Journal of Machine Learning Research 22 (2021) 1-26 Submitted 7/20; Revised 12/20; Published 4/21

Integrative High Dimensional Multiple Testing with Heterogeneity under Data Sharing Constraints

Molei Liu MOLEI_LIU@G.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health, USA

Yin Xia XIAYIN@FUDAN.EDU.CN
Department of Statistics
School of Management, Fudan University, China

Kelly Cho KELLY.CHO@VA.GOV
Massachusetts Veterans Epidemiology Research and Information Center
US Department of Veteran Affairs
Brigham and Women's Hospital, Harvard Medical School, USA

Tianxi Cai TCAI@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health, USA

Editor: David Sontag

Abstract

Identifying informative predictors in a high dimensional regression model is a critical step for association analysis and predictive modeling. Signal detection in the high dimensional setting often fails due to the limited sample size. One approach to improving power is through meta-analyzing multiple studies which address the same scientific question. However, integrative analysis of high dimensional data from multiple studies is challenging in the presence of between-study heterogeneity. The challenge is even more pronounced with additional data sharing constraints under which only summary data can be shared across different sites. In this paper, we propose a novel data shielding integrative large–scale testing (DSILT) approach to signal detection allowing between-study heterogeneity and not requiring the sharing of individual level data. Assuming the underlying high dimensional regression models of the data differ across studies yet share similar support, the proposed method incorporates proper integrative estimation and debiasing procedures to construct test statistics for the overall effects of specific covariates. We also develop a multiple testing procedure to identify significant effects while controlling the false discovery rate (FDR) and false discovery proportion (FDP). Theoretical comparisons of the new testing procedure with the ideal individual–level meta–analysis (ILMA) approach and other distributed inference methods are investigated. Simulation studies demonstrate that the proposed testing procedure performs well in both controlling false discovery and attaining power. The new

[1] Yin Xia is the corresponding author.

method is applied to a real example detecting interaction effects of the genetic variants for statins and obesity on the risk for type II diabetes.

Keywords: Debiasing; Distributed learning; False discovery rate; High dimensional inference; Integrative analysis; Multiple testing.

1. Introduction

High throughput technologies such as genetic sequencing and natural language processing have led to an increasing number and types of predictors available to assist in predictive modeling. A critical step in developing accurate and robust prediction models is to differentiate true signals from noise. A wide range of high dimensional inference procedures have been developed in recent years to achieve variable selection, hypothesis testing and interval estimation (Van de Geer et al., 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014; Chernozhukov et al., 2018, e.g.). However, regardless of the procedure, drawing precise high dimensional inference is often infeasible in practical settings where the available sample size is too small relative to the number of predictors. One approach to improve the precision and boost power is through meta-analyzing multiple studies that address the same underlying scientific problem. This approach has been widely adopted in practice in many scientific fields, including clinical trials, education, policy evaluation, ecology, and genomics (DerSimonian, 1996; Allen et al., 2002; Card et al., 2010; Stewart, 2010; Panagiotou et al., 2013, e.g.), as a tool for evidence-based decision making. Meta-analysis is particularly valuable in the high dimensional setting. For example, meta-analysis of high dimensional genomic data from multiple studies has uncovered new disease susceptibility loci for a broad range of diseases including Crohn's disease, colorectal cancer, childhood obesity and type II diabetes (Houlston et al., 2008; Bradfield et al., 2012; Franke et al., 2010; Zeggini et al., 2008, e.g.).

Integrative analysis of high dimensional data, however, is highly challenging especially with biomedical studies for several reasons. First, between study heterogeneity arises frequently due to the difference in patient population and data acquisition. Second, due to privacy and legal constraints, individual level data often cannot be shared across study sites. Instead, only summary statistics can be passed between researchers. For example, patient level genetic data linked with clinical variables extracted from electronic health records (EHR) of several hospitals are not allowed to leave the firewall of each hospital. In addition to high dimensionality, attention to both heterogeneity and data sharing constraints are needed to perform meta-analysis of multiple EHR–linked genomic studies.

The aforementioned data sharing mechanism is referred to as DataSHIELD (Data aggregation through anonymous Summary–statistics from Harmonised Individual levEL Databases) in Wolfson et al. (2010), which has been widely accepted as a useful strategy to protect patient privacy (Jones et al., 2012; Doiron et al., 2013). Several statistical approaches to integrative analysis under the DataSHILED framework have been developed for low dimensional settings (Gaye et al., 2014; Zöller et al., 2018; Tong et al., 2020, e.g.). In the absence of cross-site heterogeneity, distributed high dimensional estimation and inference procedures have also been developed that can facilitate DataSHIELD constraints (Lee et al., 2017; Battey et al., 2018; Jordan et al., 2019, e.g.). Recently, Cai et al. (2019a) proposed an integrative high dimensional sparse regression approach that accounts for heterogeneity.

However, their method is limited to parameter estimation and variable selection. To the best of our knowledge, no hypothesis testing procedures currently exist to enable identification of significant predictors with false discovery error control under the setting of interest. In this paper, we propose a data shielding integrative large–scale testing (DSILT) procedure to fill this gap.

1.1 Problem statement

Suppose there are M independent studies and the mth study contains observations on an outcome $Y^{(\mathsf{m})}$ and a p-dimensional covariate vector $\boldsymbol{X}^{(\mathsf{m})}$, where $Y^{(\mathsf{m})}$ can be binary or continuous, and without loss of generality we assume that $\boldsymbol{X}^{(\mathsf{m})}$ contains 1 as its first element. Specifically, data from the mth study consist of n_m independent and identically distributed random vectors, $\mathcal{D}^{(\mathsf{m})} = \{\mathbf{D}_i^{(\mathsf{m})} = (Y_i^{(\mathsf{m})}, \mathbf{X}_i^{(\mathsf{m})\mathsf{T}})^{\mathsf{T}}, i = 1, ..., n_m\}$. Let $N = \sum_{m=1}^{M} n_m$ and $n = N/M$. We assume a conditional mean model $\mathrm{E}(Y^{(\mathsf{m})} \mid \boldsymbol{X}^{(\mathsf{m})}) = g(\boldsymbol{\beta}_0^{(\mathsf{m})\mathsf{T}} \boldsymbol{X}^{(\mathsf{m})})$ and that the true model parameter $\boldsymbol{\beta}_0^{(\mathsf{m})}$ is the minimizer of the population loss function:

$$\boldsymbol{\beta}_0^{(\mathsf{m})} = \underset{\boldsymbol{\beta}^{(\mathsf{m})} \in \mathbb{R}^p}{\operatorname{argmin}} \mathcal{L}_m(\boldsymbol{\beta}^{(\mathsf{m})}), \text{ where } \mathcal{L}_m(\boldsymbol{\beta}^{(\mathsf{m})}) = \mathsf{E}\{f(\mathbf{X}_i^{(\mathsf{m})\mathsf{T}} \boldsymbol{\beta}^{(\mathsf{m})}, Y_i^{(\mathsf{m})})\}, \quad f(x, y) = \phi(x) - yx,$$

where $\dot{\phi}(x) \equiv d\phi(x)/dx = g(x)$. When $\phi(x) = \log(1 + e^x)$, this corresponds to a logistic model if Y is binary and a quasi-binomial model if $Y \in [0, 1]$ is a continuous probability score sometimes generated from an EHR probabilistic phenotyping algorithm. One may take $\phi(x) = e^x$ for some non-negative Y such as the count (or log-count) of a diagnostic code in EHR studies [1]. As detailed in Assumptions 2-3 of Section 3.1, our procedure allows for a broad range of models provided that $g(\cdot)$ is smooth and the residuals $Y_i^{(\mathsf{m})} - g(\boldsymbol{\beta}_0^{(\mathsf{m})\mathsf{T}} \boldsymbol{X}_i^{(\mathsf{m})})$ are sub-Gaussian, although not all generalized linear models satisfy these assumptions.

Under the DataSHIELD constraints, the individual–level data $\mathcal{D}^{(\mathsf{m})}$ is stored at the m^{th} data computer (DC) and only summary statistics are allowed to transfer from the distributed DCs to the analysis computer (AC) at the central node. Our goal is to develop procedures under the DataSHIELD constraints for testing

$$H_{0,j} : \boldsymbol{\beta}_{0,j} \equiv (\beta_{0,j}^{(1)}, \ldots, \beta_{0,j}^{(\mathsf{M})})^{\mathsf{T}} = \mathbf{0} \text{ v.s. } H_{a,j} : \boldsymbol{\beta}_{0,j} \neq \mathbf{0} \tag{1}$$

simultaneously for $j \in \mathcal{H}$ to identify $\mathcal{H}_1 = \{j \in \mathcal{H} : \boldsymbol{\beta}_{0,j} \neq \mathbf{0}\}$, while controlling the false discovery rate (FDR) and false discovery proportion (FDP), where $\mathcal{H} \subseteq \{2, \ldots, p\}$ is a user-specified subset with $|\mathcal{H}| = q \asymp p$ and $|\mathcal{A}|$ denotes the size of any set $\mathcal{A}$. Here $\boldsymbol{\beta}_{0,j} = \mathbf{0}$ indicates that X_j is independent of Y given all remaining covariates. To ensure effective integrative analysis, we assume that $\boldsymbol{\beta}_0^{(1)}, ..., \boldsymbol{\beta}_0^{(\mathsf{M})}$ are sparse and share similar support. Specifically, we assume that $|\mathcal{S}_0| \ll p$ and $s^{(\mathsf{m})} \asymp s$ for $m = 1, 2, \ldots, M$, where $\mathcal{S}_0 = \{j = 2, ..., p : \boldsymbol{\beta}_{0,j}^{(\mathsf{m})} \neq \mathbf{0}\} = \cup_{m=1}^{M} \mathcal{S}^{(\mathsf{m})}$, $\mathcal{S}^{(\mathsf{m})} = \{j = 2, ..., p : \beta_{0,j}^{(\mathsf{m})} \neq 0\}$, $s^{(\mathsf{m})} = |\mathcal{S}^{(\mathsf{m})}|$, and $s = |\mathcal{S}_0|$.

1.2 Our contribution and the related work

We propose in this paper a novel DSILT procedure with FDR and FDP control for the simultaneous inference problem (1). The proposed testing procedure consists of three major

[1] Though a Poisson distribution does not satisfy the required sub-Gaussian residual Assumption 3, the counts of EHR diagnostic codes are usually less heavy-tailed than Poisson and are accommodated by our analysis.

steps: (I) derive an integrative estimator on the AC using locally obtained summary statistics from the DCs and send the estimator back to the DCs; (II) construct a group effect test statistic for each covariate through an integrative debiasing method; and (III) develop an error rate controlled multiple testing procedure based on the group effect statistics.

The integrative estimation approach in the first step is closely related to the group inference methods in the literature. Denote by $\boldsymbol{\beta}_j = (\beta_j^{(1)}, ..., \beta_j^{(\mathsf{M})})^\intercal$, $\boldsymbol{\beta}^{(\bullet)} = (\boldsymbol{\beta}^{(1)\intercal}, \ldots, \boldsymbol{\beta}^{(\mathsf{M})\intercal})^\intercal$,

$$\widehat{\mathcal{L}}^{(\mathsf{m})}(\boldsymbol{\beta}^{(\mathsf{m})}) = n_m^{-1} \sum_{i=1}^{n_m} f(\boldsymbol{\beta}^{(\mathsf{m})\intercal}\mathbf{X}_i^{(\mathsf{m})}, Y_i^{(\mathsf{m})}) \quad \text{and} \quad \widehat{\mathcal{L}}^{(\bullet)}(\boldsymbol{\beta}^{(\bullet)}) = N^{-1} \sum_{m=1}^{M} n_m \widehat{\mathcal{L}}_m(\boldsymbol{\beta}^{(\mathsf{m})}).$$

Literature in group LASSO and multi-task learning (Huang and Zhang, 2010; Lounici et al., 2011, e.g.) established that, under the setting $s^{(\mathsf{m})} \asymp s$ as introduced in Section 1.1, the group LASSO estimator with tuning parameter λ: $\operatorname{argmin}_{\boldsymbol{\beta}^{(\bullet)}} \widehat{\mathcal{L}}^{(\bullet)}(\boldsymbol{\beta}^{(\bullet)}) + \lambda \sum_{j=2}^{p} \|\boldsymbol{\beta}_j\|_2$, benefits from the group structure and attains the optimal rate of convergence. In this paper, we adopt the same structured group LASSO penalty for integrative estimation, but under data sharing constraints. Recently, Mitra et al. (2016) proposed a group structured debiasing approach under the integrative analysis setting, where they restricted their analysis to linear models and required that the precision matrices of the covariates be group-sparse across the distributed datasets. In contrast, our method accommodates non-linear models and imposes no sparsity or homogeneity structures on the covariate distributions from different local sites (see Assumption 1 in Section 3.1).

