Some Statistical Consideration in Transcriptome-Wide Association Studies

Haoran Xue[1], Wei Pan[2], and for the Alzheimer's Disease Neuroimaging Initiative[3]

[1]School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455.
[2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455.
Email: panxx014@umn.edu. Phone: 612-624-4655. Fax: 612-626-0660.

July 25, 2019; revised October 1, 2019

[3]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List_Sep23.pdf`.

Abstract

Transcriptome-wide association study (TWAS) has become popular in integrating a reference eQTL dataset with an independent main GWAS dataset to identify (putatively) causal genes, shedding mechanistic insights to biological pathways from genetic variants to a GWAS trait mediated by gene expression. Statistically TWAS is a (two-sample) 2-stage least squares (2SLS) method in the framework of instrumental variables analysis for causal inference: in Stage 1 it uses the reference eQTL data to impute a gene's expression for the main GWAS data, then in Stage 2 it tests for association between the imputed gene expression and the GWAS trait; if an association is detected in Stage 2, a (putatively) causal relationship between the gene and the GWAS trait is claimed. If a non-linear model or a generalized linear model (GLM) is fitted in Stage 2 (e.g. for a binary GWAS trait), it is known that using only imputed gene expression, as in standard TWAS, in general does not lead to a consistent (i.e. asymptotically unbiased) estimate for the causal effect; accordingly, a variation of 2SLS, called two-stage residual inclusion (2SRI), has been proposed to yield better estimates (e.g. being consistent under suitable conditions). Our main goal is to investigate whether it is necessary or even better to apply 2SRI, instead of the standard 2SLS. In addition, due to the use of imputed gene expression (i.e. with measurement errors), it is known that in general some correction to the standard error estimate of the causal effect estimate has to be applied, while in the standard TWAS no correction is applied. Is this an issue? We also compare one-sample 2SLS with two-sample 2SLS (i.e. the standard TWAS). We used the ADNI data and simulated data mimicking the ADNI data to address the above questions. At the end, we conclude that, in practice with the large sample sizes and small effect sizes of genetic variants, the standard TWAS performs well and is recommended.

Keywords 2SLS; 2SPS; 2SRI; Causal inference; Instrumental variables; Mendelian randomization; TWAS.

1 Introduction

Genome-wide association studies (GWAS) have been successful in identifying thousands of trait-associated genetic variants, mostly single nucleotide polymorphisms (SNPs). However, since most of the identified trait-associated SNPs are in the non-coding region of the genome, there is a lack of mechanistic understanding of how these SNPs influence the traits. It is hypothesized that many genetic variants influence complex traits through transcriptional regulation (He et al., 2013), which can be used to identify *causal* genes. For this purpose, PrediXcan (Gamazon et al., 2015) and transcription-wide association study (TWAS) (Gusev et al., 2016), simply called **TWAS** from now on, were recently proposed to uncover *putatively* causal genes by integrating a main GWAS dataset with a reference gene expression or expression quantitative trait (eQTL) dataset. TWAS has since become popular and successful in applications to common diseases like T2D and cancer, and to complex traits like BMI, lipids and height, convincingly showing the power of integrating GWAS and eQTL data to gain biological insights. Statistically, TWAS applies the (two-sample) two-stage least squares (2SLS) method for causal inference (Xu, Wu, Wei & Pan, 2017a), closely related to Mendelian randomization (MR) (Zhao, Wang, Hemani, Bowden, & Small, 2019). Since TWAS is a gene-based method by testing genes one by one, for the purpose of presentation we can consider only one gene. In Stage 1, one builds a prediction model for the genetic component of the gene's expression level, called "genetically regulated expression (GReX)", by using only cis-acting genotypes around the gene based on a reference eQTL dataset. In stage 2, for a given separate main GWAS dataset, based on the genotype of each subject, we can "impute" his/her gene expression (i.e. GReX) using the predictive model built in Stage 1. Then we test the association between the imputed gene expression and the GWAS trait. If there is an association, then, under suitable modeling assumptions (Xu, Wu, Wei, & Pan, 2017; Mancuso et al., 2019; Hu et al., 2019; Wainberg et al., 2019), it is claimed that the gene is (putatively) causal to the trait: some causal SNPs affect the trait through the mediating effects of the gene's expression.

