool estimator $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{IPDpool}}$ satisfies

$$\begin{aligned}&\left\|\widehat{\mathbb{H}}^{1/2}\left(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{IPDpool}} - \boldsymbol{\beta}^{(\bullet)}_0\right)\right\|_2 = O_{\text{P}}\left(\{s_0(\log p + M)/N\}^{1/2}\right);\\ &\|\widehat{\boldsymbol{\mu}}_{\text{IPDpool}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\left\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\text{IPDpool}} - \boldsymbol{\alpha}^{(\bullet)}_0\right\|_{2,1}\\ &\quad = O_{\text{P}}\left(s_0\{(\log p + M)/N\}^{1/2}\right).\end{aligned}$$

Furthermore, for some $\lambda_\Delta = o(\tilde{\lambda})$, the IPDpool and the SHIR defined by (3) with $\lambda = \tilde{\lambda} + \lambda_\Delta$ are equivalent in prediction and estimation in the sense that

$$\begin{aligned}&\left\|\widehat{\mathbb{H}}^{1/2}\left(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{SHIR}} - \boldsymbol{\beta}^{(\bullet)}_0\right)\right\|_2 \le \left\|\widehat{\mathbb{H}}^{1/2}\left(\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{IPDpool}}\right.\right.\\ &\quad\left.\left.-\boldsymbol{\beta}^{(\bullet)}_0\right)\right\|_2 + o_{\text{P}}\left(\{s_0(\log p + M)/N\}^{1/2}\right);\\ &\|\widehat{\boldsymbol{\mu}}_{\text{SHIR}} - \boldsymbol{\mu}_0\|_1 + \lambda_g\left\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\text{SHIR}} - \boldsymbol{\alpha}^{(\bullet)}_0\right\|_{2,1} \le \|\widehat{\boldsymbol{\mu}}_{\text{IPDpool}} - \boldsymbol{\mu}_0\|_1\\ &\quad + \lambda_g\left\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\text{IPDpool}} - \boldsymbol{\alpha}^{(\bullet)}_0\right\|_{2,1} + o_{\text{P}}\left(s_0\{(\log p + M)/N\}^{1/2}\right).\end{aligned}$$

Theorem 2 demonstrates the asymptotic equivalence between $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{SHIR}}$ and $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{IPDpool}}$ with respect to estimation and prediction risks, and hence implies the optimality of the SHIR. Specifically, when $s_0 = o\left(\{N(\log p + M)\}^{1/2}/[BM\log p]\right)$, the excess risks of $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{SHIR}}$ compared to $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{IPDpool}}$ are of smaller order than those of IPDpool, that is, the minimax optimal rates (up to a logarithmic scale) for multi-task learning of high-dimensional sparse model (Huang and Zhang 2010; Lounici et al. 2011). Similar equivalence results was given in Theorem 4.8 of Battey et al. (2018) for the truncated debiased LASSO estimator. However, to the best of our knowledge, in the existing literatures, such results have not been established yet for the LASSO-type estimators obtained directly from a sparse regression model. Compared with Battey et al. (2018), our result does not require the Hessian matrix $\widehat{\mathbb{H}}_m$ to have a sparse inverse since we do not actually rely on the debiasing of $\widehat{\boldsymbol{\beta}}^{(\text{m})}_{\text{LASSO}}$. Consequently, the proofs of Theorem 2 are much more involved than those in Battey et al. (2018). The new technical skills are developed and presented in detail in the supplementary material.

4.4. Comparison With the Debiasing-based Strategy

To compare to existing approaches, we next consider an extension of the debiased LASSO-based procedures proposed in Lee et al. (2017) and Battey et al. (2018) to incorporating between study heterogeneity. Specifically, at the mth site, we derive the debiased LASSO estimator $\widehat{\boldsymbol{\beta}}^{(\text{m})}_{\text{dLASSO}}$ as defined in (4) and send it to the central site, where $\widehat{\boldsymbol{\Theta}}_m$ is obtained via nodewise LASSO (Javanmard and Montanari 2014). At the central site, compute $\widehat{\boldsymbol{\mu}}_{\text{dLASSO}} = M^{-1}\sum_{m=1}^{M}\widehat{\boldsymbol{\beta}}^{(\text{m})}_{\text{dLASSO}}$, $\widehat{\boldsymbol{\alpha}}^{(\text{m})}_{\text{dLASSO}} = \widehat{\boldsymbol{\beta}}^{(\text{m})}_{\text{dLASSO}} - \widehat{\boldsymbol{\mu}}_{\text{dLASSO}}$ and $\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\text{dLASSO}} = \left(\widehat{\boldsymbol{\alpha}}^{(1)}_{\text{dLASSO}}, \ldots, \widehat{\boldsymbol{\alpha}}^{(\text{M})}_{\text{dLASSO}}\right)^{\top}$. The final estimator for $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ can be obtained by thresholding $\widehat{\boldsymbol{\mu}}_{\text{dLASSO}}$ and $\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\text{dLASSO}}$ as

$\widehat{\boldsymbol{\mu}}_{\mathrm{L\&B}} = \mathscr{T}_{\mu}(\widehat{\boldsymbol{\mu}}_{\mathrm{dLASSO}}; \tau_1)$ and $\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathrm{L\&B}} = \mathscr{T}_{\alpha}\left(\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathrm{dLASSO}}; \mu_2\right)$, by Lee et al. (2017) and Battey et al. (2018), where

$$\begin{aligned}\mathscr{T}_{\mu}(\boldsymbol{\mu}; \tau_1) &= \left\{\mu_1, \mu_2^{h+}(\tau_1), \ldots, \mu_p^{h+}(\tau_1)\right\}^{\top} \quad \text{or} \\ &\quad \left\{\mu_1, \mu_2^{s+}(\tau_1), \ldots, \mu_p^{s+}(\tau_1)\right\}^{\top} \\ \mathscr{T}_{\alpha}\left(\boldsymbol{\alpha}^{(\bullet)}; \tau_2\right) &= \mathrm{vec}\left\{\left[\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2^{h+}(\tau_2), \ldots \boldsymbol{\alpha}_p^{h+}(\tau_2)\right]^{\top}\right\} \quad \text{or} \\ &\quad \mathrm{vec}\left\{\left[\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2^{s+}(\tau_2), \ldots \boldsymbol{\alpha}_p^{s+}(\tau_2)\right]^{\top}\right\},\end{aligned}$$

for any vector $\mathbf{x} = (x_1, \ldots, x_d)^{\top}$ and constant τ, $\mathbf{x}^{h+} = \mathbf{x}\boldsymbol{I}(\|\mathbf{x}\|_2 > \tau)$ and $\mathbf{x}^{s+} = \mathbf{x}\left(1 - \|\mathbf{x}\|_2^{-1}\tau\right)\boldsymbol{I}(\|\mathbf{x}\|_2 > \tau)$ respectively denote the hard and soft thresholded counterparts of $\mathbf{x}$, and $\mathrm{vec}(\mathbb{A})$ vectorize the matrix $\mathbb{A}$ by column.

The error rates of $\left\{\widehat{\boldsymbol{\mu}}_{\mathrm{L\&B}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathrm{L\&B}}\right\}$ can be derived by extending Lee et al. (2017) and Battey et al. (2018). We outline the results below and provide details in Section A.3.4 of the supplementary material. Denote by $\overline{\mathbb{H}}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right) = \mathrm{E}\left[\mathbb{H}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right)\right]$, $\overline{\mathbb{H}}_m = \overline{\mathbb{H}}_m\left(\boldsymbol{\beta}_0^{(\mathrm{m})}\right)$, $\overline{\boldsymbol{\Theta}}_m = \left\{\bar{\theta}_{mj\ell}\right\}_{p\times p} = \overline{\mathbb{H}}_m^{-1}$ and $s_1 = \max_{m\in[M]j\in[p]}\left|\left\{\ell \neq j : \bar{\theta}_{mj\ell} \neq 0\right\}\right|$. Then in analog to Theorem 1, one can obtain that

$$\left\|\widehat{\boldsymbol{\mu}}^{(\bullet)}_{\mathrm{L\&B}} - \boldsymbol{\mu}_0\right\|_1 + \lambda_g\left\|\widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathrm{L\&B}} - \boldsymbol{\alpha}_0^{(\bullet)}\right\|_{2,1} \tag{5}$$

$$= O_{\mathrm{P}}\left(s_0\{(\log p + M)/N\}^{1/2} + Bs_0(s_0 + s_1)M\log p/N\right), \tag{6}$$

where B is as defined in Condition 2. Compared with the error rates of SHIR as presented in Theorem 1, $\left\{\widehat{\boldsymbol{\mu}}_{\mathrm{L\&B}}, \widehat{\boldsymbol{\alpha}}^{(\bullet)}_{\mathrm{L\&B}}\right\}$ shares the same "first term", $s_0\{(\log p + M)/N\}^{1/2}$, representing the error of an individual-level empirical process. However, its second term incurred by data aggregation can be larger than that of SHIR as $s_1 = \omega(s_0)$, which could happen due to the complex design in practice.