The second step of our method, i.e., the construction of the test statistics for each of the hypotheses, relies on the group debiasing of the above integrative estimation. For debiasing of M–estimation, nodewise LASSO regression was employed in the earlier work (Van de Geer et al., 2014; Janková and Van De Geer, 2016, e.g), while the Dantzig selector type approach was proposed more recently (Belloni et al., 2018; Caner and Kock, 2018, e.g). We develop in this article a cross–fitted group Dantzig selector type debiasing method, which requires weaker inverse Hessian assumptions (see Assumption 1 in Section 3.1) than the aforementioned approaches. In addition, the proposed debiasing step achieves proper bias rate under the same model sparsity assumptions as the ideal individual–level meta–analysis (ILMA) method. Compared with the One–shot distributed inference approaches (Tang et al., 2016; Lee et al., 2017; Battey et al., 2018), the proposed method additionally considers model heterogeneity and group inference; it further reduces the bias rate by sending the integrative estimator to the DCs to derive updated summary statistics, which in turn benefits the subsequent multiple testing procedure. See Section 3.4 for detailed comparisons.

As the last step, simultaneous inference with theoretical error rates control is performed based on the group effect statistics. The test statistics are shown to be asymptotically chi-square distributed under the null, and the proposed multiple testing procedure asymptotically controls both the FDR and FDP at the pre-specified level. Multiple testing for high dimensional regression models has recently been studied in the literature (Liu and Luo, 2014; Xia et al., 2018b,a; Javanmard et al., 2019, e.g). Our testing step for FDR control as a whole differs considerably from these existing procedures in the following aspects. First, the proposed test statistics, the key input to the FDR control procedure, are brand new and the resulting estimation of false discovery proportion differs fundamentally from those of the literature. Second, we consider a more general M–estimation setting which can accommodate different types of outcomes. Third, we allow the heterogeneity in both the covariates and the coefficients. Fourth, the existing testing approaches developed for individual-level data are not suitable for the DataSHIELD framework. Last, because there are complicated dependence structures among the integrative chi-squared statistics under the DataSHIELD constraints, the theoretical derivations are technically much more involved. Hence, our proposal makes a useful addition to the general toolbox of simultaneous regression inference.

We demonstrate here via numerical experiments that the proposed DSILT procedure attains good power while maintaining error rate control. In addition, we demonstrate how our new approach outperforms existing distributed inference methods and enjoys similar performance as the ideal ILMA approach.

1.3 Outline of the paper

The rest of this paper is organized as follows. We detail the DSILT approach in Section 2. In Section 3, we present asymptotic analysis on the false discovery control of our method and compare it with the ILMA and One–shot approach. In Section 4, we summarize finite sample performance of our approach along with other methods from simulation studies. In Section 5, we apply our proposed method to a real example. Proofs of the theoretical results and additional technical lemmas and simulation results are collected in the Supplementary Material.

2. Data shielding integrative large–scale testing procedure

In this section, we study the detailed procedure of the proposed method. We start with some notation that will be used throughout the paper.

2.1 Notation

For any integer d, any vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \mathbb{R}^d$, and any set $\mathcal{S} = \{j_1, \ldots, j_k\} \subseteq [d] \equiv \{1, \ldots, d\}$, denote by $\boldsymbol{x}_\mathcal{S} = [x_{j_1}, \ldots, x_{j_k}]'$, $\boldsymbol{x}_{\text{-}j}$ the vector with its j^{th} entry removed from $\boldsymbol{x}$, $\|\boldsymbol{x}\|_q$ the ℓ_q norm of $\boldsymbol{x}$ and $\|\boldsymbol{x}\|_\infty = \max_{j\in[d]} |x_j|$. For any d-dimensional vectors $\{\boldsymbol{a}^{(\mathsf{m})} = (a_1^{(\mathsf{m})}, \ldots, a_d^{(\mathsf{m})})^\mathsf{T}, m \in [M]\}$ and $\mathcal{S} \subseteq [d]$, let $\boldsymbol{a}^{(\bullet)} = (\boldsymbol{a}^{(1)\mathsf{T}}, \ldots, \boldsymbol{a}^{(\mathsf{M})\mathsf{T}})^\mathsf{T}$, $\boldsymbol{a}_\mathcal{S}^{(\bullet)} = (\boldsymbol{a}_\mathcal{S}^{(1)\mathsf{T}}, \ldots, \boldsymbol{a}_\mathcal{S}^{(\mathsf{M})\mathsf{T}})^\mathsf{T}$, $\boldsymbol{a}_j = (a_j^{(1)}, \ldots, a_j^{(\mathsf{M})})^\mathsf{T}$, and $\boldsymbol{a}_{\text{-}j}^{(\bullet)} = (\boldsymbol{a}_{\text{-}j}^{(1)\mathsf{T}}, \ldots, \boldsymbol{a}_{\text{-}j}^{(\mathsf{M})\mathsf{T}})^\mathsf{T}$. Let $\boldsymbol{e}_j$ be the unit vector with j^{th} element being 1 and remaining elements being 0 and $\boldsymbol{e}_j^{(\bullet)} = (\boldsymbol{e}_j^\mathsf{T}, \ldots, \boldsymbol{e}_j^\mathsf{T})^\mathsf{T}$. Denote by $\|\boldsymbol{a}^{(\bullet)}\|_{2,1} = \sum_{j=1}^d \|\boldsymbol{a}_j\|_2$ and $\|\boldsymbol{a}^{(\bullet)}\|_{2,\infty} = \max_{j\in[d]} \|\boldsymbol{a}_j\|_2$ the ℓ_2/ℓ_1 and ℓ_2/ℓ_∞ norm of $\boldsymbol{a}^{(\bullet)}$ respectively. For any K–fold partition of $[n_m]$, denoted by $\{\mathcal{I}_k^{(\mathsf{m})}, k \in [K]\}$, let $\mathcal{I}_{\text{-}k}^{(\mathsf{m})} = [n_m] \backslash \mathcal{I}_k^{(\mathsf{m})}$, $\mathcal{I}_k^{(\bullet)} = \{\mathcal{I}_k^{(\mathsf{m})} : m \in [M]\}$, $\mathcal{I}_{\text{-}k}^{(\bullet)} = \{\mathcal{I}_{\text{-}k}^{(\mathsf{m})} : m \in [M]\}$. For any index set $\mathcal{I}^{(\bullet)} = \{\mathcal{I}^{(\mathsf{m})} \subseteq [n_m], m \in [M]\}$, $\mathcal{D}_{\mathcal{I}^{(\mathsf{m})}}^{(\mathsf{m})} = \{\mathbf{D}_i^{(\mathsf{m})} : i \in \mathcal{I}^{(\mathsf{m})}\}$, and $\mathcal{D}_{\mathcal{I}^{(\bullet)}}^{(\bullet)} = \{\mathcal{D}_{\mathcal{I}^{(\mathsf{m})}}^{(\mathsf{m})} : m \in [M]\}$. Let $\ddot{\phi}(\theta) = d^2\phi(\theta)/d\theta^2 \geq 0$. Denote by $\boldsymbol{\beta}_{0,j}$ and $\boldsymbol{\beta}_0^{(\bullet)}$ the true values of $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}^{(\bullet)}$ respectively. For any $\mathcal{I}^{(\bullet)}$ and $\boldsymbol{\beta}^{(\bullet)}$, define the sample measure operators $\widehat{\mathscr{P}}_{\mathcal{I}^{(\mathsf{m})}} \eta_{\boldsymbol{\beta}^{(\mathsf{m})}} = |\mathcal{I}^{(\mathsf{m})}|^{-1} \sum_{i\in\mathcal{I}^{(\mathsf{m})}} \eta_{\boldsymbol{\beta}^{(\mathsf{m})}}(\mathbf{D}_i^{(\mathsf{m})})$ and $\widehat{\mathscr{P}}_{\mathcal{I}^{(\bullet)}} \eta_{\boldsymbol{\beta}^{(\bullet)}} = |\mathcal{I}^{(\bullet)}|^{-1} \sum_{m=1}^M \sum_{i\in\mathcal{I}^{(\mathsf{m})}} \eta_{\boldsymbol{\beta}^{(\mathsf{m})}}(\mathbf{D}_i^{(\mathsf{m})})$, and the population measure operator $\mathscr{P}^{(\mathsf{m})} \eta_{\boldsymbol{\beta}^{(\mathsf{m})}} = \mathsf{E}\eta_{\boldsymbol{\beta}^{(\mathsf{m})}}(\mathbf{D}_i^{(\mathsf{m})})$, for all integrable functions $\eta_{\boldsymbol{\beta}^{(\bullet)}} = \{\eta_{\boldsymbol{\beta}^{(\mathsf{m})}}, m \in [M]\}$ parameterized by $\boldsymbol{\beta}^{(\bullet)}$ or $\boldsymbol{\beta}^{(\mathsf{m})}$.

For any given $\boldsymbol{\beta}^{(\mathsf{m})}$, we define $\theta_i^{(\mathsf{m})} = \mathbf{X}_i^{(\mathsf{m})\intercal}\boldsymbol{\beta}^{(\mathsf{m})}$, $\theta_{0,i}^{(\mathsf{m})} = \mathbf{X}_i^{(\mathsf{m})\intercal}\boldsymbol{\beta}_0^{(\mathsf{m})}$, and the residual $\epsilon_i^{(\mathsf{m})} := Y_i^{(\mathsf{m})} - \dot{\phi}(\theta_{0,i}^{(\mathsf{m})})$. Similar to Cai et al. (2019b) and Ma et al. (2020), given coefficient $\boldsymbol{\beta}^{(\mathsf{m})}$, we can express $Y_i^{(\mathsf{m})} \sim \mathbf{X}_i^{(\mathsf{m})}$ in an approximately linear form:

$$Y_i^{(\mathsf{m})} - \dot{\phi}(\theta_i^{(\mathsf{m})}) + \ddot{\phi}(\theta_i^{(\mathsf{m})})\theta_i^{(\mathsf{m})} = \ddot{\phi}(\theta_i^{(\mathsf{m})})\mathbf{X}_i^{(\mathsf{m})\intercal}\boldsymbol{\beta}_0^{(\mathsf{m})} + \epsilon_i^{(\mathsf{m})} + R_i^{(\mathsf{m})}(\theta_i^{(\mathsf{m})}),$$

where $R_i^{(\mathsf{m})}(\theta_i^{(\mathsf{m})})$ is the reminder term and $R_i^{(\mathsf{m})}(\theta_{0,i}^{(\mathsf{m})}) = 0$. For a given observation set $\mathbf{D}$ and coefficient $\boldsymbol{\beta}$, we let $\theta = \mathbf{X}^\intercal\boldsymbol{\beta}$, $Y_{\boldsymbol{\beta}} = \ddot{\phi}^{-\frac{1}{2}}(\theta)\left\{Y - \dot{\phi}(\theta) + \ddot{\phi}(\theta)\theta\right\}$, $\mathbf{X}_{\boldsymbol{\beta}} = \ddot{\phi}^{\frac{1}{2}}(\theta)\mathbf{X}$. Note that for the logistic model, we have $\mathrm{Var}(Y_{\boldsymbol{\beta}}|\mathbf{X}_{\boldsymbol{\beta}}) = 1$, and $\mathbf{X}_{\boldsymbol{\beta}}$ and $Y_{\boldsymbol{\beta}}$ can be viewed as the covariates and responses adjusted for the heteroscedasticity of the residuals.

2.2 Outline of the proposed testing procedure

We first outline in this section the DSILT procedure in Algorithm 1 and then study the details of each key step later in Sections 2.3 to 2.5. The procedure involves partitioning of $\mathcal{D}^{(\mathsf{m})}$ into K folds $\{\mathcal{I}_k^{(\mathsf{m})} : k \in [K]\}$ for $m \in [M]$, where without loss of generality we let $K \geq 2$ be an even number. With a slight abuse of notation, we write $\mathcal{D}_{[k]}^{(\mathsf{m})} = \mathcal{D}_{\mathcal{I}_k^{(\mathsf{m})}}^{(\mathsf{m})}$, $\mathcal{D}_{[k]}^{(\bullet)} = \mathcal{D}_{\mathcal{I}_k^{(\bullet)}}^{(\bullet)}$, $\mathcal{D}_{[\text{-}k]}^{(\mathsf{m})} = \mathcal{D}_{\mathcal{I}_{\text{-}k}^{(\mathsf{m})}}^{(\mathsf{m})}$, and $\mathcal{D}_{[\text{-}k]}^{(\bullet)} = \mathcal{D}_{\mathcal{I}_{\text{-}k}^{(\bullet)}}^{(\bullet)}$.

Algorithm 1 DSILT Algorithm.

Input: $\mathcal{D}^{(\mathsf{m})}$ at the m^{th} DC for $m \in [M]$.