As to be reviewed next, if the GWAS trait is quantitative, a linear model is usually used in Stage 2 to test the association between the imputed gene expression level and the trait, which is exactly 2SLS. However, in practice, a non-linear model (with an additive error term) may be used; or, more often, since in many GWAS the trait is not quantitative, e.g. being binary as an indicator of a disease status, a generalized linear model (GLM), e.g. a logistic regression model, is instead fitted. For such a non-linear model in Stage 2, it is known that the usual 2SLS, specifically called Two-Stage Predictor Substitution (2SPS), may not be consistent; instead, for a non-linear model with an additive error term, Two-Stage Residual Inclusion (2SRI) is consistent and should be applied (Terza, Basu, & Rathouz, 2008; MacKenzie, Tosteson, Morden, Stukel & O'Malley, 2014), while for a logistic regression model, an equivalent method to 2SRI has been proposed (Palmer et al., 2008, 2011). Hence, an important question is whether it is indeed suitable to apply 2SPS for binary traits in Stage 2 as in the current practice of TWAS. From now on, we use 2SLS as a *generic* term covering both 2SPS and 2SRI. In addition, TWAS is based on a two-sample 2SLS (2S-2SLS), in which two separate and independent datasets are used in Stages 1 and 2 respectively; there is a corresponding one-sample 2SLS (1S-2SLS) with the two datasets in the two stages collected from the same set of subjects (Angrist & Krueger, 1991). This is convenient with two separate eQTL and GWAS datasets. However, in practice, there are situations when both eQTL and GWAS are collected on the same set (or largely overlapping sets) of subjects. In these situations, should be split the dataset into two

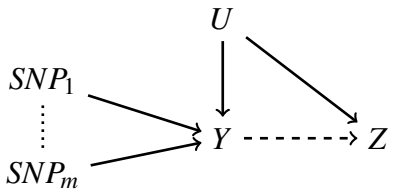
non-overlapping subsets before applying 2S-2SLS, or apply 1S-2SLS to the whole dataset? To answer this question, we need to investigate how 1S-2SLS performs in the context of TWAS. We will use both a real dataset and simulated data to address these questions.

2 Methods

2.1 The True Models in Instrumental Variables Analysis

The true (causal) model is illustrated with a directed acyclic graph (DAG) in Figure 1, where Y, Z and U are the gene, GWAS trait and unobserved confounders respectively, and a directed edge solid between U and Z, U and Y, SNP_j and Y (for $j = 1,...,m$) represents a direct causal effect. Our goal is to test whether Y has a direct effect on Z, and possibly estimate its effect size.

Figure 1: True model.



For individual i, $i = 1,...,n$, with gene expression level Y_i, SNPs $SNP_{1,i}$, ... ,$SNP_{m,i}$, and binary trait Z_i. Let p_i be the probability of $Z_i = 1$. Our the true model for Stage 1 is:

$$Y_i = \beta_0 + \beta_1 \cdot SNP_{1,i} + ... + \beta_m \cdot SNP_{m,i} + U_i, \tag{1}$$

where, as usual, throughout, we use additive coding for each SNP: $SNP_{j,i} = 0, 1$ or 2 for $j = 1,...,m$ and $i = 1,...,n$. The true model for Stage 2 is:

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 \cdot Y_i + \alpha_2 \cdot U_i \tag{2}$$

with $p_i = \Pr(Y_i = 1)$. A main challenge is that we do not observe the confounder U_i that is correlated with both X_i and Y_i.

The above true model is the typical one adopted in the instrumental variables (IVs) analysis for causal inference, in which the SNPs are taken as IVs. The focus is statistical inference on the causal effect α_1. A main benefit is consistent estimation and inference for α_1 at the expense of three assumptions with IVs, as in MR: 1) the SNPs/IVs are associated with Y; 2) the SNPs/IVs are not associated with U; 3) Conditional on Y, the SNPs/IVs are not associated with Z. If any of the above three modeling assumptions is violated, biased inference for α_1 results, as discussed in the context of TWAS (Mancuso et al., 2019; Wainberg et al., 2019). Since it is not the focus of this paper while it is quite challenging to deal with, we will assume that the three assumptions hold in the following as in standard scenarios.

2.2 One-sample versus Two-sample Approaches

All methods we are going to introduce consist of two stages. If we have both eQTL and GWAS data from the same sample of subjects, we have two choice of how to use the data. First, we can use the whole sample for both Stages 1 and 2, which is denoted as **one-sample** approach. Alternatively, we can randomly split the whole sample into half-half (or in whatever desired ratio), using the first half for Stage 1, and the other half for Stage 2, which is denoted as **two-sample** approach.