In addition, SHIR could be more efficient than the debiasing-based strategy even when the impact of the additional error term, which depends on s_1 in (6), is asymptotically negligible. Consider the setting when all $\boldsymbol{\beta}^{(\mathrm{m})}$'s are the same, that is, $\boldsymbol{\beta}^{(\mathrm{m})} = \boldsymbol{\beta}$, and p is moderate or small so that the regularization is unnecessary and the maximum likelihood estimator (MLE) for $\boldsymbol{\beta}$ is feasible and asymptotically Gaussian. In this case, SHIR can be viewed as the inverse variance weight estimation with asymptotic variance $\boldsymbol{\Sigma}_{\mathrm{SHIR}} = \left\{\sum_{m=1}^{M} n_m\overline{\boldsymbol{\Theta}}_m^{-1}\right\}^{-1}$, while the debiasing-based approach outputs an estimator of variance $\boldsymbol{\Sigma}_{\mathrm{L\&B}} = M^{-2}\sum_{m=1}^{M} n_m^{-1}\overline{\boldsymbol{\Theta}}_m$. It is not hard to show that $\boldsymbol{\Sigma}_{\mathrm{SHIR}} \preceq \boldsymbol{\Sigma}_{\mathrm{L\&B}}$, where the equality holds only if all $\overline{\boldsymbol{\Theta}}_m$'s are in certain proportion. Thus, SHIR is strictly more efficient than debiasing-based approach under the low-dimensional setting with heterogeneous $\overline{\boldsymbol{\Theta}}_m$, which commonly arises in metaanalysis as the distributions of $\mathbb{X}^{(\mathrm{m})}$'s are typically heterogeneous across the local sites. In the high-dimensional setting, similarly, SHIR is expected to benefit from the "inverse variance weight" construction, and our simulation results in Section 5 support this point.

4.5. Sparsistency

In this section, we present theoretical results concerning the variable selection consistency of the SHIR. We begin with some extra sufficient conditions for the sparsistency result.

Condition 5.—For all $\delta_2 = \Theta\left\{(s_0 M \log p/N)^{1/2}\right\}$ and $\boldsymbol{\beta}^{(\mathrm{m})}$ satisfying $\boldsymbol{\beta}^{(\mathrm{m})} \in \mathscr{B}_{\delta_2}(\boldsymbol{\beta}_0^{(\mathrm{m})})$, there exists $C_{\min} = \Theta(1)$ such that $\Lambda_{\min}\left\{\mathbb{H}_{m,\mathscr{S}_0}\left(\boldsymbol{\beta}^{(\mathrm{m})}\right)\right\} > C_{\min}$, where $\mathbb{H}_{m,\mathscr{S}_0}\left(\boldsymbol{\beta}^{(\mathrm{m})}\right)$ denotes the submatrix of $\mathbb{H}_m\left(\boldsymbol{\beta}^{(\mathrm{m})}\right)$ with its rows and columns corresponding to $\mathscr{S}_0$.

Condition 6.—For all $\delta_3 = \Theta\left\{(s_0 M \log p/N)^{1/2}\right\}$ and $\boldsymbol{\beta}^{(\bullet)} = \left(\boldsymbol{\beta}^{(1)\top}, \ldots, \boldsymbol{\beta}^{(\mathrm{M})\top}\right)^\top$ satisfying $\boldsymbol{\beta}^{(\mathrm{m})} \in \mathscr{B}_{\delta_3}(\boldsymbol{\beta}_0^{(\mathrm{m})})$, the weighted design matrix $\mathbb{W}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ satisfies the Irrepresentable Condition $\mathscr{C}_{\mathrm{Irrep}}$ on $\mathscr{S}_{\mathrm{full}}$ with parameter $\epsilon > 0$, where $\mathbb{W}\left(\boldsymbol{\beta}^{(\bullet)}\right)$ is defined in Section A.2 and $\mathscr{C}_{\mathrm{Irrep}}$ is given in Definition A2 of the supplementary material.

Condition 7.—Let $\upsilon = \min\left\{\min_{j\in\mathscr{S}_\mu}|\mu_{0j}|, M^{-1/2}\min_{j\in\mathscr{S}_\alpha}\|\boldsymbol{\alpha}_{0j}\|_2\right\}$. For the ϵ defined in Condition 6, $\left[\{s_0(\log p + M)/N\}^{1/2} + Bs_0^{3/2}M\log p/N\right]/(\upsilon\epsilon) \to 0$, as $N \to \infty$.

Remark 8.—Conditions 5–7 are sparsistency assumptions similar to those of Zhao and Yu (2006) and Nardi et al. (2008). Condition 5 requires the eigenvalues for the covariance matrix of the weighted design matrix corresponding to $\mathscr{S}_0$ to be bounded away from zero, so that its inverse behaves well. Condition 6 adopts the commonly used Irrepresentable Condition (Zhao and Yu 2006) to our mixture penalty setting. Roughly speaking, it requires that the weighted design corresponding to $\mathscr{S}_{\mathrm{full}}$ cannot be represented well by the weighted design for $\mathscr{S}_{\mathrm{full}}^c$. Compared to Nardi et al. (2008), $\mathscr{C}_{\mathrm{Irrep}}$ is less intuitive but essentially weaker. We justify such condition on several common correlation structures and compare it with Zhao and Yu (2006) in Section A.2 of the supplementary material. Condition 7 assumes that the minimum magnitude of the coefficients is large enough to make the nonzero coefficients recognizable. It requires essentially weaker assumption on the minimum magnitude than local LASSO (Zhao and Yu 2006). This is because we leverage the group structure of $\boldsymbol{\beta}_0^{(\mathrm{m})}$'s to improve the efficiency of variable selection.

Theorem 3.—*(Sparsistency)* Let $\widehat{\mathscr{S}}_\mu = \{j: \hat{\mu}_{\mathrm{SHIR},j} \neq 0\}$ and $\widehat{\mathscr{S}}_\alpha = \{j: \|\hat{\boldsymbol{\alpha}}_{\mathrm{SHIR},j}\|_2 \neq 0\}$. Denote by the event $\mathscr{O}_\mu = \left\{\widehat{\mathscr{S}}_\mu = \mathscr{S}_\mu\right\}$ and $\mathscr{O}_\alpha = \left\{\widehat{\mathscr{S}}_\alpha = \mathscr{S}_\alpha\right\}$. Under Conditions 1–7 and assume that

$$\lambda = o\left(\upsilon/s_0^{1/2}\right) \quad \text{and}$$

$$\lambda = \epsilon^{-1}\omega\left(\{(\log p + M)/N\}^{1/2} + Bs_0 M\log p/N\right),$$

we have $\mathrm{P}(\mathscr{O}_\mu \cap \mathscr{O}_\alpha) \to 1$ as $N \to \infty$.

Theorem 3 establishs the sparsistency of SHIR. When $s_0 = o\left(\{N(\log p + M)\}^{1/2}/[BM\log p]\right)$, Condition 7 turns out to be $v\epsilon = \omega\left(\{s_0(\log p + M)/N\}^{1/2}\right)$, the corresponding sparsistency assumption for the IPDpool estimator. In contrast, a similar condition, which could be as strong as $v\epsilon = \omega\left\{(s_0 M\log p/N)^{1/2}\right\}$, is required for the local LASSO estimator (Zhao and Yu 2006). Compared with the local one, our integrative analysis procedure can recognize smaller signal under some sparsity assumptions. In this sense, the structure of $\boldsymbol{\beta}_0^{(\bullet)}$ helps us to improve the selection efficiency over the local LASSO estimator. Different from the existing work, we need carefully address the mixture penalty ρ and the aggregation noise of the SHIR, which introduce technical difficulties to our theoretical analysis.

In both Theorems 2 and 3, we allow M, the number of studies, to diverge while still preserving theoretical properties. The growing rate of M is allowed to be

$$M = \min\left(o\left\{(N/\log p)^{1/2}/(Bs_0)\right\}, o\left\{N/(Bs_0\log p)^2\right\}\right)$$

for the equivalence result in Theorem 2 and

$$M = \min\left(o\left\{N\epsilon v/\left(Bs_0^{3/2}\log p\right)\right\}, o\left\{N(\epsilon v)^2/s_0\right\}\right)$$

for the sparsistency result in Theorem 3.