1. For each $k \in [K]$, fit **integrative sparse regression under DataSHIELD** with $\mathcal{D}_{[\text{-}k]}^{(\bullet)}$:
   (a) At the m^{th} DC, construct cross-fitted summary statistics based on local LASSO estimator, and send them to the AC;
   (b) Obtain the integrative estimator $\widetilde{\boldsymbol{\beta}}_{[\text{-}k]}^{(\bullet)}$ at AC and send them back to each DC.
2. **Obtain debiased group test statistics**:
   (a) For each k, at the m^{th} DC, obtain the updated summary statistics based on $\widetilde{\boldsymbol{\beta}}_{[\text{-}k]}^{(\bullet)}$ and $\mathcal{D}_{[k]}^{(\mathsf{m})}$, and send them to the AC;
   (b) At the AC, construct cross-fitted debiased group estimators $\{\breve{\zeta}_j, j \in \mathcal{H}\}$.
3. Construct a multiple testing procedure based on the test statistics from Step 2.

2.3 Step 1: Integrative sparse regression

As a first step, we fit integrative sparse regression under DataSHIELD with $\mathcal{D}_{[\text{-}k]}^{(\bullet)}$ following similar strategies as given in Cai et al. (2019a). To carry out Step 1(a) of Algorithm 1, we split the index set $\mathcal{I}_{\text{-}k}^{(\mathsf{m})}$ into K' folds $\mathcal{I}_{\text{-}k,1}^{(\mathsf{m})}, \ldots, \mathcal{I}_{\text{-}k,K'}^{(\mathsf{m})}$. For $k \in [K]$ and $k' \in [K']$, we construct local LASSO estimator with tuning parameter $\lambda^{(\mathsf{m})}$: $\widehat{\boldsymbol{\beta}}_{[\text{-}k,\text{-}k']}^{(\mathsf{m})} = \mathrm{argmin}_{\boldsymbol{\beta}^{(\mathsf{m})} \in \mathbb{R}^p}\, \widehat{\mathscr{P}}_{\mathcal{I}_{\text{-}k}^{(\mathsf{m})} \setminus \mathcal{I}_{\text{-}k,k'}^{(\mathsf{m})}} f(\mathbf{X}^\intercal\boldsymbol{\beta}^{(\mathsf{m})}, Y) + \lambda^{(\mathsf{m})}\|\boldsymbol{\beta}_{\text{-}1}^{(\mathsf{m})}\|_1$. With $\mathcal{D}_{[\text{-}k]}^{(\mathsf{m})}$, we then derive summary data $\mathcal{S}^{(\mathsf{m})}_{[\text{-}k]} = \{|\mathcal{I}^{(\mathsf{m})}_{\text{-}k}|, \widehat{\boldsymbol{\xi}}^{(\mathsf{m})}_{[\text{-}k]}, \widehat{\mathbb{H}}^{(\mathsf{m})}_{[\text{-}k]}\}$, where

$$\widehat{\boldsymbol{\xi}}^{(\mathsf{m})}_{[\text{-}k]} = K'^{-1} \sum_{k'=1}^{K'} \widehat{\mathscr{P}}_{\mathcal{I}^{(\mathsf{m})}_{\text{-}k,k'}} \mathbf{X}_{\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k,\text{-}k']}} Y_{\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k,\text{-}k']}}, \quad \widehat{\mathbb{H}}^{(\mathsf{m})}_{[\text{-}k]} = K'^{-1} \sum_{k'=1}^{K'} \widehat{\mathscr{P}}_{\mathcal{I}^{(\mathsf{m})}_{\text{-}k,k'}} \mathbf{X}_{\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k,\text{-}k']}} \mathbf{X}^{\mathsf{T}}_{\widehat{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k,\text{-}k']}}. \tag{2}$$

In Step 1(b) of Algorithm 1, for $k \in [K]$, we aggregate the M sets of summary data $\{\mathcal{S}^{(\mathsf{m})}_{[\text{-}k]}, m \in [M]\}$ at the central AC and solve a regularized quasi–likelihood problem to obtain the integrative estimator with tuning parameter λ:

$$\widetilde{\boldsymbol{\beta}}^{(\bullet)}_{[\text{-}k]} = \underset{\boldsymbol{\beta}^{(\bullet)}}{\operatorname{argmin}} |\mathcal{I}^{(\bullet)}_{\text{-}k}|^{-1} \sum_{m=1}^{M} |\mathcal{I}^{(\mathsf{m})}_{\text{-}k}| \left(\boldsymbol{\beta}^{(\mathsf{m})\mathsf{T}} \widehat{\mathbb{H}}^{(\mathsf{m})}_{[\text{-}k]} \boldsymbol{\beta}^{(\mathsf{m})} - 2\boldsymbol{\beta}^{(\mathsf{m})\mathsf{T}} \widehat{\boldsymbol{\xi}}^{(\mathsf{m})}_{[\text{-}k]}\right) + \lambda \|\boldsymbol{\beta}^{(\bullet)}_{\text{-}1}\|_{2,1}. \tag{3}$$

These K sets of estimators, $\{\widetilde{\boldsymbol{\beta}}^{(\bullet)}_{[\text{-}k]}, k \in [K]\}$, are then sent back to the DCs. The summary statistics introduced in (2) can be viewed as the covariance terms of $\mathcal{D}^{(\mathsf{m})}_{[\text{-}k]}$ with the local LASSO estimator plugged-in to adjust for the heteroscedasticity of the residuals. Cross–fitting is used to remove the dependence of the observed data and the fitted outcomes - a strategy frequently employed in high dimensional inference literatures (Chernozhukov et al., 2016, 2018). As in Cai et al. (2019a), the integrative procedure can also be viewed in such a way that $\boldsymbol{\beta}^{(\mathsf{m})\mathsf{T}} \widehat{\mathbb{H}}^{(\mathsf{m})}_{[\text{-}k]} \boldsymbol{\beta}^{(\mathsf{m})} - 2\boldsymbol{\beta}^{(\mathsf{m})\mathsf{T}} \widehat{\boldsymbol{\xi}}^{(\mathsf{m})}_{[\text{-}k]}$ provides a second order one–step approximation to the individual–level data loss function $2\widehat{\mathscr{P}}_{\mathcal{I}^{(\mathsf{m})}_{\text{-}k}} f(\mathbf{X}^{\mathsf{T}}\boldsymbol{\beta}^{(\mathsf{m})}, Y)$ initializing with the local LASSO estimators. In contrast to Cai et al. (2019a), we introduce a cross–fitting procedure at each local DC to reduce fitting bias and this in turn relaxes their uniformly- bounded assumption on $\mathbf{X}_i^{(\mathsf{m})\mathsf{T}} \boldsymbol{\beta}^{(\mathsf{m})}$ for each i and m, i.e., Condition 4(i) of Cai et al. (2019a).

2.4 Step 2: Debiased group test statistics

We next derive group effect test statistics in Step 2 by constructing debiased estimators for $\boldsymbol{\beta}_0^{(\bullet)}$ and estimating their variances. In Step 2(a), we construct updated summary statistics

$$\widetilde{\boldsymbol{\xi}}^{(\mathsf{m})}_{[k]} = \widehat{\mathscr{P}}_{\mathcal{I}^{(\mathsf{m})}_{k}} \mathbf{X}_{\widetilde{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k]}} Y_{\widetilde{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k]}}, \ \widetilde{\mathbb{H}}^{(\mathsf{m})}_{[k]} = \widehat{\mathscr{P}}_{\mathcal{I}^{(\mathsf{m})}_{k}} \mathbf{X}_{\widetilde{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k]}} \mathbf{X}^{\mathsf{T}}_{\widetilde{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k]}} \text{ and } \widetilde{\mathbb{J}}^{(\mathsf{m})}_{[k]} = \widehat{\mathscr{P}}_{\mathcal{I}^{(\mathsf{m})}_{k}} \mathbf{X}\mathbf{X}^{\mathsf{T}} \left\{Y - \dot{\phi}(\mathbf{X}^{\mathsf{T}} \widetilde{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k]})\right\}^2$$

at the mth DC, for $k \in [K]$. These mK sets of summary statistics are then sent to the AC in Step 2(b) to be aggregated and debiased. Specifically, for each $j \in \mathcal{H}$ and $k \in [K]$, we solve the group Dantzig selector type optimization problem:

$$\widehat{\boldsymbol{u}}^{(\bullet)}_{j,[k]} = \underset{\boldsymbol{u}^{(\bullet)}}{\operatorname{argmin}} \max_{m \in [M]} \|\boldsymbol{u}^{(\mathsf{m})}\|_1 \quad \text{s.t.} \quad \|\widetilde{\mathbb{H}}^{(\bullet)}_{[k]} \boldsymbol{u}^{(\bullet)} - \boldsymbol{e}^{(\bullet)}_{j}\|_{2,\infty} \leq \tau, \tag{4}$$

to obtain a vector of projection directions for some tuning parameter τ, where $\widetilde{\mathbb{H}}^{(\bullet)}_{[k]} = \mathsf{diag}\{\widetilde{\mathbb{H}}^{(1)}_{[k]}, \ldots, \widetilde{\mathbb{H}}^{(\mathsf{M})}_{[k]}\}$. Combining across the K splits, we construct the cross–fitted group debiased estimator for $\beta^{(\mathsf{m})}_j$ by $\breve{\beta}^{(\mathsf{m})}_j = K^{-1} \sum_{k=1}^{K} \left\{\widetilde{\beta}^{(\mathsf{m})}_{j,[\text{-}k]} + \widehat{\boldsymbol{u}}^{(\mathsf{m})\mathsf{T}}_{j,[k]} (\widetilde{\boldsymbol{\xi}}^{(\mathsf{m})}_{[k]} - \widetilde{\mathbb{H}}^{(\mathsf{m})}_{[k]} \widetilde{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k]})\right\}$.

In Section 3.2, we show that the distribution of $n_m^{1/2}(\breve{\beta}^{(\mathsf{m})}_j - \beta_{0,j})$ is approximately normal with mean 0 and variance $(\sigma^{(\mathsf{m})}_{0,j})^2$, estimated by $(\widehat{\sigma}^{(\mathsf{m})}_j)^2 = K^{-1} \sum_{k=1}^{K} \widehat{\boldsymbol{u}}^{(\mathsf{m})\mathsf{T}}_{j,[k]} \widetilde{\mathbb{J}}^{(\mathsf{m})}_{[k]} \widehat{\boldsymbol{u}}^{(\mathsf{m})}_{j,[k]}$. Finally, we test for the group effect of the j-th covariate across M studies based on the standardized sum of square type statistics

$$\breve{\zeta}_j = \sum_{m=1}^{M} n_m \{\breve{\beta}_j^{(\mathsf{m})}/\widehat{\sigma}_j^{(\mathsf{m})}\}^2, \text{ for } j \in \mathcal{H}.$$

We show in Section 3.2 that, under mild regularity assumptions, $\breve{\zeta}_j$ is asymptotically chi-square distributed with degree of freedom M under the null. This result is crucial to ensure the error rate control for the downstream multiple testing procedure.

2.5 Step 3: Multiple testing

To construct an error rate controlled multiple testing procedure for

$$H_{0,j} : \boldsymbol{\beta}_{0,j} = \mathbf{0} \text{ versus } H_{a,j} : \boldsymbol{\beta}_{0,j} \neq \mathbf{0}, \quad j \in \mathcal{H} \subseteq \{2, \ldots, p\},$$

we first take a normal quantile transformation of $\breve{\zeta}_j$, namely $\mathcal{N}_j = \bar{\Phi}^{-1}\left\{\bar{\mathbb{F}}_{\chi^2_M}(\breve{\zeta}_j)/2\right\}$, where Φ is the standard normal cumulative distribution function, $\bar{\Phi} = 1 - \Phi$, and $\bar{\mathbb{F}}_{\chi^2_M}(\cdot)$ is the survival function of χ^2_M. Based on the asymptotic χ^2_M distribution of $\breve{\zeta}_j$ as will be shown in Theorem 1, we present in the proof of Theorem 2 that $\mathcal{N}_j$ asymptotically has the same distribution as the absolute value of a standard normal random variable. Thus, to test a single hypothesis of $H_{0,j} : \boldsymbol{\beta}_{0,j} = \mathbf{0}$, we reject the the null at nominal level $\alpha > 0$ whenever $\Psi_{\alpha,j} = 1$, where $\Psi_{\alpha,j} = I\left\{\mathcal{N}_j \geq \bar{\Phi}^{-1}(\alpha/2)\right\}$.