Denote $I_1, I_2 \subseteq \{1, \cdots, n\}$ be the index sets for samples used in Stages 1 and 2 respectively. For the one-sample strategy, $I_1 = I_2 = \{1, \cdots, n\}$; for the two-sample approach, we have $I_1 \cap I_2 = \emptyset$.

2.3 TWAS: Four Methods to Implement It

We implement TWAS in four different ways in Stage 2, depending on whether a LM or GLM is fitted to the binary GWAS trait and how to use imputed gene expression, leading to four different methods, denoted as GLM 2SPS, GLM 2SRI, LM 2SPS and LM 2SRI. The (standard) TWAS corresponds to LM 2SPS. All of these four methods are variants of 2SLS, consisting of two stages; they share fitting the same LM in Stage 1, but differ in fitting different models in Stage 2.

In Stage 1, we fit a LM by regressing gene expression Y on the SNPs, SNP_1, ... , SNP_m:

$$Y_i = \beta_0 + \beta_1 \cdot SNP_{1,i} + ... + \beta_m \cdot SNP_{m,i} + \varepsilon_i, \qquad i \in I_1, \tag{3}$$

where ε_i is assumed to be a random noise with mean 0 and independent of the SNPs. We obtain the OLS estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m$, and predict (or impute) gene expression $\hat{Y}_i$'s for $i \in I_2$. We also estimate $\hat{U}_i = Y_i - \hat{Y}_i$ for confounders U_i, $i \in I_2$.

In Stage 2, for GLM-2SPS, we fit a logistic regression model using $\hat{Y}_i$:

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 \cdot \hat{Y}_i. \tag{4}$$

In contrast, for GLM-2SRI, we fit a logistic model using both Y_i and $\hat{U}_i$:

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 \cdot Y_i + \alpha_2 \cdot \hat{U}_i. \tag{5}$$

Note the use of Y_i, not $\hat{Y}_i$, in the above model, which will not be possible for usual two-sample scenarios.

As alternatives, we can fit two LMs in Stage 2. The first corresponds to LM-2SPS:

$$Z_i = \alpha_0 + \alpha_1 \cdot \hat{Y}_i + e_i, \tag{6}$$

where e_i is assumed to be a random noise with mean 0 and independent of the SNPs. The second is for LM-2SRI:

$$Z_i = \alpha_0 + \alpha_1 \cdot Y_i + \alpha_2 \cdot \hat{U}_i + e_i. \tag{7}$$

For each method, we draw inference on the causal effect α_1 after obtaining its estimate $\hat{\alpha}_1$ and standard error $\text{se}(\hat{\alpha}_1)$. For GLM-2SRI and LM-2SPS, there are existing methods for correct $\text{se}(\hat{\alpha}_1)$ as to be discussed next, while for others we will use the output from standard software fitting their corresponding models in Stage 2.

It is noted that, although we have one-sample ADNI data with the availability of the genotype, gene expression and GWAS trait (AD) data for each subject, to mimic a two-sample approach, we split the whole sample of the ADNI data into two non-overlapping samples. An advantage of the two-sample approach is the relaxed assumption of the availability of the gene expression and the GWAS trait in the two samples respectively, to which the two-sample LM- or GLM-2SPS method (as the standard TWAS) can be applied; however, the two-sample LM- or GLM-2SRI does require the availability of the gene expression data in both samples (but only the availability of the GWAS trait in the second sample).

Given the true data generating models (1) and (2), and the working models (3) and (4), although the estimation of the causal effect α_1 in (2) is not consistent in general, the test for whether $\alpha_1 = 0$ is valid and consistent (Dai & Zhang, 2015). Furthermore, we note that GLM-2SRI with model (5) is equivalent to the so called "Adjusted IV Estimator" (Palmer, Thompson, Tobin, Sheehan & Burton, 2008), also called "Control Function Estimator" (Palmer et al., 2011) for binary traits, in which Y_i is replaced by $\hat{Y}_i$ in (5).