5. Simulation Study

We present simulation results in this section to evaluate the performance of our proposed SHIR estimator and compare it with several other approaches. The simulation codes are available at https://github.com/moleibobliu/SHIR. Let $M \in \{4, 8\}$ and $p \in \{100, 800, 1500\}$ and set $n_m = n = 400$ for each m. For each configuration, we summarize results based on 200 simulated datasets. We consider three data-generating mechanisms:

i. *Sparse precision and correctly specified model (strong and sparse signal):* Across all studies, let $\mathcal{S}_\mu = \{1, 2, \ldots, 6\}$ for $\boldsymbol{\mu}$, $\mathcal{S}_\alpha = \{3, 4, \ldots, 8\}$ for $\boldsymbol{\alpha}$, $\mathcal{S} = \mathcal{S}_\mu \cup \mathcal{S}_\alpha$ and $\mathcal{S}^c = [p]\backslash\mathcal{S}$. For each $m \in [M]$, we generate $\mathbf{X}^{(\mathrm{m})}$ from a zero-mean multivariate normal distribution with covariance $\mathbb{C}^{(\mathrm{m})}$, where $\mathbb{C}^{(\mathrm{m})}_{\mathcal{S}^c\mathcal{S}^c} = \mathbb{R}_{p-8}(r_m)$, $\mathbb{C}^{(\mathrm{m})}_{\mathcal{S}^c\mathcal{S}} = \mathbb{R}_{p-8}(r_m)\boldsymbol{\Gamma}_{p-8,8}(r_m, 15)$, $\mathbb{C}^{(m)}_{\mathcal{S}\mathcal{S}} = \mathbb{I}_8 + \boldsymbol{\Gamma}^\top_{p-8,8}(r_m, 15)\mathbb{R}_{p-8}(r_m)\boldsymbol{\Gamma}_{p-8,8}(r_m, 15)$, $\mathbb{I}_q$ denotes the $q \times q$ identity matrix, $\mathbb{R}_q(r)$ denotes the $q \times q$ correlation matrix of AR(1) with correlation coefficient r, $\boldsymbol{\Gamma}_{q_1,q_2}(r, s_1)$ denotes the $q_1 \times q_2$ matrix with each of its column having randomly picked s_1 entries set as r or $-r$ in random and the remaining being 0, and $r_m = 0.4(m-1)/M + 0.15$. Given $\mathbf{X}^{(\mathrm{m})}$, we generate $Y^{(\mathrm{m})}$ from the logistic model $P\left(Y^{(\mathrm{m})} = 1 \mid \mathbf{X}^{(\mathrm{m})}\right) = \mathrm{expit}\left\{\mathbf{X}^{(\mathrm{m})\top}_{\mathcal{S}_\mu}\boldsymbol{\mu}_{\mathcal{S}_\mu} + \mathbf{X}^{(\mathrm{m})\top}_{\mathcal{S}_\alpha}\boldsymbol{\alpha}^{(\mathrm{m})}_{\mathcal{S}_\alpha}\right\}$ with $\boldsymbol{\mu}_{\mathcal{S}_\mu} = 0.5(1, -1, 1, -1, 1, -1)^\top$ and $\boldsymbol{\alpha}^{(m)}_{\mathcal{S}_\alpha} = 0.35(-1)^m \cdot (1, 1, 1, -1, -1, -1)^\top$.

ii. *Sparse precision and correctly specified model (weak and sparse signal):* Use the same data-generation mechanism as in (i) but relatively weak signals $\boldsymbol{\mu}_{\mathcal{S}_\mu} = 0.2(1, -1, 1, -1, 1, -1)^\top$ and $\boldsymbol{\alpha}^{(m)}_{\mathcal{S}_\alpha} = 0.15(-1)^m \cdot (1, 1, 1, -1, -1, -1)^\top$.

iii. *Sparse precision and correctly specified model (strong and dense signal):* Use the same mechanism as in (i) but denser supports: $\mathcal{S}_\mu = \{1, 2, \ldots, 18\}$, and $\mathcal{S}_\alpha = \{7, 8, \ldots, 24\}$, and more heterogeneous coefficients across the sites (see their specific values in Section A.5 of the supplementary material).

iv. *Sparse precision and correctly specified model (weak and dense signal):* Use the same mechanism as in (iii) but weaker signals (see Section A.5 of the supplementary material).

v. *Dense precision and wrongly specified model:* Let $\mathcal{S} = \{1, 2, \ldots, 5\}$, $\mathcal{S}' = \{6, \ldots, 50\}$, and $\mathcal{S}'' = [p] \backslash (\mathcal{S} \cup \mathcal{S}')$. For each $m \in [M]$,, we generate $\mathbf{X}^{(m)}$ from zero-mean multivariate normal with covariance matrix $\mathbb{C}^{(m)}$, where $\mathbb{C}^{(m)}_{(\mathcal{S}' \cup \mathcal{S}'')(\mathcal{S}' \cup \mathcal{S}'')} = \text{bdiag}\{\mathbb{R}_{45}(r_m), \mathbb{R}_{p-50}(r_m)\}$, $\mathbb{C}^{(m)}_{\mathcal{S}\mathcal{S}''} = \mathbf{0}$, $\mathbb{C}^{(m)}_{\mathcal{S}'\mathcal{S}} = \mathbb{R}_{45}(r_m)\boldsymbol{\Gamma}_{45,5}(r_m, 45)$ and $\mathbb{C}^{(m)}_{\mathcal{S}\mathcal{S}} = \mathbb{I}_5 + \boldsymbol{\Gamma}^\top_{45,5}(r_m, 45)\mathbb{R}_{45}(r_m)\boldsymbol{\Gamma}_{45,5}(r_m, 45)$. Given $\mathbf{X}^{(m)}$, we generate $Y^{(m)}$ from a logistic model with $P\left(Y^{(m)} = 1 \mid \mathbf{X}^{(m)}\right) = \text{expit}\left\{\sum_{j=1}^5 \{0.25 + 0.15(-1)^m\}\left\{X_j^{(m)} + 0.2\left(X_j^{(m)}\right)^3\right\}. + 0.1\sum_{j=1}^4 X_j^{(m)} X_{j+1}^{(m)}\right\}$

Across all settings, the distributions of $\mathbf{X}^{(m)}$ and model parameters of $Y^{(m)} \mid \mathbf{X}^{(m)}$ differ across the M sites, which mimic the heterogeneity of the covariates and models. The heterogeneity of $\mathbf{X}^{(m)}$ is driven by the study-specific correlation coefficient r_m in its covariance matrix $\mathbb{C}^{(m)}$. Under Settings (i)–(iv), the fitted logistic loss corresponds to the likelihood under a correctly specified model with the support of $\boldsymbol{\mu}$ and that of $\boldsymbol{\alpha}^{(m)}$ overlapping but not exactly the same. Under Setting (v), the fitted loss corresponds to a misspecified model but the true target parameter $\boldsymbol{\beta}^{(m)}$ remains approximately sparse with only first 5 elements being relatively large, 45 close to zero and remaining exactly zero. For each $j \in \mathcal{S}$, there are 15 nonzero coefficients on average in the jth column of the precision $\boldsymbol{\Theta}_m$ under Settings (i)–(iv), and 45 nonzero coefficients under Setting (v). So we can use Settings (i)–(iv) to simulate the scenario with sparse precision on the active set and use Setting (v) to simulate relatively dense precision.

For each simulated dataset, we obtain the SHIR estimator as well as the following alternative estimators: (a) the IPDpool estimator $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{IPDpool}} = \text{argmin}_{\boldsymbol{\beta}^{(\bullet)}} \widehat{\mathcal{Q}}\left(\boldsymbol{\beta}^{(\bullet)}\right)$; (b) the SMA estimator (He et al. 2016), following the sure independent screening procedure (Fan and Lv 2008) that reduces the dimension to $n/(3\log n)$ as recommended by He et al. (2016); and (c) the debiasing-based estimator $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{L\&B}}$ as introduced in Section 4.4, denoted by $\text{Debias}_{\text{L\&B}}$. For $\widehat{\boldsymbol{\beta}}^{(\bullet)}_{\text{L\&B}}$, we used the soft thresholding to be consistent with the penalty used by IPDpool, SMA and SHIR. We used the BIC to choose the tuning parameters for all methods.