However, for simultaneous inference across q hypotheses $\{H_{0,j}, j \in \mathcal{H}\}$, we shall further adjust the multiplicity of the tests as follows. For any threshold level t, let $R_0(t) = \sum_{j\in\mathcal{H}_0} I(\mathcal{N}_j \geq t)$ and $R(t) = \sum_{j\in\mathcal{H}} I(\mathcal{N}_j \geq t)$ respectively denote the total number of false positives and the total number of rejections associated with t, where $\mathcal{H}_0 = \{j \in \mathcal{H} : \boldsymbol{\beta}_{0,j} = \mathbf{0}\}$. Then the FDP and FDR for a given t are respectively defined as

$$\mathsf{FDP}(t) = \frac{R_0(t)}{R(t) \vee 1} \quad \text{and} \quad \mathsf{FDR}(t) = \mathsf{E}\{\mathsf{FDP}(t)\}.$$

The smallest t such that $\mathsf{FDP}(t) \leq \alpha$, namely $t_0 = \inf\left\{0 \leq t \leq (2\log q)^{1/2} : \mathsf{FDP}(t) \leq \alpha\right\}$ would be a desirable threshold since it maximizes the power under the FDP control. However, since the null set is unknown, we estimate $R_0(t)$ by $2\bar{\Phi}(t)|\mathcal{H}_0|$ and conservatively estimate $|\mathcal{H}_0|$ by q because of the model sparsity. We next calculate

$$\hat{t} = \inf\left\{0 \leq t \leq t_q : \frac{2q\bar{\Phi}(t)}{R(t) \vee 1} \leq \alpha\right\} \quad \text{where} \quad t_q = (2\log q - 2\log\log q)^{\frac{1}{2}} \tag{5}$$

to approximate the ideal threshold t_0. If (5) does not exist, we set $\hat{t} = (2\log q)^{1/2}$. Finally, we obtain the rejection set $\{j : \mathcal{N}_j \geq \hat{t}, j \in \mathcal{H}\}$ as the output of Algorithm 1. The theoretical analysis of the asymptotic error rates control of the proposed multiple testing procedure will be studied in Section 3.3.

Remark 1. *Our testing approach is different from the BH procedure (Benjamini and Hochberg, 1995) in that, the latter obtains the rejection set* $\{j : \mathcal{N}_j \geq \hat{t}_{BH}, j \in \mathcal{H}\}$ *with*

$\hat{t}_{BH} = \inf\left\{t \geq 0 :\ 2q\bar{\Phi}(t)/\{R(t) \vee 1\} \leq \alpha\right\}$. *Note that, first, the range* $[0, t_q]$ *in our procedure is critical, because when* $t \geq (2\log q - \log\log q)^{\frac{1}{2}}$, $R_0(t)$ *is no longer consistently estimated by* $2q\bar{\Phi}(t)$. *As a result, the BH may not able to control the FDP with positive probability. Second, in the proposed approach, if* $\hat{t}$ *is not attained in the range, it is crucial to threshold it at* $(2\log q)^{1/2}$, *instead of* t_q, *because the latter will cause too many false rejections, and as a result the FDR cannot be properly controlled.*

2.6 Tuning parameter selection

In this section, we detail data-driven procedures for selecting the tuning parameters $\boldsymbol{\eta} = \{\boldsymbol{\lambda}^{(\bullet)} = (\lambda^{(1)}, \ldots, \lambda^{(\mathsf{M})})^{\mathsf{T}}, \lambda, \tau\}$. Since our primary goal is to perform simultaneous testing, we follow a similar strategy as that of Xia et al. (2018a) and select tuning parameters to minimize a ℓ_2 distance between $\widehat{R}_0(t)/\{2|\mathcal{H}_0|\bar{\Phi}(t)\}$ and its expected value of 1, where $\widehat{R}_0(t)$ is an estimate of $R_0(t)$ from the testing procedure. However, unlike Xia et al. (2018a), it is not feasible to tune $\boldsymbol{\eta}$ simultaneously due to DataSHIELD constraints. We instead tune $\boldsymbol{\lambda}^{(\bullet)}$, λ and τ sequentially as detailed below. Furthermore, based on the theoretical analyses of the optimal rates for $\boldsymbol{\eta}$ given in Section 3, we select $\boldsymbol{\eta}$ within a set of candidate values that are of the same order as their respective optimal rates.

First for $\boldsymbol{\lambda}^{(\bullet)}$ in Algorithm 1, we tune $\lambda^{(\mathsf{m})}$ via cross validation within the mth DC. Second, to select λ for the integrative estimation in (3), we minimize an approximated generalized information criterion that only involves derived data from M studies. Specifically, we choose λ as the minimizer of $\text{GIC}\left(\lambda, \widetilde{\boldsymbol{\beta}}^{(\bullet)}_{[\text{-}k],\lambda}\right) = \text{Dev}\left(\widetilde{\boldsymbol{\beta}}^{(\bullet)}_{[\text{-}k],\lambda}\right) + \gamma\text{DF}\left(\lambda, \widetilde{\boldsymbol{\beta}}^{(\bullet)}_{[\text{-}k],\lambda}\right)$, where γ is some pre-specified scaling parameter, $\widetilde{\boldsymbol{\beta}}^{(\mathsf{m})}_{[\text{-}k],\lambda}$ is the estimator obtained with λ,

$$\text{Dev}\left(\boldsymbol{\beta}^{(\bullet)}\right) = |\mathcal{I}_{\text{-}k}|^{-1}\sum_{m=1}^{M}|\mathcal{I}^{(\mathsf{m})}_{\text{-}k}|\left(\boldsymbol{\beta}^{(\mathsf{m})\mathsf{T}}\widehat{\mathbb{H}}^{(\mathsf{m})}_{[\text{-}k]}\boldsymbol{\beta}^{(\mathsf{m})} - 2\boldsymbol{\beta}^{(\mathsf{m})\mathsf{T}}\widehat{\boldsymbol{\xi}}^{(\mathsf{m})}_{[\text{-}k]}\right) \quad \text{and}$$

$$\text{DF}\left(\lambda, \boldsymbol{\beta}^{(\bullet)}\right) = \left[\partial^2_{\widehat{\mathcal{S}}}\left\{\text{Dev}\left(\boldsymbol{\beta}^{(\bullet)}\right) + \lambda\|\boldsymbol{\beta}^{(\bullet)}_{\text{-}1}\|_{2,1}\right\}\right]^{-1}\left[\partial^2_{\widehat{\mathcal{S}}}\text{Dev}\left(\boldsymbol{\beta}^{(\bullet)}\right)\right],$$

are respectively the approximated deviance and degree of freedom measures, $\widehat{\mathcal{S}}$ is the set of non-zero elements in $\boldsymbol{\beta}^{(\bullet)}$ and the operator $\partial^2_{\widehat{\mathcal{S}}}$ denotes the second order partial derivative with respect to $\boldsymbol{\beta}^{(\bullet)}_{\widehat{\mathcal{S}}}$. Common choices of γ include $2|\mathcal{I}_{\text{-}k}|^{-1}$ (AIC), $|\mathcal{I}_{\text{-}k}|^{-1}\log|\mathcal{I}_{\text{-}k}|$ (BIC), $|\mathcal{I}_{\text{-}k}|^{-1}\log|\mathcal{I}_{\text{-}k}|\log\log p$ (Wang et al., 2009, modified BIC) and $2|\mathcal{I}_{\text{-}k}|^{-1}\log|\mathcal{I}_{\text{-}k}|\log p$ (Foster and George, 1994, RIC). For numerical studies in Sections 4 and 5, we use BIC which appears to perform well across settings.

At the last step, we tune τ by minimizing an ℓ_2 distance between $\widehat{R}_{0,\mathsf{null}}(t \mid \tau)/\{2q\bar{\Phi}(t)\}$ and 1, where $\widehat{R}_{0,\mathsf{null}}(t \mid \tau)$ is an estimate of $R_0(t)$ with a given tuning parameter τ, and we replace $\mathcal{H}_0$ by q as in Xia et al. (2018a). Our construction of $\widehat{R}_{0,\mathsf{null}}(t \mid \tau)$ differs from that of Xia et al. (2018a) in that we estimate $R_0(t)$ under the complete null to better approximate the denominator of $2q\bar{\Phi}(t)$. As detailed in Algorithm 2, we construct $\breve{\beta}^{(\mathsf{m})}_{j,\mathsf{null}}$ as the difference between the estimator obtained with the first $K/2$ folds of data and the corresponding estimator obtained using the second $K/2$ folds of data, which is always centered around 0 rather than $\beta^{(\mathsf{m})}_{0j}$. Since the accuracy of $\widehat{R}_{0,\mathsf{null}}(t \mid \tau)$ for large t is most relevant to the error control, we construct the distance measure $\widehat{d}(\tau)$ in Algorithm 2 focusing on t around $\bar{\Phi}^{-1}[\bar{\Phi}\{(2\log q)^{1/2}\}\iota]$ for some values of $\iota \in (0,1]$.

Algorithm 2 Selection of τ for multiple testing.

1. For any given τ and each $j \in \mathcal{H}$, calculate $\breve{\zeta}_{j,\mathsf{null}}(\tau) = \sum_{m=1}^{M} n_m \{\breve{\beta}_{j,\mathsf{null}}^{(\mathsf{m})}(\tau)/\widehat{\sigma}_j^{(\mathsf{m})}\}^2$ with

$$\breve{\beta}_{j,\mathsf{null}}^{(\mathsf{m})}(\tau) = K^{-1}\sum_{k=1}^{K}(-1)^{k>K/2}\left\{\widetilde{\beta}_{j,[\text{-}k]}^{(\mathsf{m})} + \widehat{\boldsymbol{u}}_{j,[k]}^{(\mathsf{m})\mathsf{T}}(\tau)\left(\widetilde{\boldsymbol{\xi}}_{[k]}^{(\mathsf{m})} - \widetilde{\mathbb{H}}_{[k]}^{(\mathsf{m})}\widetilde{\boldsymbol{\beta}}_{[\text{-}k]}^{(\mathsf{m})}\right)\right\},$$

   where $\widehat{\boldsymbol{u}}_{j,[k]}^{(\bullet)}(\tau)$ is the debiasing projection direction obtained at tuning value τ.

2. Define $\widehat{R}_{0,\mathsf{null}}(t \mid \tau) = \sum_{j\in\mathcal{H}} I[\bar{\mathbb{F}}_{\chi_M^2}\{\breve{\zeta}_{j,\mathsf{null}}(\tau)\} \le 2\bar{\Phi}(t)]$ and a modified measure

$$\widehat{d}(\tau) = \int_0^1 \left[\widehat{R}_{0,\mathsf{null}}\{\bar{\Phi}^{-1}(x) \mid \tau\}/(2qx) - 1\right]^2 d\widehat{\omega}(x),$$

   where $\widehat{\omega}(x) = H^{-1}\sum_{h=1}^{H} I(\bar{\Phi}\{(2\log q)^{1/2}\}h/H \le x)$ and $H > 0$ is some specified constant.

3. Theoretical Results

In this section, we present the asymptotic analysis results of the proposed method and compare it with alternative approaches.

3.1 Notation and assumptions

For any semi–positive definite matrix $\mathbb{A} \in \mathbb{R}^{d\times d}$ and $i,j \in [d]$, denote by $\mathbb{A}_{ij}$ the $(i,j)^{\mathsf{th}}$ element of $\mathbb{A}$ and $\mathbb{A}_j$ its j^{th} row, $\Lambda_{\min}(\mathbb{A})$ and $\Lambda_{\max}(\mathbb{A})$ the smallest and largest eigenvalue of $\mathbb{A}$. Define the sub-gaussian norms of a random variable X and a d-dimensional random vector $\boldsymbol{X}$, respectively by $\|X\|_{\psi_2} := \sup_{q\ge1} q^{-1/2}(\mathsf{E}|X|^q)^{1/q}$ and $\|\boldsymbol{X}\|_{\psi_2} := \sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \|\boldsymbol{x}^\mathsf{T}\boldsymbol{X}\|_{\psi_2}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. For $c > 0$ and a scalar or vector $\boldsymbol{x}$, define $\mathcal{B}(\boldsymbol{x}, c) := \{\boldsymbol{x}' : \|\boldsymbol{x}' - \boldsymbol{x}\|_1 \le c\}$ as its ℓ_1 neighbor with radius c. Denote by $\boldsymbol{\Sigma}_0^{(\mathsf{m})} = \mathscr{P}^{(\mathsf{m})}\mathbf{X}\mathbf{X}^\mathsf{T}$, $\mathbb{H}_{\boldsymbol{\beta}}^{(\mathsf{m})} = \mathscr{P}^{(\mathsf{m})}\mathbf{X}_{\boldsymbol{\beta}}\mathbf{X}_{\boldsymbol{\beta}}^\mathsf{T}$, $\mathbb{J}_{\boldsymbol{\beta}}^{(\mathsf{m})} = \mathscr{P}^{(\mathsf{m})}\mathbf{X}\mathbf{X}^\mathsf{T}\{Y - \dot{\phi}(\mathbf{X}^\mathsf{T}\boldsymbol{\beta})\}^2$ and $\mathbb{U}_{\boldsymbol{\beta}}^{(\mathsf{m})} = \{\mathbb{H}_{\boldsymbol{\beta}}^{(\mathsf{m})}\}^{-1}$. For simplicity, let $\mathbb{H}_0^{(\mathsf{m})} = \mathbb{H}_{\boldsymbol{\beta}_0^{(\mathsf{m})}}^{(\mathsf{m})}$, $\mathbb{J}_0^{(\mathsf{m})} = \mathbb{J}_{\boldsymbol{\beta}_0^{(\mathsf{m})}}^{(\mathsf{m})}$ and denote by $\boldsymbol{u}_{0,j}^{(\mathsf{m})}$ the j^{th} row of $\mathbb{U}_{\boldsymbol{\beta}_0^{(\mathsf{m})}}^{(\mathsf{m})}$. In our following analysis, we assume that the cross–fitting folds $K', K = O(1)$, $n_m \asymp N/M \equiv n$ for all $m \in [M]$. Here and in the sequel we use $O(1)$ and $O_{\mathsf{P}}(1)$ denote order 1. Next, we introduce assumptions for our theoretical results. For Assumption 4, we only require either 4(a) or 4(b) to hold.