In Figures 1 and 2, we present the relative average absolute estimation error (rAEE), $\left\|\boldsymbol{\beta}^{(\bullet)} - \boldsymbol{\beta}_0^{(\bullet)}\right\|_1$, and the relative prediction error (rPE), $\left\|\mathbb{X}\left(\boldsymbol{\beta}^{(\bullet)} - \boldsymbol{\beta}_0^{(\bullet)}\right)\right\|_2$, for each estimator compared to the IPDpool estimator, respectively. Consistent with the theoretical equivalence results, the SHIR estimator attains very close estimation and prediction accuracy as those of theidealizedIPDpoolestimator,withrPEandrAEEaround1.03 under Setting (i), 1.02 under (ii), 1.04 under (iii), 1.03 under (iv), and 1.07 under (v). The SHIR estimator is substantially more efficient than the SMA under all the settings, with about 45% reduction in both AEE and PE on average. This can be attributed to the improved performance of the local LASSO estimator $\hat{\boldsymbol{\beta}}_{\text{LASSO}}^{(\text{m})}$ over the MLE $\check{\boldsymbol{\beta}}^{(\text{m})}$ on sparse models. The superior performance is more pronounced for large p such as 800 and 1500, because the screening procedure does not work well in choosing the active set, especially in the presence of correlations among the covariates. Compared with $\text{Debias}_{\text{L\&B}}$, SHIR also demonstrates its gain in efficiency. Specifically, relative to SHIR, $\text{Debias}_{\text{L\&B}}$ has 15% ~ 29% higher AEE and 18% − 42% higher PE under the five settings. This is consistent with our theoretical results presented in Section 4.4 that SHIR has smaller error compared to $\text{Debias}_{\text{L\&B}}$ due to the heterogeneous Hessians and aggregation errors. In addition, compared to Settings (i)–(iv), the excessive error of $\text{Debias}_{\text{L\&B}}$ is larger in Setting (v) where the the inverse Hessian $\overline{\boldsymbol{\Theta}}_m$ is relatively dense. This is consistent with conclusion in Section 4.4.

In Figure 3, we present the average number of misclassifications on the support of $\boldsymbol{\beta}^{(\bullet)}$, that is, $\sum_{j=1}^{p} I\left(\hat{\beta}_j = 0\right) \neq I(\beta_{0,j} = 0)$ for $\hat{\boldsymbol{\beta}}$ obtained via different methods under Settings (i)–(iv) where the model for Y is correctly specified. SMA performs poorly and has more misclassification numbers under nearly all the settings, specially for $p = 800, 1500$ and dense signals. Both IPDpool and SHIR have good support recovery performance with the misclassification numbers below 2.5 under all settings with sparse signal, and below 7.5 under those with dense signal. These two methods attain similar misclassification numbers with the absolute differences less than 0.8 across all settings. Compared to IPDpool and SHIR, $\text{Debias}_{\text{L\&B}}$ has significantly worse performance for all the settings with $p \in \{800, 1500\}$. For weak signal, $M = 4$ and $p \in \{800, 1500\}$, the misclassification numbers of $\text{Debias}_{\text{L\&B}}$ are about two to four times as large as those of IPDpool and SHIR. For strong signal or $M = 8$, the gap between $\text{Debias}_{\text{L\&B}}$ and SHIR is still visible though a bit smaller. For example, under Setting (i) with $M = 8$, $\text{Debias}_{\text{L\&B}}$ has about 60% more misclassifications than SHIR when $p = 800$, and 110% more misclassifications when $p = 1500$ on average. In Figures A1 and A2 of the supplementary material, we present the average true positive rate (TPR) and false discovery rate (FDR) for recovering the support of $\boldsymbol{\beta}^{(\bullet)}$. When $p = 100$, the estimator $\text{Debias}_{\text{L\&B}}$ tends to have smaller FDR than those of SHIR. However, this is achieved at the expense of substantially lower TPR. On the other hand, when p is larger ($p \in \{800, 1500\}$), SHIR attains lower FPR than $\text{Debias}_{\text{L\&B}}$ while attaining higher or comparable TPR. In summary, SHIR achieves similar performance as IPDpool and better performance than $\text{Debias}_{\text{L\&B}}$ in support recovery.

6. Application to EHR Phenotyping in Multiple Disease Cohorts

Linking EHR data with biorepositories containing "-omics" information has expanded the opportunities for biomedical research (Kho et al. 2011). With the growing availability of these high-dimensional data, the bottleneck in clinical research has shifted from a paucity of biologic data to a paucity of high-quality phenotypic data. Accurately and efficiently annotating patients with disease characteristics among millions of individuals is a critical step in fulfilling the promise of using EHR data for precision medicine. Novel machine learningbased phenotyping methods leveraging a large number of predictive features have improved the accuracy and efficiency of existing phenotyping methods (Liao et al. 2015; Yu et al. 2015).

While the portability of phenotyping algorithms across multiple patient cohorts is of great interest, existing phenotyping algorithms are often developed and evaluated for a specific patient population. To investigate the portability issue and develop EHR phenotyping algorithms for CAD useful for multiple cohorts, Liao et al. (2015) developed a CAD algorithm using a cohort of rheumatoid arthritis (RA) patients and applied the algorithm to other disease cohorts using EHR data from Partner's Healthcare System. Here, we performed integrative analysis of multiple EHR disease cohorts to jointly develop algorithms for classifying CAD status for four disease cohorts including type 2 diabetes mellitus (DM), inflammatory bowel disease (IBD), multiple sclerosis (MS), and RA. Under the DataSHIELD constraint, our proposed SHIR algorithm enables us to let the data determine if a single CAD phenotyping algorithm can perform well across four disease cohorts or disease specific algorithms are needed.

For algorithm training, clinical investigators have manually curated gold standard labels on the CAD status used as the response Y, for $n_1 = 172$ DM patients, $n_2 = 230$ IBD patients, $n_3 = 105$ MS patients, and $n_4 = 760$ RA patients. There are a total of $p = 533$ candidate features including both codified features, narrative features extracted via natural language processing (NLP) (Zeng et al. 2006), as well as their two-way interactions. Examples of codified features include demographic information, lab results, medication prescriptions, counts of International Classification of Diseases (ICD) codes and Current Procedural Terminology (CPT) codes. Since patients may not have certain lab measurements and missingness is highly informative, we also create missing indicators for the lab measurements as additional features. Examples of NLP terms include mentions of CAD, current smoking (CSMO), nonsmoking (NSMO) and CAD related procedures. Since the count variables such as the total number of CAD ICD codes are zero-inflated and skewed, we take $\log(x + 1)$ transformation and include $\mathbf{I}(x > 0)$ as additional features for each count variable x.

For each cohort, we randomly select 50% of the observations to form the training set for developing the CAD algorithms and use the remaining 50% for validation. We trained CAD algorithms based on SHIR, $\text{Debias}_{\text{L\&B}}$ and SMA. Since the true model parameters are unknown, we evaluate the performance of different methods based on the prediction performance of the trained algorithms on the validation set. We consider several standard accuracy measures including the area under the receiver operating characteristic curve (AUC), the brier score defined as the mean squared residuals on the validation data, as

well as the F-score at threshold value chosen to attain a false-positive rate of 5% ($F_{5\%}$) and 10% ($F_{10\%}$), where the F-score is defined as the harmonic mean of the sensitivity and positive predictive value. The standard errors of the estimated prediction performance measures are obtained by bootstrapping the validation data. We only report results based on tuning parameters selected with BIC as in the simulation studies but note that the results obtained from AIC are largely similar in terms of prediction performance. Furthermore, to verify the improvement of the performance by combining the four datasets, we include the LASSO estimator for each local dataset (Local) as a comparison.

In Table 1, we present the estimated coefficients for variables that received nonzero coefficients by at least one of the included methods. Interestingly, all integrative analysis methods set all heterogeneous coefficients to zero, suggesting that a single CAD algorithm can be used across all cohorts although different intercepts were used for different disease cohorts. The magnitude of the coefficients from SHIR largely agree with the published algorithm with most important features being NLP mentions and ICD codes for CAD as well as total number of ICD codes which serves as a measure of healthcare utilization. The SMA set all variables to zero except for age, nonsmoker and the NLP mentions and ICD codes for CAD, while $\text{Debias}_{\text{L\&B}}$ has more similar support to SHIR.

The point estimates along with their 95% bootstrap confidence intervals of the accuracy measures are presented in Figure 4. The results suggest that SHIR has the best performance across all methods, nearly on all datasets and across all measures. Among the integrative methods, SMA and $\text{Debias}_{\text{L\&B}}$ performed much worse than SHIR on all accuracy measures. For example, the AUC with its 95% confidence interval of the CAD algorithm for the RA cohorts trained via SHIR, SMA and $\text{Debias}_{\text{L\&B}}$ is respectively 0.93 (0.90,0.95), 0.88 (0.84,0.92), and 0.86 (0.82,0.90). Compared to the local estimator, SHIR also performs substantially better. For example, the AUC of SHIR and Local for the IBD cohort is 0.93 (0.88,0.97) and 0.90 (0.84,0.95). The difference between the integrative procedures and the local estimator is more pronounced for the DM cohort with AUC being around 0.95 for SHIR and 0.90 for the local estimator trained using DM data only. The local estimator fails to produce an informative algorithm for the MS cohort due to the small size of the training set. These results again demonstrate the power of borrowing information across studies via integrative analysis.

7. Discussion

In this article, we proposed a novel approach, the SHIR, for integrative analysis of high dimensional data under the DataSHIELD framework, where only summary statistics are allowed to be transferred from the local sites to the central node to protect the individual-level data. As we demonstrated via both theoretical analyses and numerical studies, the SHIR estimator is considerably more efficient than the estimators obtained based on the debiasing-based strategies considered in literatures (Lee et al. 2017; Battey et al. 2018). Also, our method accommodates heterogeneity among the design matrices, as well as the coefficients of the local sites, which is not adequately handled under the ultra high-dimensional regime in existing literature. Our approach only solves the LASSO problem

once in each local site without requiring the computation of $\widehat{\boldsymbol{\Theta}}^{(m)}$ or debiasing. Note that, SHIR aims at an ℓ_1/ℓ_2-consistent estimation and is not asymptotically unbiased. Consequently, it cannot be directly used for hypothesis testing or confidence interval construction, for example, Caner and Kock (2018a,b). Future work lies on developing statistical approaches for such purposes under DataSHIELD, high-dimensionality and heterogeneity. In addition, sparsistency of our estimator relies on the Irrepresentable Condition (Condition 6) that has been commonly used in the literature (see, e.g.,Yuan and Lin 2006; Nardi et al. 2008), but its rigorous verification for random design or nonlinear models is technically highly challenging. To achieve variable selection consistency without such condition, one may use non-concave (group) sparse penalty like group adaptive lasso (Wang and Leng 2008) or group bridge (Zhou and Zhu 2010) in our framework.

For the choice of penalty, in the current article, we focus primarily on the mixture penalty, $\rho\left(\boldsymbol{\beta}^{(\bullet)}\right) = \sum_{j=2}^{p}|\mu_j| + \lambda_g\sum_{j=2}^{p}\|\boldsymbol{\alpha}_j\|_2$. Nevertheless, other penalty functions, such as group lasso (Huang and Zhang 2010) and hierarchical lasso (Zhou and Zhu 2010), can be incorporated into our framework provided that they effectively leverage certain prior knowledge. Similar techniques used for deriving the theoretical results of SHIR with the mixture penalty can be used for other penalty functions, with some technical details varying according to different choices on $\rho(\cdot)$. See Section A.4 of the supplementary material for further justifications.

For the consistency result in Theorem 1, SHIR requires $s_0 = o\left\{(N/M\log p)^{1/2}\right\}$. Although this sparsity assumption is already weaker than those in the existing literature (Battey et al. 2018, e.g.) as shown in Section 4.4, it may be strong in practical applications. For example, $(N/M\log p)^{1/2} \approx 7$ in the EHR example which suggests that the sparsity assumption may not hold. Nevertheless, the resulting SHIR algorithm appear to perform well in terms of out-of-sample classification accuracy. On the other hand, it is of interests to explore the possibilities of relaxing such assumption. One potential way is to use multiple rounds of communications such as Fan, Guo, and Wang (2019). Detailed analysis of this approach warrants future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

The research of Yin Xia was supported in part by NSFC Grants 12022103, 11771094, and 11690013. The research of Tianxi Cai and Molei Liu were partially supported by the Translational Data Science Center for a Learning Health System at Harvard Medical School and Harvard T.H. Chan School of Public Health.

References

Akaike H. (1974), “A New Look at the Statistical Model Identification,” IEEE Transactions on Automatic Control, 19, 716–723. [2108]

Battey H, Fan J, Liu H, Lu J, Zhu Z. (2018), "Distributed Testing and Estimation Under Sparse High Dimensional Models," The Annals of Statistics, 46, 1352–1382. [2105,2106,2108,2110,2117,2118] [PubMed: 30034040]

Bhat HS, and Kumar N. (2010), On the Derivation of the Bayesian Information Criterion, Los Angels: School of Natural Sciences, University of California. [2108]

Bühlmann P, and Van De Geer S. (2011), Statistics for High-dimensional Data: Methods, Theory and Applications. Springer Science & Business Media. [2109]

Caner M, and Kock AB (2018a), "Asymptotically Honest Confidence Regions for High Dimensional Parameters by the Desparsified Conservative Lasso," Journal of Econometrics, 203, 143–168. [2117]

Caner M, and Kock AB (2018b), "High Dimensional Linear GMM," arXiv preprint arXiv:1811.08779. [2117]

Chen X, and Xie M. g. (2014), "A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data," Statistica Sinica, 24, 1655–1684. [2106]

Chen Y, Dong G, Han J, Pei J, Wah BW, and Wang J. (2006), "Regression Cubes With Lossless Compression and Aggregation," IEEE Transactions on Knowledge and Data Engineering, 18, 1585–1599. [2105]

Cheng X, Lu W, and Liu M. (2015), "Identification of Homogeneous and Heterogeneous Variables in Pooled Cohort Studies," Biometrics, 71, 397–403. [2107] [PubMed: 25732747]

Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BH, Perola M, Stolk RP, Foco L, Minelli C, Waldenberger M, Holle R, Kvaløy K, Hillege HL, Tassé A-M, Ferretti V, Fortier I. (2013), "Data Harmonization and Federated Analysis of Population-based Studies: The BioSHaRE Project," Emerging Themes in Epidemiology, 10, 12. [2105] [PubMed: 24257327]

Duan R, Boland MR, Liu Z, Liu Y, Chang HH, Xu H, Chu H, Schmid CH, Forrest CB, Holmes JH, Schuemie MJ, Berlin JA, Moore JH, Chen Y. (2020), "Learning From Electronic Health Records Across Multiple Sites: A Communication-efficient and Privacypreserving Distributed Algorithm," Journal of the American Medical Informatics Association, 27, 376–385. [2105] [PubMed: 31816040]

Duan R, Boland MR, Moore JH, and Chen Y. (2019), "Odal: A One-shot Distributed Algorithm to Perform Logistic Regressions on Electronic Health Records Data From Multiple Clinical Sites," in PSB R. Altman B, Dunker AK, Hunter L, Ritchie MD, Murray T. and Klein TE, Kohala Coast, Hawaii, USA: World Scientific Publishing Conference, pp. 30–41. [2105]

Fan J, Guo Y, and Wang K. (2019), "Communication-efficient Accurate Statistical Estimation," arXiv preprint arXiv:1906.04870. [2106,2118]

Fan J, and Lv J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," Journal of the Royal Statistical Society, Series B, 70, 849–911. [2112]

Foster DP, and George EI (1994), "The Risk Inflation Criterion for Multiple Regression," The Annals of Statistics, 1947–1975. [2108]

Friedman J, Hastie T, and Tibshirani R. (2010), "A Note on the Group Lasso and a Sparse Group Lasso," arXivpreprintarXiv:1001.0736. [2107]

Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J, Boyd AW, Newby CJ, Nuotio M-L, Wilson R, Butters O, Murtagh B, Demir I, Doiron D, Giepmans L, Wallace SE, Budin-Ljøsne I, Oliver Schmidt C, Boffetta P, Boniol M, Bota M, Carter KW, deKlerk N, Dibben C, Francis RW, Hiekkalinna T, Hveem K, Kvaløy K, Millar S, Perry IJ, Peters A, Phillips CM, Popham F, Raab G, Reischl E, Sheehan N, Waldenberger M, Perola M, van den Heuvel E, Macleod J, Knoppers BM, Stolk RP, Fortier I, Harris JR, Woffenbuttel BH, Murtagh MJ, Ferretti V, Burton PR(2014), "DataSHIELD: Taking the Analysis to the Data, not the Data to the Analysis," International Journal of Epidemiology, 43, 1929–1944. [2105] [PubMed: 25261970]