Assumption 1 (Regular covariance). *(i) There exists absolute constant $C_\Lambda > 0$ such that for all $m \in [M]$, $C_\Lambda^{-1} \le \Lambda_{\min}(\boldsymbol{\Sigma}_0^{(\mathsf{m})}) \le \Lambda_{\max}(\boldsymbol{\Sigma}_0^{(\mathsf{m})}) \le C_\Lambda$, $C_\Lambda^{-1} \le \Lambda_{\min}(\mathbb{H}_0^{(\mathsf{m})}) \le \Lambda_{\max}(\mathbb{H}_0^{(\mathsf{m})}) \le C_\Lambda$ and $C_\Lambda^{-1} \le \Lambda_{\min}(\mathbb{J}_0^{(\mathsf{m})}) \le \Lambda_{\max}(\mathbb{J}_0^{(\mathsf{m})}) \le C_\Lambda$. (ii) There exist $C_\Omega > 0$ and $\delta > 0$ that for all $m \in [M]$ and $\boldsymbol{\beta} \in \mathscr{B}(\boldsymbol{\beta}_0^{(\mathsf{m})}, \delta)$, ℓ_1 norm of each row of $\mathbb{U}_{\boldsymbol{\beta}}^{(\mathsf{m})}$ is bounded by C_Ω.*

Assumption 2 (Smooth link function)**.** *There exists a constant $C_L > 0$ such that for all $\theta, \theta' \in \mathbb{R}$, $|\ddot{\phi}(\theta) - \ddot{\phi}(\theta')| \le C_L|\theta - \theta'|$.*

Assumption 3 (Sub-Gaussian residual)**.** *For any $x \in \mathbb{R}^p$, $\epsilon_i^{(\mathsf{m})}$ is conditional sub-Gaussian, i.e. there exists $\kappa(x)$ such that $\|\epsilon_i^{(\mathsf{m})}\|_{\psi_2} < \kappa(x)$ given $\mathbf{X}_i^{(\mathsf{m})} = x$. In addition, there exists some absolute constant $C_\epsilon > 0$ such that, almost surely for $m = 1, 2 \ldots, M$, $\kappa(\mathbf{X}_i^{(\mathsf{m})}) \le C_\epsilon$ and $\ddot{\phi}^{-1}(\mathbf{X}_i^{(\mathsf{m})\mathsf{T}}\boldsymbol{\beta}_0^{(\mathsf{m})})\kappa^2(\mathbf{X}_i^{(\mathsf{m})}) \le C_\epsilon$.*

Assumption 4(a) (Sub-Gaussian design)**.** *$\mathbf{X}_i^{(\mathsf{m})}$ is sub-Gaussian, i.e. there exists some constant $\kappa > 0$ that $\|\mathbf{X}_i^{(\mathsf{m})}\|_{\psi_2} < \kappa$.*

Assumption 4(b) (Bounded design)**.** *$\|\mathbf{X}_i^{(\mathsf{m})}\|_\infty$ is almost surely bounded by some absolute constant.*

Remark 2. *Assumptions 1 (i) and 4(a) (or 4(b)) are commonly used technical conditions in high dimensional inference in order to guarantee rate optimality of the regularized regression and debiasing approach (Negahban et al., 2012; Javanmard and Montanari, 2014). Assumptions 4(a) and 4(b) are typically unified by the sub-Gaussian design assumption (Negahban et al., 2012). In our analysis, they are separately studied, since $\|\mathbf{X}_i^{(\mathsf{m})}\|_\infty$ affects the bias rate, which leads to different sparsity assumptions under different design types. Similar conditions as our Assumption 1 (ii) were used in the context of high dimensional precision matrix estimation (Cai et al., 2011) and debiased inference (Chernozhukov et al., 2018; Caner and Kock, 2018; Belloni et al., 2018). Compared with their exact or approximate sparsity assumption imposed on the inverse Hessian, this ℓ_1 boundness assumption is essentially less restrictive. As an important example in our analysis, logistics model satisfies the smoothness conditions for $\phi(\cdot)$ presented by Assumption 2. As used in Lounici et al. (2011) and Huang and Zhang (2010), Assumption 3 regularizes the tail behavior of the residuals and is satisfied in many common settings like logistic model.*

3.2 Asymptotic properties of the debiased estimator

We next study the asymptotic properties of the group effect statistics $\breve{\zeta}_j$, $j \in \mathcal{H}$. We shall begin with some important prerequisite results on the convergence properties of $\widetilde{\boldsymbol{\beta}}_{[\text{-}k]}^{(\bullet)}$ and the debiased estimators $\{\breve{\beta}_j^{(\mathsf{m})}, j \in \mathcal{H}, m \in [M]\}$ as detailed in Lemmas 1 and 2.

Lemma 1. *Under Assumptions 1-3, 4(a) or 4(b), and that $s = o\{n(\log p)^{-1}\}$, there exist a sequence of the tuning parameters*

$$\lambda_n^{(\mathsf{m})} \asymp \frac{(\log p)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \quad \text{and} \quad \lambda_N \asymp \frac{(M + \log p)^{\frac{1}{2}}}{n^{\frac{1}{2}} M} + \frac{sM^{-\frac{1}{2}}(\log p + \log N)^{a_0}\log p}{n},$$

with $a_0 = 1/2$ under Assumption 4(a) and $a_0 = 0$ under Assumption 4(b), such that, for each $k \in [K]$, the integrative estimator satisfies

$$\|\widetilde{\boldsymbol{\beta}}_{[\text{-}k]}^{(\bullet)} - \boldsymbol{\beta}_0^{(\bullet)}\|_{2,1} = O_{\mathsf{P}}(sM\lambda_N), \quad \text{and} \quad \|\widetilde{\boldsymbol{\beta}}_{[\text{-}k]}^{(\bullet)} - \boldsymbol{\beta}_0^{(\bullet)}\|_2^2 = O_{\mathsf{P}}(sM^2\lambda_N^2).$$

Remark 3. *Lemma 1 provides the estimation rates of the integrative estimator $\widetilde{\boldsymbol{\beta}}^{(\bullet)}_{[\text{-}k]}$. In contrast to the ILMA method, the second term in the expression of λ_N quantifies the additional noise incurred by using summary data under the DataSHIELD constraint. Similar results can be observed through debiasing truncation in distributed learning (Lee et al., 2017; Battey et al., 2018) or integrative estimation under DataSHIELD (Cai et al., 2019a). When $s = o\{n^{1/2}(\log p + \log N)^{-a_0}(M + \log p)^{-1}(\log p)^{-1/2}\}$ as assumed in Lemma 2, the above mentioned error term becomes negligible. The DSILT method allows for any degree of heterogeneity across sites with respect to both the magnitude and support of $\boldsymbol{\beta}^{(\mathsf{m})}_0$. However, the cross-site similarity in the support determines the estimation rates as shown in Lemma 1 above. Specifically, the DSILT estimator for $\boldsymbol{\beta}^{(\bullet)}$ attains a rate-M improvement over the local methods (Lounici et al., 2011; Huang and Zhang, 2010, e.g.) if $s \asymp s^{(\mathsf{m})}$ and has the same rate as that of the local estimators if $s \asymp \sum_{m=1}^{M} s^{(\mathsf{m})}$.*

We next present the theoretical properties of the group debiased estimators.

Lemma 2. *Under the same assumptions of Lemma 1 and assume that*

$$s = o\left\{\frac{n^{\frac{1}{2}}}{(\log p + \log N)^{a_0}(M + \log p)(\log p)^{\frac{1}{2}}} \wedge \frac{n}{M^4(\log p)^4(M + \log p)}\right\},$$

we have $\breve{\beta}^{(\mathsf{m})}_j - \beta^{(\mathsf{m})}_{0,j} = V^{(\mathsf{m})}_j + \Delta^{(\mathsf{m})}_j$ with $V^{(\mathsf{m})}_j = K^{-1}\sum_{k=1}^{K} \widehat{\mathscr{P}}_{\mathcal{I}^{(\mathsf{m})}_k} \boldsymbol{u}^{(\mathsf{m})\mathsf{T}}_{0,j}\mathbf{X}\epsilon$ converging to a normal random variable with mean 0 and variance $n_m^{-1}(\sigma^{(\mathsf{m})}_{0,j})^2$, where $(\sigma^{(\mathsf{m})}_{0,j})^2 = \boldsymbol{u}^{(\mathsf{m})\mathsf{T}}_{0,j}\mathbb{J}^{(\mathsf{m})}_0\boldsymbol{u}^{(\mathsf{m})}_{0,j}$. In addition, there exists $\tau \asymp (M + \log p)^{1/2}n^{-1/2}$ such that, simultaneously for all $j \in \mathcal{H}$, the bias term $\Delta^{(\mathsf{m})}_j$ and the variance estimator $(\widehat{\sigma}^{(\mathsf{m})}_j)^2$ satisfy that

$$|\Delta^{(\mathsf{m})}_j| \leq \sum_{m=1}^{M} |\Delta^{(\mathsf{m})}_j| = o_{\mathsf{P}}\left\{(n\log p)^{-\frac{1}{2}}\right\} \quad \textit{and} \quad \left|(\widehat{\sigma}^{(\mathsf{m})}_j)^2 - (\sigma^{(\mathsf{m})}_{0,j})^2\right| = o_{\mathsf{P}}\left\{(\log p)^{-1}\right\}.$$

Remark 4. *The sparsity assumption in Lemma 2 is weaker than the existing debiased estimators for M–estimation where s is only allowed to diverge in a rate dominated by $N^{\frac{1}{3}}$ (Janková and Van De Geer, 2016; Belloni et al., 2018; Caner and Kock, 2018). This is benefited by the cross–fitting technique, through which we can get rid of the dependence on the convergence rate of $\|\boldsymbol{u}^{(\mathsf{m})}_{0,j} - \widehat{\boldsymbol{u}}^{(\mathsf{m})}_{j,[k]}\|_1$.*

Finally, we establish in Theorem 1 the main result of this section regarding to the asymptotic distribution of the group test statistic $\breve{\zeta}_j$ under the null.

Theorem 1. *Under all assumptions in Lemma 2, simultaneously for all $j \in \mathcal{H}_0$, we have $\breve{\zeta}_j = S_j + o_{\mathsf{P}}(1)$, where $S_j = \sum_{m=1}^{M} n_m[V^{(\mathsf{m})}_j/\sigma^{(\mathsf{m})}_{0,j}]^2$. Furthermore, if $M \leq C\log p$ and $\log p = o(n^{1/C'})$ for some constants $C > 0$ and $C' > 6$, we have*

$$\sup_t |\mathsf{P}(S_j \leq t) - \mathsf{P}(\chi^2_M \leq t)| \to 0, \ \textit{as } n, p \to \infty.$$

The above theorem shows that, the group effect test statistics $\breve{\zeta}_j$ is asymptotically chi-squared distributed under the null and its bias is uniformly negligible for $j \in \mathcal{H}_0$.

3.3 False discovery control

We establish theoretical guarantees for the error rate control of the multiple testing procedure described in Section 2.5 in the following two theorems.

Theorem 2. *Assume that $q_0 = |\mathcal{H}_0| \asymp q$. Then under all assumptions in Lemma 2 with $\log p = o(n^{1/10})$ and $M = O(\log p)$, we have*

$$\limsup_{(N,p)\to\infty} \mathsf{FDR}(\hat{t}) \leq \alpha, \text{ and } \lim_{(N,p)\to\infty} P\{\mathsf{FDP}(\hat{t}) \leq \alpha + \epsilon\} = 1 \text{ for any } \epsilon > 0.$$

Remark 5. *Assumption 1 (i) ensures that most of the group estimates $\{\breve{\zeta}_j, j \in \mathcal{H}_0\}$ are not highly correlated with each other. Thus the the variance of $\widehat{R}_0(t)$ can be appropriately controlled, which in turn guarantees the control of FDP. It is possible to further relax the condition $\log p = o(n^{1/10})$ to $\log p = o(n^{\zeta})$ for some $0 < \zeta < 3/23$, See, for example, Liu and Shao (2014) and Belloni et al. (2018), where they used moderate deviation technique to have tighter truncations and normal approximations for t-statistics. Because we used chi-squared type test statistics with growing M, the technical details on moderate deviation are much more involved and warrant future research.*

As described in Section 2.5, if $\hat{t}$ in equation (5) is not attained in the range $[0, (2\log q - 2\log\log q)^{1/2}]$, then it is thresholded at $(2\log q)^{1/2}$. The following theorem states a weak condition to ensure the existence of $\hat{t}$ in such range. As a result, the FDP and FDR will converge to the pre-specified level α asymptotically.