Han J, and Liu Q. (2016), "Bootstrap Model Aggregation for Distributed Statistical Learning," in Advances in Neural Information Processing Systems, Lee D, Sugiyama M, Luxburg U, Guyon I. and Garnett R, San Diego, CA, USA: Curran Associates, Inc., pp. 1795–1803. [2105]

He Q, Zhang HH, Avery CL, and Lin D. (2016), "Sparse Meta-analysis With High-dimensional Data," Biostatistics, 17, 205–220. [2105,2107,2112] [PubMed: 26395907]

Huang C, and Huo X. (2015), "A Distributed One-step Estimator," arXiv preprint arXiv:1511.01443. [2105]

Huang J, and Zhang T. (2010), "The Benefit of Group Sparsity," The Annals of Statistics, 38, 1978–2004. [2109,2110,2118]

Janková J, and Van De Geer S. (2016), "Confidence Regions for High-dimensional Generalized Linear Models Under Sparsity," arXiv preprint arXiv:1610.01353. [2108]

Javanmard A, and Montanari A. (2014), "Confidence Intervals and Hypothesis Testing for High-dimensional Regression," The Journal of Machine Learning Research, 15, 2869–2909. [2105,2110]

Jones E, Sheehan N, Masca N, Wallace S, Murtagh M, and Burton P. (2012), "DataSHIELD–shared Individual-level Analysis Without Sharing the Data: A Biostatistical Perspective," Norsk Epidemiologi, 21.[2105]

Jordan MI, Lee JD, and Yang Y. (2019), "Communication-efficient Distributed Statistical Inference," Journal of the American Statistical Association, 526, 668–681. [2106]

Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Crane PK, Pathak J, Chute CG, Bielinski SJ, Kullo IJ, Li R, Manolio TA, Chisholm RL, Denny JC (2011), "Electronic Medical Records for Genetic Research: Results of the Emerge Consortium," Science Translational Medicine, 3, 79re1–79re1. [2116]

Kim Y, Kwon S, and Choi H. (2012), "Consistent Model Selection Criteria on High Dimensions," Journal of Machine Learning Research, 13, 1037–1057. [2108]

Lee JD, Liu Q, Sun Y, and Taylor JE (2017), "Communication-efficient Sparse Regression," Journal of Machine Learning Research, 18, 1–30. [2105,2106,2108,2110,2117]

Li W, Liu H, Yang P, and Xie W. (2016), "Supporting Regularized Logistic Regression Privately and Efficiently," PloS One, 11, e0156479. [2106]

Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, Goryachev S, Chen P, Savova GK, Agniel D, et al. (2015). "Methods to Develop an Electronic Medical Record Phenotype Algorithm to ComparetheRisk of CoronaryArteryDiseaseAcross3 Chronic Disease Cohorts," PLoS One, 10, e0136651. [2116]

Liu M, Xia Y, Cai T, and Cho K. (2020), "Integrative High Dimensional Multiple Testing With Heterogeneity Under Data Sharing Constraints," arXiv preprint arXiv:2004.00816.

Liu Q, and Ihler AT (2014), "Distributed Estimation, Information Loss and Exponential Families," in Advances in Neural Information Processing Systems, pp. 1098–1106. [2105]

Lounici K, Pontil M, Van De Geer S, Tsybakov AB (2011), "Oracle Inequalities and Optimal Inference Under Group Sparsity," The Annals of Statistics, 39, 2164–2204. [2110]

Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, and Ohno-Machado L. (2015), "Webdisco: A Web Service for Distributed Cox Model Learning Without Patient-level Data Sharing," Journal of the American Medical Informatics Association, 22, 1212–1219. [2105,2106] [PubMed: 26159465]

Maity S, Sun Y, and Banerjee M. (2019), "Communication-efficient Integrative Regression in High-dimensions," arXiv preprint arXiv:1912.11928. [2106]

Minsker S. (2019), "Distributed Statistical Estimation and Rates of Convergence in Normal Approximation," Electronic Journal of Statistics, 13, 5213–5252. [2106]

Nardi Y, Rinaldo A. (2008), "On the Asymptotic Properties of the Group Lasso Estimator for Linear Models," Electronic Journal of Statistics, 2, 605–633. [2111,2117]

Negahban SN, Ravikumar P, Wainwright MJ, Yu B. (2012), "A Unified Framework for High-dimensional Analysis of m-estimators With Decomposable Regularizers," Statistical Science, 27, 538–557. [2109]

Raskutti G, Wainwright MJ, and Yu B. (2011), "Minimax Rates of Estimation for High-dimensional Linear Regression Over l_q-Balls," IEEE Transactions on Information Theory, 57, 6976–6994. [2109]

Rivasplata O. (2012), "Subgaussian Random Variables: An Expository Note," Internet publication, PDF. [2109]

Tang L, Zhou L, and Song PX-K (2016), "Method of Divide-andcombine in Regularized Generalized Linear Models for Big Data," arXiv preprint arXiv:1611.06208. [2105]

Vaiter S, Deledalle C, Peyré G, Fadili J, and Dossal C. (2012), "The Degrees of Freedom of the Group Lasso," arXiv preprint arXiv:1205.1481. [2108]

Van de Geer S, Bühlmann P, Ritov Y, Dezeure R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-dimensional Models," The Annals of Statistics, 42, 1166–1202. [2105,2107,2108]

Van de Geer SA (2008), "High-dimensional Generalized Linear Models and the Lasso," The Annals of Statistics, 36, 614–645. [2109]

Vershynin R. (2018), High-dimensional Probability: An Introduction With Applications in Data Science, Vol. 47. Cambridge, UK: Cambridge University Press. [2108]

Wang H, and Leng C. (2007), "Unified Lasso Estimation by Least Squares Approximation," Journal of the American Statistical Association, 102, 1039–1048. [2108]

Wang H, and Leng C. (2008), "A Note on Adaptive Group Lasso,"Computational Statistics & Data Analysis, 52, 5277–5286. [2118]

Wang H, Li B, and Leng C. (2009), "Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters," Journal of the Royal Statistical Society, Series B, 71, 671–683. [2108]

Wang J, Kolar M, Srebro N, and Zhang T. (2017), "Efficient Distributed Learning With Sparsity," in Proceedings of the 34th International Conference on Machine Learning, Vol. 70, pp. 3636–3645. JMLR. org. [2106]

Wang X, Peng P, and Dunson DB (2014), "Median Selection Subset Aggregation for Parallel Inference," In Advances in Neural Information Processing Systems, pp. 2195–2203. [2106]

Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, LaFlamme P, Tobin MD, Macleod J, Little J, Fortier I, Knoppers BM, Burton PR (2010), "DataSHIELD: Resolving a Conflict in Contemporary Biosciencepeforming a Pooled Analysis of Individual-level Data Without Sharing the Data," International Journal of Epidemiology, 39, 1372–1382. [2105] [PubMed: 20630989]

Wu Y, Jiang X, Kim J, and Ohno-Machado L. (2012), "Grid Binary Logistic Regression (glore): Building Shared Models Without Sharing Data," Journal of the American Medical Informatics Association, 19, 758–764. [2105] [PubMed: 22511014]

Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, and Cai T. (2015), "Toward High-throughput Phenotyping: Unbiased Automated Feature Extraction and Selection From Knowledge Sources," Journal of the American Medical Informatics Association, 22, 993–1000. [2116] [PubMed: 25929596]

Yuan M, and Lin Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," Journal of the Royal Statistical Society, Series B, 68, 49–67. [2117]

Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, and Lazarus R. (2006), "Extracting Principal Diagnosis, Co-morbidity and Smoking Status for Asthma Research: Evaluation of a Natural Language Processing System," BMC Medical Informatics and Decision Making, 6, Article no. 30. [2116]

Zhang C-H, and Zhang SS (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," Journal of the Royal Statistical Society, Series B, 76, 217–242. [2105]

Zhang Y, Li R, and Tsai C-L (2010), "Regularization Parameter Selections Via Generalized Information Criterion," Journal of the American Statistical Association, 105, 312–323. [2108] [PubMed: 20676354]

Zhao P, and Yu B. (2006), "On Model Selection Consistency of Lasso," Journal of Machine Learning Research, 7, 2541–2563. [2111]

Zhou N, and Zhu J. (2010), "Group Variable Selection Via a Hierarchical Lasso and Its Oracle Property," arXiv preprint arXiv:1006.2871. [2118]

Zöller D, Lenz S, and Binder H. (2018), "Distributed Multivariable Modeling for Signature Development Under Data Protection Constraints," arXiv preprint arXiv:1803.00422. [2105]

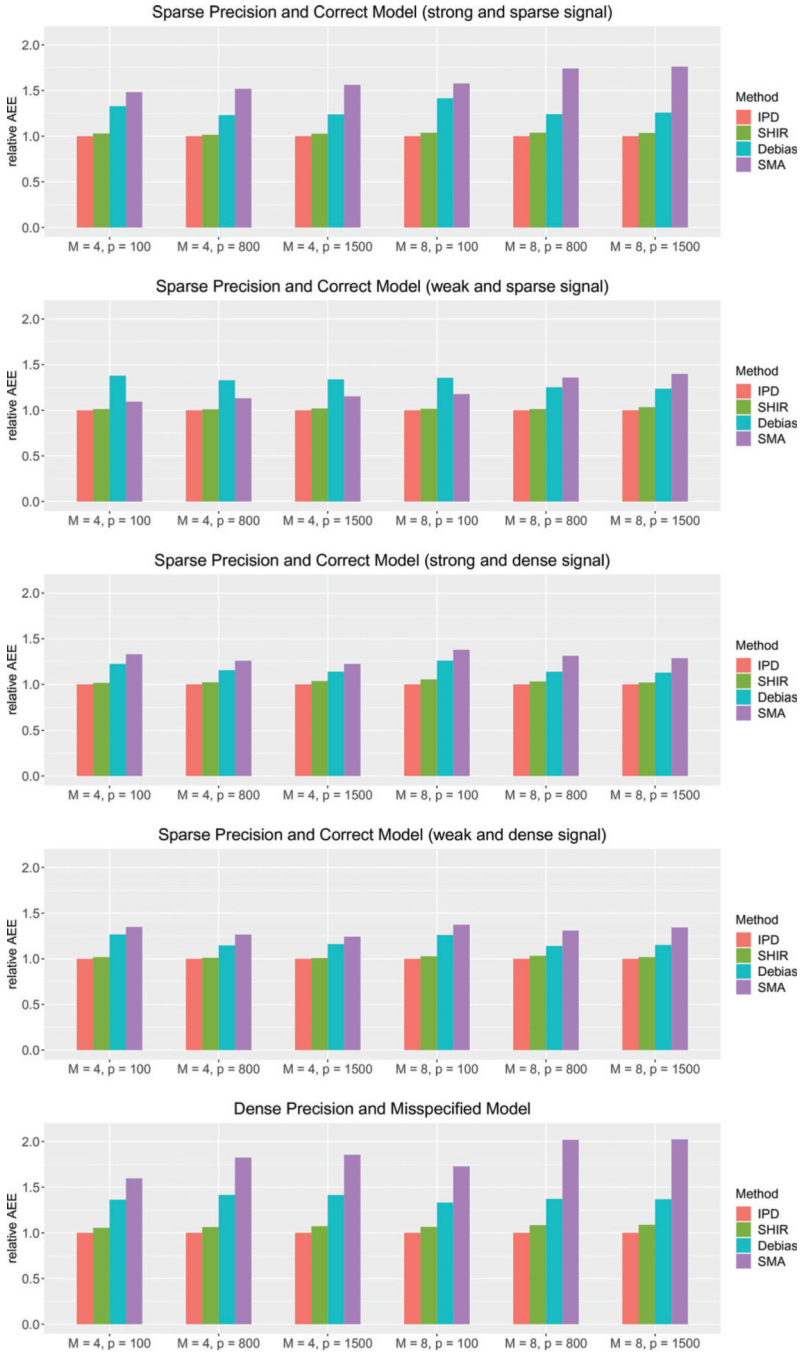

Figure 1.

The relative average absolute estimation error (AEE) of IPDpool (IPD), SHIR, $\text{Debias}_{\text{L\&B}}$ (Debias) and SMA compared to those of IPDpool underdifferent $M \in \{4, 8\}$, $p \in \{100, 800, 1500\}$ and data-generation mechanisms (i)–(v) introduced in Section 5.

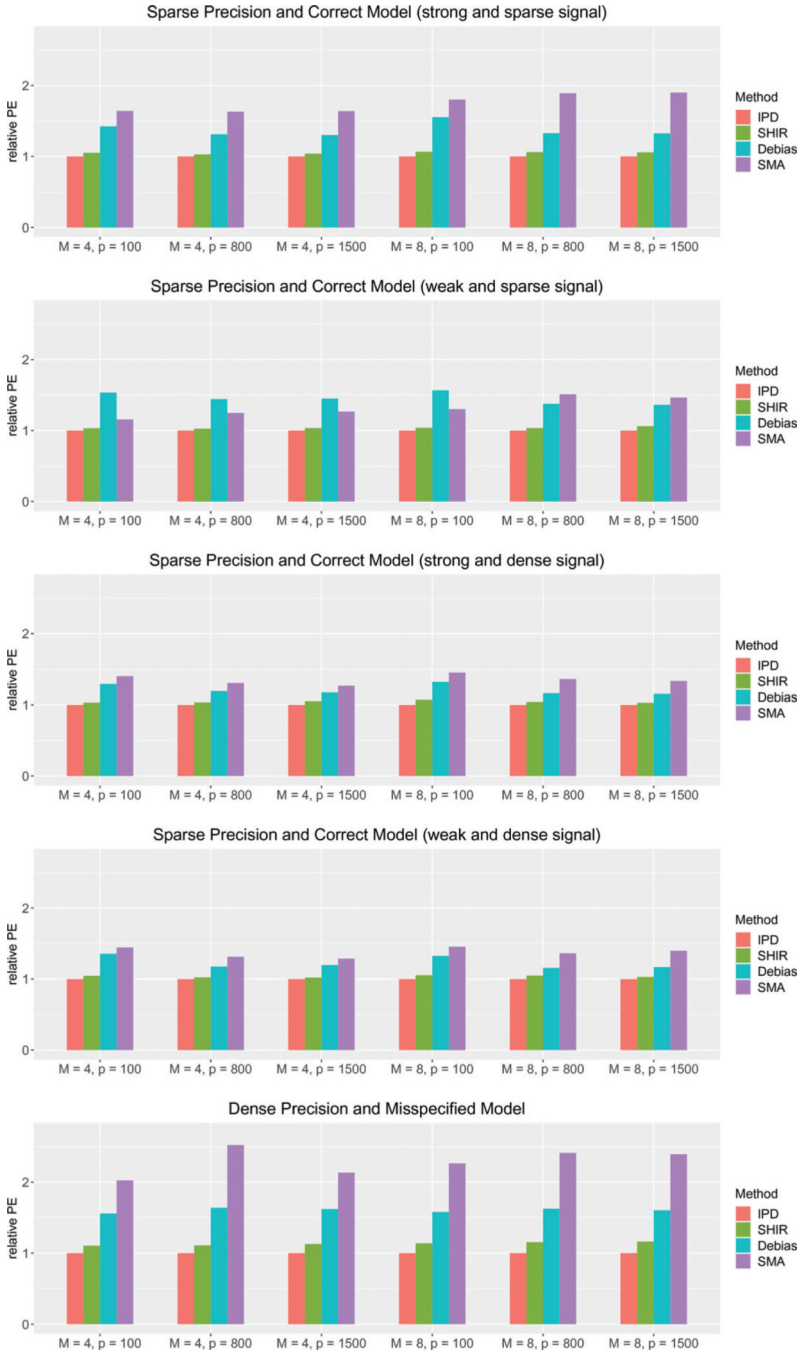

Figure 2.

The relative prediction error (PE) of IPDpool (IPD), SHIR, $\text{Debias}_{\text{L\&B}}$ (Debias), and SMA compared to those of IPDpool under different $M \in \{4, 8\}$, $p \in \{100, 800, 1500\}$ and data-generation mechanisms (i)–(v) introduced in Section 5.