Theorem 3. *Let $\mathcal{S}_\rho = \left\{j \in \mathcal{H} : \sum_{m=1}^{M} n_m [\beta_{0,j}^{(\mathsf{m})}]^2 \geq (\log q)^{1+\rho}\right\}$. Suppose for some $\rho > 0$ and some $\delta > 0$, $|\mathcal{S}_\rho| \geq \{1/(\pi^{1/2}\alpha) + \delta\}(\log q)^{1/2}$. Then under the same conditions as in Theorem 2, we have, as $(N, p) \to \infty$,*

$$\frac{\mathsf{FDR}(\hat{t})}{\alpha q_0/q} \to 1, \quad \frac{\mathsf{FDP}(\hat{t})}{\alpha q_0/q} \to 1 \text{ in probability.}$$

In the above theorem, the condition on $\mathcal{S}_\rho$ only requires very few covariates to have the signal sum of squares across the studies $\sum_{m=1}^{M}[\beta_{0,j}^{(\mathsf{m})}]^2$ exceeding the rate $(\log q)^{1+\rho}/n_m$ for some $\rho > 0$, and is thus a very mild assumption.

3.4 Comparison with alternative approaches

To study the advantage of our testing approach and the impact of the DataSHIELD constraint, we next compare the proposed DSILT method to a One–shot approach and the ILMA approach, as described in Algorithms 3 and 4, through a theoretical perspective. The One–shot approach in Algorithm 3 is inspired by existing literature in distributed learning (Lee et al., 2017; Battey et al., 2018, e.g.) , and is a natural extension of existing methods to the problem of multiple testing under the DataSHIELD constraint. The debiasing step of the One–shot approach is performed locally as in the existing literature.

Following similar proofs of Lemma 2 and Theorems 2 and 3, the One–shot, ILMA, and DSILT can attain the same error rate control results under the sparsity assumptions of

$$s = o(\gamma_1 \wedge \gamma_2), \quad \text{(One–shot)} \qquad \text{and} \qquad s = o\{(\gamma_1 M) \wedge \gamma_2\} \quad \text{(ILMA/DSILT)},$$

Algorithm 3 One–shot approach.

```
1. At each DC, obtain the cross–fitted debiased estimator by solving a Dantzig selector
   problem locally, where $\boldsymbol{\beta}^{(\mathsf{m})}$ is estimated by local LASSO.
2. Send the debiased estimators to the AC and obtain the group statistics.
3. Perform multiple testing procedure as described in Section 2.5.
```

Algorithm 4 Individual–level meta–analysis (ILMA).

```
1. Integrate all individual–level data at the AC.
2. Construct the cross–fitted debiased estimator by (4) using individual–level integrative
   estimator analog to (3), and then obtain the overall effect statistics.
3. Perform multiple testing procedure in Section 2.5.
```

where under the high dimensional regime of $\log n = O(\log p)$ and the assumptions of $M = O(\log p)$ and $\log p = o(n^{1/10})$ as required in Theorems 2 and 3,

$$\gamma_1 = \frac{n^{\frac{1}{2}}}{M(\log p + \log n)^{a_0}(\log p)^{\frac{3}{2}}} \asymp \frac{n^{\frac{1}{2}}}{M(\log p)^{a_0+\frac{3}{2}}}, \quad \gamma_2 = \frac{n}{M^4(\log p)^5},$$

and $a_0 = 1/2$ for sub-Gaussian design and $a_0 = 0$ for bounded design as in Lemma 1. If additionally $M = o\{n^{1/6}(\log p)^{a_0/3-7/6}\}$ which directly implies $\gamma_1 = o(\gamma_2)$, then the respective sparsity conditions for One–shot and ILMA/DSILT reduce to $s = o(\gamma_1)$ and $s = o\{(\gamma_1 M) \wedge \gamma_2\}$. Hence, when M grows with n and p at a slower rate of $M = o\{n^{1/6}(\log p)^{a_0/3-7/6}\}$, we have $\gamma_1 = o\{(\gamma_1 M) \wedge \gamma_2\}$, which implies that the ILMA and DSILT methods require strictly weaker sparsity assumption than the One–shot approach. On the other hand, if $M = o(n^{1/6}(\log p)^{a_0/3-7/6})$ is not satisfied, then the rate γ_2 dominates the rate of s and the three methods share the same sparsity condition $s = o(\gamma_2)$. Besides the sparsity condition comparisons in terms of the validity of tests, we learn from Cai et al. (2019a) that the estimation error rate of our integrative sparse regression in Step 1 is equivalent to the idealized method with all raw data and is smaller than the local estimator. Hence, we anticipate the power gain of the DSILT over the One–shot approach in finite-sample studies as the former uses more accurate estimator than the latter to derive statistics for debiasing. This advantage is also verified in our simulation studies in Section 4. Moreover, it is possible to follow the debiasing strategies proposed in Zhu et al. (2018) and Dukes and Vansteelandt (2019) that adapts to model sparsity, and construct a corresponding DSILT procedure with additional theoretical power gain compared with the One-shot method.

Remark 6. *Our DSILT approach involves transferring data twice from the DCs to the AC and once from the AC to the DCs, which requires more communication efforts compared to the One–shot approach. The additional communication gains lower bias rate than the*

One–shot approach while only requiring the same sparsity assumption as the ILMA method as discussed above. Under its sparsity condition, each method is able to draw inference that is asymptotically valid and has the same power as the ideal case when one uses the true parameters in construction of the group test statistics. This further implies that to construct a powerful and valid multiple testing procedure, there is no necessity to adopt further sequential communications between the DCs and the AC as in the distributed methods of Li et al. (2016) and Wang et al. (2017).

4. Simulation Study

We evaluate the empirical performance of the DSILT procedure and compare it with the One–shot and the ILMA methods. Throughout, we let $M = 5$, $n_m = 500$, and vary p from 500 to 1000. For each setting, we perform 200 replications and set the number of sample splitting folds $K = 2$, $K' = 5$ and false discovery level $\alpha = 0.1$. The tuning strategies described in Section 2.6 are employed with $H = 10$.

The covariate $\boldsymbol{X}$ of each study is generated from either the (i) Gaussian auto–regressive (AR) model of order 1 and correlation coefficient 0.5; or (ii) Hidden Markov model (HMM) with binary hidden variables and binary observed variables with the transition probability and the emission probability both set as 0.2. We choose $\{\boldsymbol{\beta}_0^{(\mathsf{m})}\}$ to be heterogeneous in magnitude across studies but to share the same support with

$$\boldsymbol{\beta}_0^{(\mathsf{m})} = \mu \left\{(\nu_1^{(\mathsf{m})}+1)\psi_1, (\nu_2^{(\mathsf{m})}+1)\psi_2, \ldots, (\nu_s^{(\mathsf{m})}+1)\psi_s, \mathbf{0}_{p-s}\right\}^{\mathsf{T}}$$

where the sparsity level s is set to be 10 or 50, and $\{\psi_1, ..., \psi_s\}$ are independently drawn from $\{-1, 1\}$ with equal probability and are shared across studies, while the local signal strength $\nu_j^{(\mathsf{m})}$'s vary across studies and are drawn independently from $\mathrm{N}\{0, (\mu/2)^2\}$. To ensure the procedures have reasonable power magnitudes for comparison, we set the overall signal strength μ to be in the range of $[0.21, 0.42]$ for $s = 10$, mimicking a sparse and strong signal setting; and $[0.14, 0.35]$ for $s = 50$, mimicking a dense and weak signal setting. We then generate binary responses $Y^{(\mathsf{m})}$ from $\mathrm{logit}P(Y^{(\mathsf{m})} = 1 \mid \mathbf{X}^{(\mathsf{m})}) = \boldsymbol{\beta}_0^{(\mathsf{m})\mathsf{T}}\mathbf{X}^{(\mathsf{m})}$.

In Figure 1, we report the empirical FDR and power of the three methods with varying p, s, and μ under the Gaussian design. Results for the HMM design have almost the same pattern and are included in the Supplementary Material. Across all settings, DSILT achieves almost the same performance as the ideal ILMA in both error rate control and power. All the methods successfully control the desired FDR at $\alpha = 0.1$. When $s = 10$ or the signal strength μ is weak, all the methods have conservative error rates compared to the nominal level. While for $s = 50$ with relatively strong signal, our method and the ideal ILMA become close to the exact error rate control empirically. This is consistent with Theorem 3 that if the number of relatively strong signals is large enough, our method tends to achieve exact FDR control. In contrast, the One–shot method fails to borrow information across the studies, and hence requires stronger signal magnitude to achieve exact FDR control. As a result, we observe consistently conservative empirical error rates for the One–shot approach.

In terms of the empirical power, the difference between DSILT and ILMA is less than 1% in all cases. This indicates that the proposed DSILT can accommodates the DataSHIELD constraint at almost no cost in power compared to ideal method. This is consistent with our theoretical result in Section 3.4 that the two methods require the same sparsity assumption for simultaneous inference. Furthermore, the DSILT and ILMA methods dominate the One–shot strategy in terms of statistical power. Under every single scenario, the power of the former two methods is around 15% higher than that of the One–shot approach in the dense case, i.e., $s = 50$, and 6% higher in the sparse case, i.e, $s = 10$. By developing testing procedures using integrative analysis rather than local estimations, both DSILT and ILMA methods use the group sparsity structure of the model parameters $\boldsymbol{\beta}^{(\bullet)}$ more adequately than the One–shot approach, which leads to the superior power performance of these two methods. The power advantage is more pronounced as the sparsity level s grows from 10 to 50. This is due to the fact that, to achieve the same result, the One–shot approach requires a stronger sparsity assumption than the other two methods, and is thus much more easily impacted by the growth of s. In comparison, the performance of our method and the ILMA method is less sensitive to sparsity growth because the integrative estimator employed in these two methods is more stable than the local estimator under the dense scenario.

5. Real Example

Statins are the most widely prescribed drug for lowering low–density lipoprotein (LDL) and the risk of cardiovascular disease (CVD), with over a quarter of adults 45 years or older receiving the drug in the United States. Statins lower LDL by inhibiting 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGCR) (Nissen et al., 2005). The treatment effect of statins can also be causally inferred based on the effect of the HMGCR variant *rs17238484* – patients carrying the *rs17238484*-G allele have profiles similar to individuals receiving statin, with lower LDL and lower risk of CVD (Swerdlow et al., 2015). While the benefit of statins have been consistently observed, they are not without risk. There has been increasing evidence that statins increase the risk of type II diabetes (T2D) (Rajpathak et al., 2009; Carter et al., 2013). Swerdlow et al. (2015) demonstrated via both meta analysis of clinical trials and genetic analysis of the *rs17238484* variant that statins are associated with a slight increase of T2D risk. However, the adverse effect of statins on T2D risk appears to differ substantially depending on the number of T2D risk factors patients have prior to receiving the statin, with adverse risk higher among patients with more risk factors (Waters et al., 2013).

To investigate potential genetic determinants of statin treatment effect heterogeneity, we studied interactive effects of the *rs17238484* variant and 256 SNPs associated with T2D, LDL, high–density lipoprotein (HDL) cholesterol, and the coronary artery disease (CAD) gene which plays a central role in obesity and insulin sensitivity (Kozak and Anunciado-Koza, 2009; Rodrigues et al., 2013). A significant interaction between SNP j and the statin variant *rs17238484* would indicate that SNP j modifies the effect of statin. Since the LDL, CAD and T2D risk profiles differ greatly between different racial groups and between male and female, we focus the analysis on the black sub-population and fit separate models for female and male subgroups.