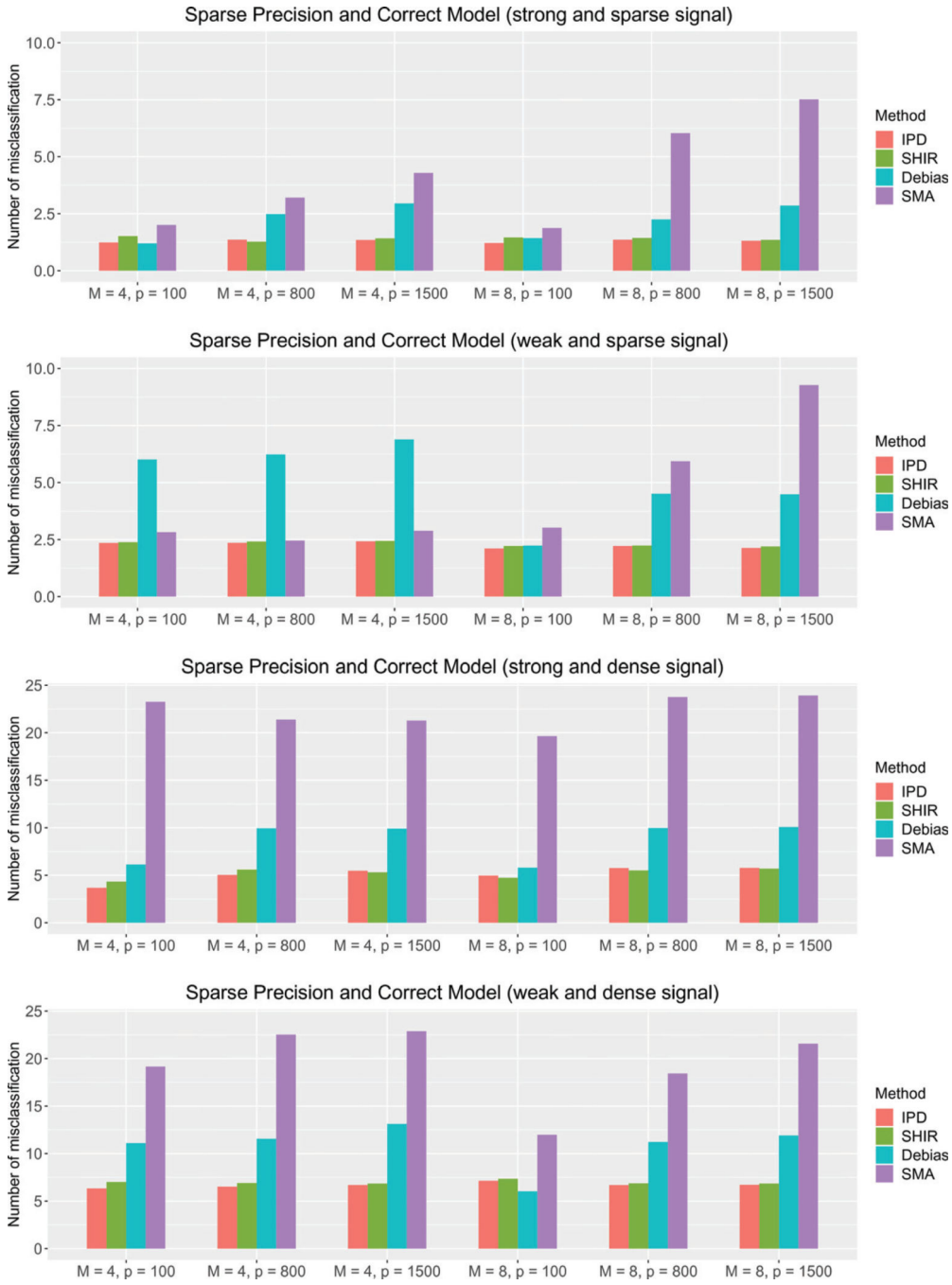

Figure 3.

The average number of misclassifications on $\{I(\beta_j \neq 0), j = 1, \ldots, p\}$ based on IPDpool (IPD), SHIR, $\text{Debias}_{\text{L\&B}}$ (Debias), and SMA under different $M \in \{4, 8\}$, $p \in \{100, 800, 1500\}$ and data-generation mechanisms (i)–(iv) introduced in Section 5.

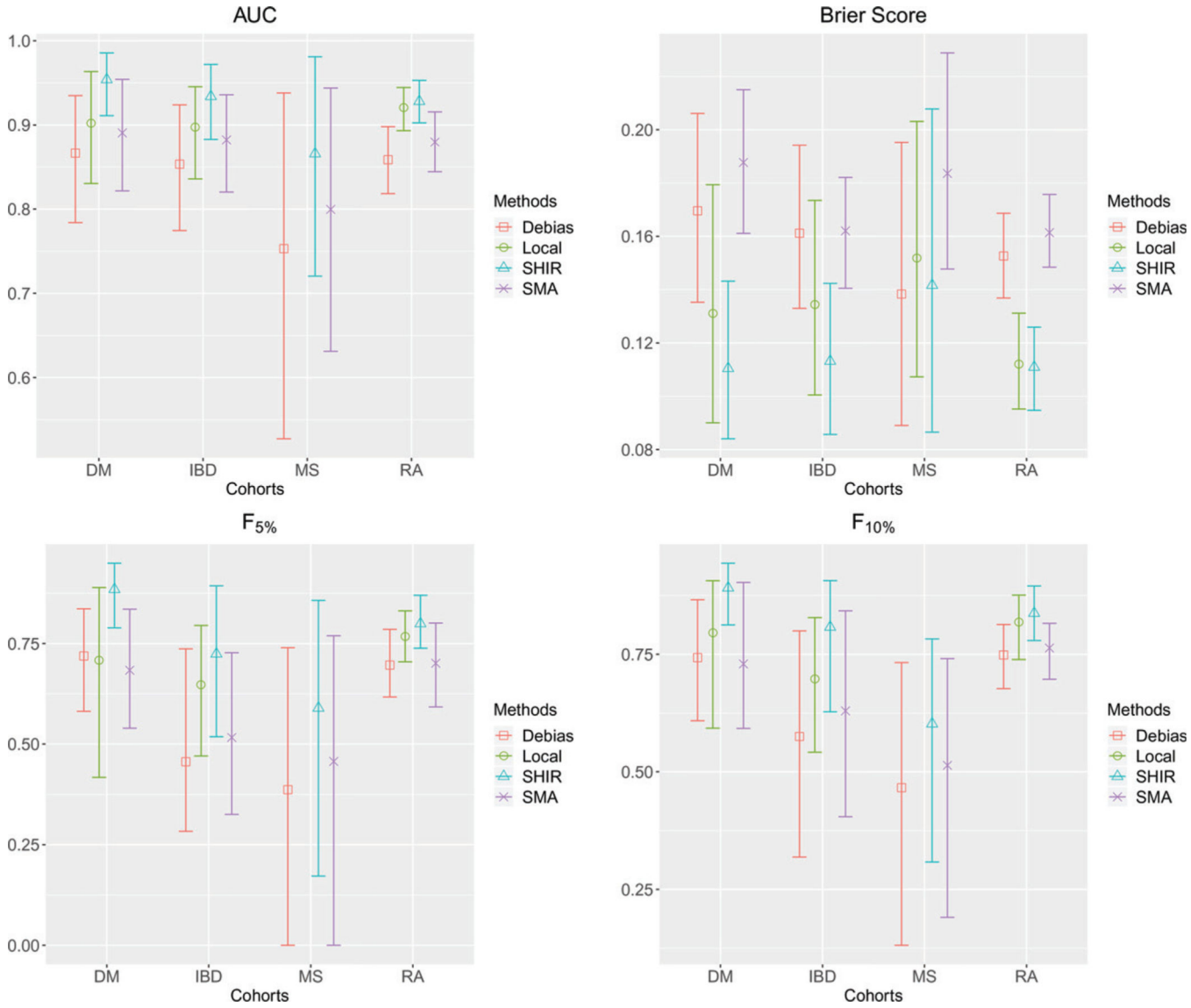

Figure 4.

The mean and 95% bootstrap confidence interval of AUC, Brier Score, $F_{5\%}$ and $F_{10\%}$ of $\text{Debias}_{\text{L\&B}}$, Local, SHIR and SMA on the validation data from the four studies.

Table 1.

Detected variables and magnitudes of their fitted coefficients for homogeneous effect μ.

Variable	$\text{Debias}_{\text{L\&B}}$	SHIR	SMA
Prescription count of statin	0.14	0.07	0
Age	0.09	0.26	0.28
Total ICD counts	−0.38	−0.75	0
NLP count of CAD	0.97	1.34	0.81
NLP count of CAD procedure related concepts	0	0.02	0
NLP count of nonsmoker	−0.07	−0.25	− 0.42
NLP count of nonsmoker > 0	−0.53	0	0
NLP count of current-smoker	0	−0.03	0
NLP count of CAD related diagnosis or procedure ≥ 1	0.06	0.05	0
ICD count for CAD	1.00	0.67	0.35
CPT count for stent or CABG	0	0.05	0
CPT count for echo	0	−0.10	0
ICD count for CAD × CPT count for echo	0	−0.04	0
NLP count of non-smoker × Oncall	0.09	0	0
NLP count of CAD × NLP count of possible-smoker	0	−0.02	0

NOTE: A×B denotes the interaction term of variables A and B. The $\log(x+1)$ transformation is taken on the count data and the covariates are normalized.