To efficiently identify genetic risk factors that significantly interact with *rs17238484*, we performed an integrative analysis of data from 3 different studies, including the Million Vetern Project (MVP) from the Veteran Health Administration (Gaziano et al., 2016), Partners Healthcare Biobank (PHB) and the UK Biobank (UKB). Within each study, we have both a male subgroup indexed by subscript m, and a female subgroup indexed by subscript f, leading to $M = 6$ datasets denoted by $\mathrm{MVP}_{\mathsf{F}}, \mathrm{MVP}_{\mathsf{M}}, \mathrm{PHB}_{\mathsf{F}}, \mathrm{PHB}_{\mathsf{M}}, \mathrm{UKB}_{\mathsf{F}}$ and $\mathrm{UKB}_{\mathsf{M}}$. Since T2D prevalence within the datasets varies greatly from 0.05% to 0.15%, we performed a case control sampling with 1:1 matching so each dataset has equal numbers of T2D cases and controls. Since MVP has a substantially larger number of male T2D cases than all other studies, we down sampled its cases to match the number of female cases in MVP so that the signals are not dominated by the male population. This leads to sample sizes of 216, 392, 606, 822, 3120 and 3120 at $\mathrm{PHB}_{\mathsf{M}}$, $\mathrm{PHB}_{\mathsf{F}}$, $\mathrm{UKB}_{\mathsf{M}}$, $\mathrm{UKB}_{\mathsf{F}}$, $\mathrm{MVP}_{\mathsf{M}}$ and $\mathrm{MVP}_{\mathsf{F}}$, respectively. The covariate vector $\boldsymbol{X} = (\boldsymbol{X}_{\mathsf{main}}^{\mathsf{T}}, \boldsymbol{X}_{\mathsf{int}}^{\mathsf{T}})^{\mathsf{T}}$ is of dimension $p = 516$, where $\boldsymbol{X}_{\mathsf{main}}$ consists of the main effects of *rs17238484*, age and the aforementioned 256 SNPs, and $\boldsymbol{X}_{\mathsf{int}}$ consists of the interactions between *rs17238484* and age, as well as each of the 256 SNPs. All SNPs are encoded such that the higher value is associated with higher risk of T2D. We implemented the proposed testing method along with the One–shot approach as a benchmark to perform multiple testing of $q = 256$ coefficients corresponding to the interaction terms in $\boldsymbol{X}_{\mathsf{int}}$ at nominal level of $\alpha = 0.1$ with the model chosen as logistics regression and the sample splitting folds $K = 2$ and $K' = 5$.

As shown in Table 1, our method identifies 5 SNPs significantly interacting with the statin SNP while the One–shot approach detects only 3 SNPs, all of which belong to the set of SNPs identified by our method. The presence of non-zero interactive effects demonstrates that the adverse effect of statin SNP *rs17238484*-G on the risk of T2D can differ significantly among patients with different levels of genetic predisposition to T2D. In Figure 2, we also present 90% confidence intervals obtained within each dataset for the interactive effects between *rs17238484*-G and each of these 5 detected SNPs. The SNP *rs581080*-G in the TTC39B gene has the strongest interactive effect with the statin SNP and has all interactive effects estimated as positive for most studies, suggesting that the adverse effect of statin is generally higher for patients with this mutation compared to those without. Interestingly, a previous report finds that a SNP in the TTC39B gene is associated with statin induced response to LDL particle number (Chu et al., 2015), suggesting that the effect of statin can be modulated by the *rs581080*-G SNP.

Results shown in Figure 2 also suggest some gender differences in the interactive effects. For example, the adverse effect of the statin is lower for female patients carrying the *rs12328675*-T allele compare to female patients without the allele. On the other hand, the effect of the statin appear to be higher for male patients with the *rs12328675*-T allele compared to those without genetic variants associated with a various of phenotypes related to T2D. The variation in the effect sizes across different data sources illustrates that it is necessary to properly account for heterogeneity of $\boldsymbol{\beta}$ in the modeling procedure. Comparing the lengths of confidence intervals obtained based on the One–shot approach to those from the proposed method, we find that the DSILT approach generally yields shorter confidence intervals, which translates to higher power in signal detection. It is important to note that since MVP has much larger sample sizes, the width of the confidence intervals from MVP are much smaller than those of UKB and PHB. However, the effect sizes obtained from MVP also tend to be much smaller in magnitude and consequently, using MVP alone would only detect 2 of the 5 SNPs by multiple testing with level 0.1. This demonstrates the utility of the integrative testing involving $M = 6$ data sources.

6. Discussion

In this paper, we propose a DSILT method for simultaneous inference of high dimensional covariate effects in the presence of between-study heterogeneity under the DataSHIELD framework. The proposed method is able to properly control the FDR and FDP in theory asymptotically, and is shown to have similar performance as the ideal ILMA method and to outperform the One–shot approach in terms of the required assumptions and the statistical power for multiple testing. Our method allows most distributional properties of the data $\mathcal{D}^{(\mathsf{m})}$ to differ across the M sites, such as the marginal distribution of $\boldsymbol{X}^{(\mathsf{m})}$, the conditional variance of $\boldsymbol{Y}^{(\mathsf{m})}$ given $\boldsymbol{X}^{(\mathsf{m})}$, and the magnitude of each $\beta_j^{(\mathsf{m})}$. The support $\mathcal{S}^{(\mathsf{m})}$ is also allowed to vary across the sites as well, but the DSILT method is more powerful when $\mathcal{S}^{(1)}$, ..., $\mathcal{S}^{(\mathsf{M})}$ are more similar to each other. We demonstrate that the sparsity assumptions of the proposed method are equivalent to those for the ideal method but strictly weaker than those for the One–shot approach. As the price to pay, our method requires one more round of data transference between the AC and the DCs than the One–shot approach. Meanwhile, the sparsity condition equivalence between the proposed method and ILMA method implies that there is no need to include in our method further rounds of communications or adopt iterative procedures as in Li et al. (2016) and Wang et al. (2017), which saves a great deal of human effort in practice.

The proposed approach also adds technical contributions to existing literature in several aspects. First, our debiasing formulation helps to get rid of the group structure assumption on the covariates $\mathbf{X}^{(\mathsf{m})}$ at different distributed sites. Such an assumption is not satisfied in our real data setting, but is unavoidable if one uses the node-wise group LASSO (Mitra et al., 2016) or group structured inverse regression (Xia et al., 2018a) for debiasing. Second, compared with the existing work on joint testing of high dimensional linear models (Xia et al., 2018a), our method considers model heterogeneity and allows the number of studies M to diverge under the data sharing constraint, resulting in substantial technical difficulties in characterizing the asymptotic distribution of our proposed test statistics $\breve{\zeta}_j$ and their correlation structures for simultaneous inference.

We next discuss the limitation and possible extension of the current work. First, the proposed procedure requires transferring of Hessian matrix with $O(p^2)$ complexity from each DC to the AC. To the best of our knowledge, there is no natural way to reduce the order of complexity for the group debiasing step, i.e., Step 2, as introduced in Section 2.4. Nevertheless, it is worthwhile to remark that, for the integrative estimation step, i.e., Step 1, the communication complexity can be reduced to $O(p)$ only, by first transferring the locally debiased LASSO estimators from each DC to the AC and then integrating the debiased estimators with a group structured truncation procedure (Lee et al., 2017; Battey et al., 2018, e.g.) to obtain an integrative estimator with the same error rate as $\widetilde{\boldsymbol{\beta}}^{(\bullet)}_{[-k]}$. However, such a procedure requires greater efforts in deriving the data at each DC, which is not easily accomplished in some situations such as in our real example. Second, we assume $q = |\mathcal{H}| \asymp p$ in the current paper as we have $q = p/2$ in the real example of Section 5. We can further extend our results to the cases when q grows slower than p. In such scenarios, the error rate control results in Theorems 2 and 3 still hold. Meanwhile, the model sparsity assumptions and the conditions on p and N can be further relaxed because we have fewer number of hypotheses to test in total and as a result the error rate tolerance for an individual test $H_{0,j}$ can be weakened. Third, for the limiting null distribution of the test statistics $\breve{\zeta}_j$ and the subsequent simultaneous error rate control, we require $M = O(\log p)$ and $\log p = o(n^{1/10})$. Such an assumption is naturally satisfied in many situations as in our real example. However, when the collaboration is of a larger scale, say $M \gg \log p$ or $M > n_m$, developing an adaptive and powerful overall effect testing procedure (such as the ℓ_∞–type test statistics), particularly under DataSHIELD constraints, warrants future research. Fourth, the sub-Gaussian residual Assumption 3 in our theoretical analysis does not hold for Poisson or negatively binomial response. Inspired by existing work (Jia et al., 2019; Xie and Xiao, 2020, e.g.), our framework can be potentially generalized to accommodate more types of outcome models. Last, our method may be modified by perturbing the weighted covariates $\mathbf{X}^{(\mathsf{m})}_{\widehat{\boldsymbol{\beta}}}$ and response $\mathbf{Y}^{(\mathsf{m})}_{\widehat{\boldsymbol{\beta}}}$, and transferring the summary statistics derived from the perturbed data. Designing such a method with more convincing privacy guarantees, as well as similar estimation and testing performance as in our current framework warrants future research.

Acknowledgments

Part of this research is based on data from the Million Veteran Program (Gaziano et al., 2016), Office of Research and Development, Veterans Health Administration, and was supported by awards #MVP000 and #MVP001. This publication does not represent the views of the Department of Veterans Affairs or the United States Government. 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. We thank Zeling He and Nick Link for the help of running the real data, and Amanda King for the writing polish of the paper.

References

Mike Allen, John Bourhis, Nancy Burrell, and Edward Mabry. Comparing student satisfaction with distance education to traditional classrooms in higher education: A meta-analysis. *The American Journal of Distance Education*, 16(2):83–97, 2002.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, Ziwei Zhu, et al. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3): 1352–1382, 2018.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. High-dimensional econometrics and regularized gmm. *arXiv preprint arXiv:1806.01888*, 2018.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300, 1995.

Jonathan P Bradfield, H Rob Taal, Nicholas J Timpson, André Scherag, Cecile Lecoeur, Nicole M Warrington, Elina Hypponen, Claus Holst, Beatriz Valcarcel, Elisabeth Thiering, et al. A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature genetics*, 44(5):526, 2012.

Tianxi Cai, Molei Liu, and Yin Xia. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *arXiv preprint arXiv:1902.06115*, 2019a.

Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

TT Cai, H Li, J Ma, and Y Xia. Differential markov random field analysis with an application to detecting differential microbial community networks. *Biometrika*, 106(2):401–416, 2019b.

Mehmet Caner and Anders Bredahl Kock. High dimensional linear gmm. *arXiv preprint arXiv:1811.08779*, 2018.

David Card, Jochen Kluve, and Andrea Weber. Active labour market policy evaluations: A meta-analysis. *The economic journal*, 120(548):F452–F477, 2010.

Aleesa A Carter, Tara Gomes, Ximena Camacho, David N Juurlink, Baiju R Shah, and Muhammad M Mamdani. Risk of incident diabetes among patients treated with statins: population based study. *Bmj*, 346:f2610, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney K Newey. Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, 2016.

Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018.

Audrey Y Chu, Franco Giulianini, Bryan J Barratt, Bo Ding, Fredrik Nyberg, Samia Mora, Paul M Ridker, and Daniel I Chasman. Differential genetic effects on statin-induced changes across low-density lipoprotein–related measures. *Circulation: Cardiovascular Genetics*, 8(5):688–695, 2015.

REBECCA DerSimonian. Meta-analysis in the design and monitoring of clinical trials. *Statistics in medicine*, 15(12):1237–1248, 1996.

Dany Doiron, Paul Burton, Yannick Marcon, Amadou Gaye, Bruce HR Wolffenbuttel, Markus Perola, Ronald P Stolk, Luisa Foco, Cosetta Minelli, Melanie Waldenberger, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerging themes in epidemiology*, 10(1):12, 2013.

Oliver Dukes and Stijn Vansteelandt. Uniformly valid confidence intervals for conditional treatment effects in misspecified high-dimensional models. *arXiv preprint arXiv:1903.10199*, 2019.

Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

Andre Franke, Dermot PB McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, et al. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118, 2010.

Amadou Gaye, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, Andrew Turner, Elinor M Jones, Joel Minion, Andrew W Boyd, Christopher J Newby, Marja-Liisa Nuotio, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *International journal of epidemiology*, 43(6):1929–1944, 2014.

John Michael Gaziano, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, Jennifer Deen, Colleen Shannon, Donald Humphries, et al. Million veteran program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*, 70:214–223, 2016.

Richard S Houlston, Emily Webb, Peter Broderick, Alan M Pittman, Maria Chiara Di Bernardo, Steven Lubbe, Ian Chandler, Jayaram Vijayakrishnan, Kate Sullivan, Steven Penegar, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet*, pages 1426–35, 2008.

Junzhou Huang and Tong Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

Jana Janková and Sara Van De Geer. Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv preprint arXiv:1610.01353*, 2016.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Adel Javanmard, Hamid Javadi, et al. False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253, 2019.

Jinzhu Jia, Fang Xie, Lihu Xu, et al. Sparse poisson regression with penalized weighted score function. *Electronic Journal of Statistics*, 13(2):2898–2920, 2019.

EM Jones, NA Sheehan, N Masca, SE Wallace, MJ Murtagh, and PR Burton. DataSHIELD–shared individual-level analysis without sharing the data: a biostatistical perspective. *Norsk epidemiologi*, 21(2), 2012.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 526(114):668–681, 2019.

LP Kozak and R Anunciado-Koza. Ucp1: its involvement and utility in obesity. *International journal of obesity*, 32(S7):S32, 2009.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Wenfa Li, Hongzhe Liu, Peng Yang, and Wei Xie. Supporting regularized logistic regression privately and efficiently. *PloS one*, 11(6):e0156479, 2016.

W.D. Liu and S. Luo. Hypothesis testing for high-dimensional regression models. *Technical report*, 2014.

Weidong Liu and Qi-Man Shao. Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics*, 42(5):2003–2025, 2014.

Karim Lounici, Massimiliano Pontil, Sara Van De Geer, Alexandre B Tsybakov, et al. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

Rong Ma, T Tony Cai, and Hongzhe Li. Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, pages 1–15, 2020.

Ritwik Mitra, Cun-Hui Zhang, et al. The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics*, 10(2):1829–1873, 2016.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Steven E Nissen, E Murat Tuzcu, Paul Schoenhagen, Tim Crowe, William J Sasiela, John Tsai, John Orazem, Raymond D Magorien, Charles O'Shaughnessy, and Peter Ganz. Statin therapy, ldl cholesterol, c-reactive protein, and coronary artery disease. *New England Journal of Medicine*, 352(1):29–38, 2005.

Orestis A Panagiotou, Cristen J Willer, Joel N Hirschhorn, and John PA Ioannidis. The power of meta-analysis in genome-wide association studies. *Annual review of genomics and human genetics*, 14:441–465, 2013.

Swapnil N Rajpathak, Dharam J Kumbhani, Jill Crandall, Nir Barzilai, Michael Alderman, and Paul M Ridker. Statin therapy and risk of developing type 2 diabetes: a meta-analysis. *Diabetes care*, 32(10):1924–1929, 2009.

Alice Cristina Rodrigues, B Sobrino, Fabiana Dalla Vecchia Genvigir, Maria Alice Vieira Willrich, Simone Sorkin Arazi, Egidio Lima Dorea, Marcia Martins Silveira Bernik, Marcelo Bertolami, André Arpad Faludi, MJ Brion, et al. Genetic variants in genes related to lipid metabolism and atherosclerosis, dyslipidemia and atorvastatin response. *Clinica Chimica Acta*, 417:8–11, 2013.

Gavin Stewart. Meta-analysis in applied ecology. *Biology letters*, 6(1):78–81, 2010.

Daniel I Swerdlow, David Preiss, Karoline B Kuchenbaecker, Michael V Holmes, Jorgen EL Engmann, Tina Shah, Reecha Sofat, Stefan Stender, Paul CD Johnson, Robert A Scott, et al. Hmg-coenzyme a reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *The Lancet*, 385(9965):351–361, 2015.

Lu Tang, Ling Zhou, and Peter X-K Song. Method of divide-and-combine in regularized generalized linear models for big data. *arXiv preprint arXiv:1611.06208*, 2016.

Jiayi Tong, Rui Duan, Ruowang Li, Martijn J Scheuemie, Jason H Moore, and Yong Chen. Robust-odal: Learning from heterogeneous health systems without sharing patient-level data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 25, page 695. World Scientific, 2020.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3636–3645. JMLR. org, 2017.

David D Waters, Jennifer E Ho, S Matthijs Boekholdt, David A DeMicco, John JP Kastelein, Michael Messig, Andrei Breazna, and Terje R Pedersen. Cardiovascular event reduction versus new-onset diabetes during atorvastatin therapy: effect of baseline risk factors for diabetes. *Journal of the American College of Cardiology*, 61(2):148–152, 2013.

Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, et al. DataSHIELD: resolving a conflict in contemporary bioscience–performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5):1372–1382, 2010.

Yin Xia, T Tony Cai, and Hongzhe Li. Joint testing and false discovery rate control in high-dimensional multivariate regression. *Biometrika*, 105(2):249–269, 2018a.

Yin Xia, Tianxi Cai, and T Tony Cai. Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Statistica Sinica*, 28:63, 2018b.

Fang Xie and Zhijie Xiao. Consistency of ℓ_1 penalized negative binomial regressions. *Statistics & Probability Letters*, page 108816, 2020.

Eleftheria Zeggini, Laura J Scott, Richa Saxena, Benjamin F Voight, Jonathan L Marchini, Tianle Hu, Paul IW de Bakker, Gonçalo R Abecasis, Peter Almgren, Gitte Andersen, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5):638, 2008.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Yinchu Zhu, Jelena Bradic, et al. Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12(2):3312–3364, 2018.

Daniela Zöller, Stefan Lenz, and Harald Binder. Distributed multivariable modeling for signature development under data protection constraints. *arXiv preprint arXiv:1803.00422*, 2018.

Figure 1: The empirical FDR and power of our DSILT method, the One–shot approach and the ILMA method under the Gaussian design, with $\alpha = 0.1$. The horizontal axis represents the overall signal magnitude μ.

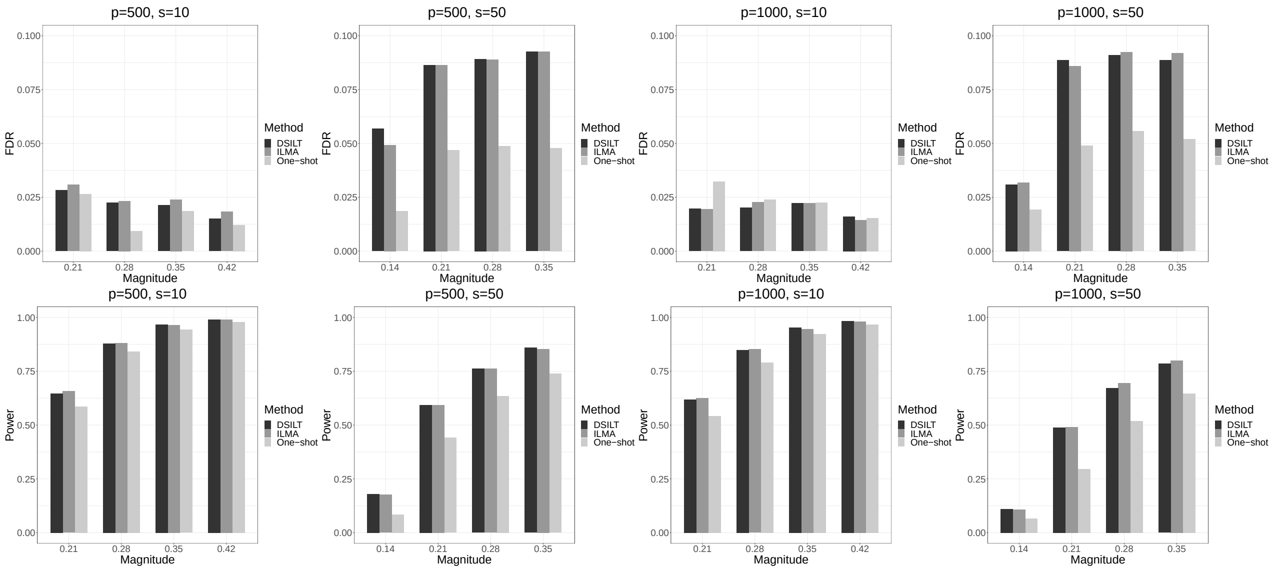

Figure 2: Debiased estimates of the log odds ratios and their 90% confidence intervals in each local site for the interaction effects between *rs17238484*-G and the 5 SNPs detected by DSILT, obtained respectively based on the One–shot and the DSILT approaches.

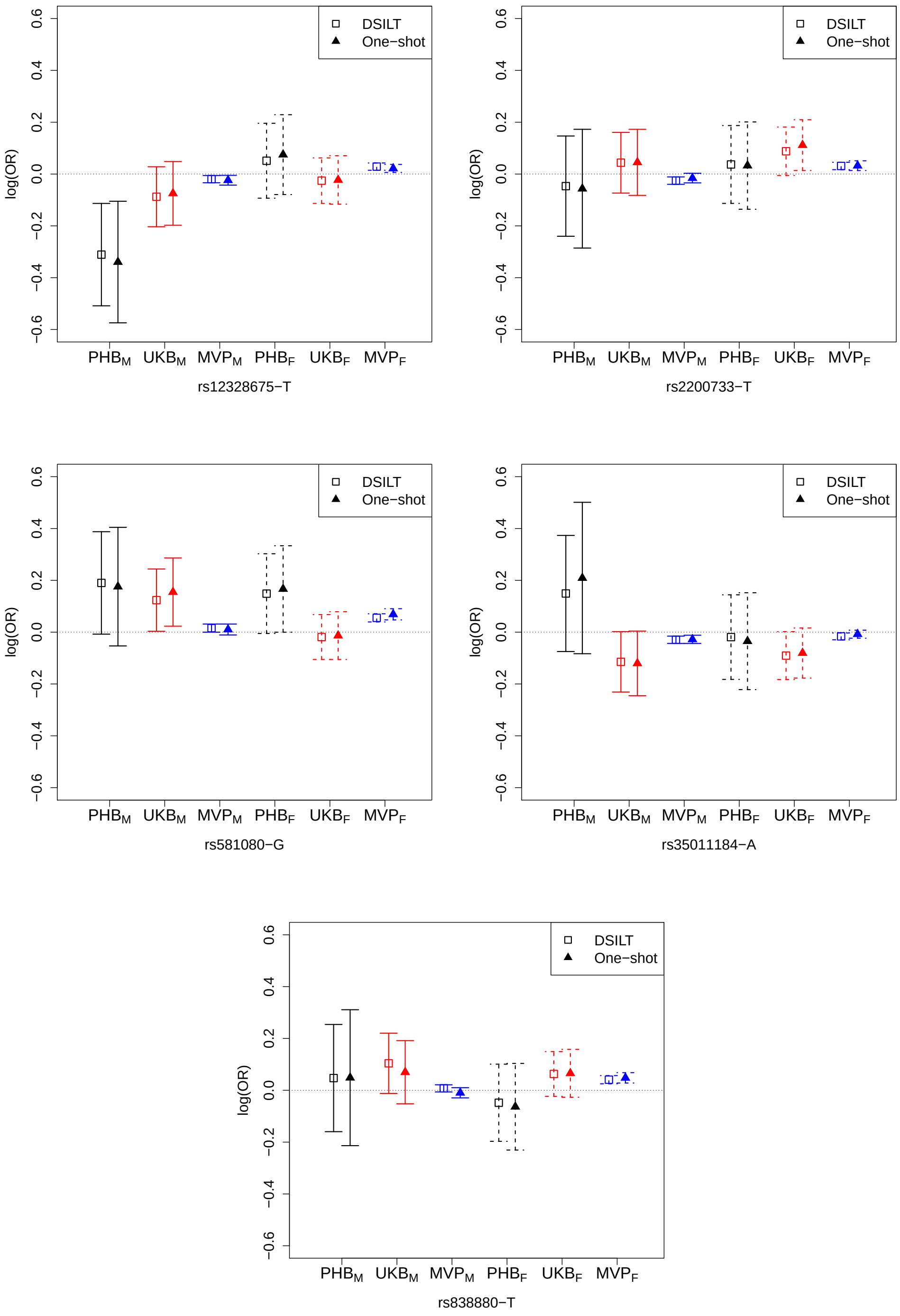

Table 1: SNPs identified by DSILT to interact with the statin genetic variants *rs17238484*-G on the risk for T2D. The second column presents the name of the gene where the SNP locates. The third column presents the minor allele frequency (MAF) of each SNP averaged over the three sites. The last three columns respectively present the p–values obtained using One–shot approach with all the $M = 6$ studies, One–shot with solely the datasets MVP$_f$ and MVP$_m$ and the proposed method with all the $M = 6$ studies. The p–values shown in black fonts represent the SNPs selected by each method.

SNP	Gene	MAF	One–shot	MVP–only	DSILT
rs12328675-T	COBLL1	0.13	$\mathbf{1.1 \times 10^{-3}}$	2.3×10^{-3}	$\mathbf{6.0 \times 10^{-4}}$
rs2200733-T	LOC729065	0.18	3.7×10^{-2}	5.7×10^{-3}	$\mathbf{6.2 \times 10^{-4}}$
rs581080-G	TTC39B	0.22	$\mathbf{3.6 \times 10^{-6}}$	$\mathbf{1.1 \times 10^{-6}}$	$\mathbf{2.6 \times 10^{-6}}$
rs35011184-A	TCF7L2	0.22	1.9×10^{-2}	5.2×10^{-2}	$\mathbf{8.6 \times 10^{-4}}$
rs838880-T	SCARB1	0.36	$\mathbf{6.7 \times 10^{-4}}$	$\mathbf{6.0 \times 10^{-5}}$	$\mathbf{6.2 \times 10^{-4}}$