2.4 Corrections for se($\hat{\alpha}_1$)

For one-sample GLM 2SRI, we correct the $se(\hat{\alpha}_1)$ with following equation (Terza, 2018):

$$V_c(\hat{\alpha}) = V(\hat{\alpha}) + V(\hat{\alpha})AV(\hat{\beta})A'V(\hat{\alpha}). \tag{8}$$

Here $V(\hat{\beta})$ is the original covariance matrix of $\hat{\beta}$'s in the first stage, and $V(\hat{\alpha})$ is the original covariance matrix of $\hat{\alpha}$'s in the second stage; throughout, the "original" one means the usual covariance matrix directly output from fitting a standard model in Stage 1 or 2. $V_c(\hat{\alpha})$ is the corrected covariance matrix for $\hat{\alpha}$. We define:

$$f^*(Z|Y,SNPs;\alpha,\beta) = f(Z|Y,U = Y - SNPs\cdot\beta;\alpha) = \frac{[exp(\alpha_0+\alpha_1\cdot Y+\alpha_2\cdot U)]^Z}{1+exp(\alpha_0+\alpha_1\cdot Y+\alpha_2\cdot U)}, \tag{9}$$

and let $f_i^* = f(Z_i|Y_i,U_i = Y_i - SNPs_i\cdot\beta;\alpha)$ and $A_i^{(\alpha,\beta)}$:

$$A_i^{(\alpha,\beta)} = \nabla_\alpha log(f_i^*)\cdot\nabla_\beta log(f_i^*)'. \tag{10}$$

Then A is calculated as

$$A = \sum_{i=1}^{n} A_i^{(\hat{\alpha},\hat{\beta})}. \tag{11}$$

We can notice that the corrected $se_c(\hat{\alpha}_1)$ is larger than the original $se(\hat{\alpha}_1)$. Though the correction was originally designed for one-sample GLM-2SRI, we will also apply it to two-sample GLM-2SRI.

For one-sample LM-2SPS, the usual standard error estimate requires the homoskedasticity assumption on the error term in a LM; to allow heterskedastic errors, a robust or sandwich-type standard error estimator has been proposed (Baiocchi, Cheng, & Small, 2014; Imbens & Angrist, 1994; Angrist & Pischke, 2009). We can use the `robust.se()` function in R package **ivpack** to obtain the robust standard error of $\hat{\alpha}_1$ in LM 2SPS (Baiocchi, Cheng, & Small, 2014).

For two-sample LM-2SPS, to account for the statistical uncertainty or estimation error of estimating/imputing Y_i as $\hat{Y}_i$ in Stage 1, we can correct the usual $se(\hat{\alpha}_1)$ by inflating it with a

factor no smaller than 1 (Inoue, Atsushi & Solon, 2010):

$$V_c(\hat{\alpha}) = V(\hat{\alpha})(1 + \hat{\alpha}_1^2 \cdot \frac{n_2}{n_1} \cdot \frac{\hat{\sigma_1}^2}{\hat{\sigma_2}^2}), \tag{12}$$

where $V(\hat{\alpha})$ is the original covariance matrix of $\hat{\alpha}$, n_1 and n_2 are the sample sizes in Stages 1 and 2 respectively, and $\hat{\sigma_1}^2$ and $\hat{\sigma_2}^2$ are the estimated variances for the error terms in the two LMs in the two stages respectively.

We thus obtain the corrected standard errors (SEs) for both one-sample and two-sample LM-2SPS, and for both one-sample and two-sample GLM-2SRI. For other methods, we are not aware of any existing methods to correct their SEs and thus just use the standard/uncorrected SEs from the output from fitting their corresponding models in Stage 2.

2.5 ADNI Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a 60 million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California — San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see `www.adni-info.org`.

For real data analysis and for generating simulated data, we used the data from Alzheimer's Disease Neuroimaging Initiative (ADNI) (Shen et al., 2014), including its gene expression, whole genome sequencing (WGS) and trait data. After cleaning and merging, we had a sample size 712. To mimic a case-control study, we treated 247 Cognitively Normal (CN) individual as controls, and the remaining 465 individuals with Alzheimer's Disease (AD) or Mild Cognitive Impairment (MCI) as cases. The expression levels of 17256 genes on the autosomes were used. For each gene, we defined its cis-region by expanding 100kb upstream and downstream its coding region (i.e. from its TSS and TES) respectively. We excluded *SNP*s with MAF ≤ 0.05 or with missing values. We also pruned the SNPs to ensure that any of their pairwise Pearson correlations in absolute values was no more than 0.9. Finally, if there were still more than 30 *SNP*s in the cis-region of the gene, we chose and only kept the top 30 SNPs (as IVs)

with the largest absolute values of the correlations with the gene's expression level; if there were less than or equal to 30 *SNP*s, we kept all of them as IVs. In this way, we had $m \leq 30$ *SNP*s as IVs so that the ordinary least squares (OLS) estimation could be applied in Stage 1.

For some genes, if their expression levels are not associated with their cis-SNPs, then using their cis-SNPs to predict their expression levels in Stage 1 would violate the first IV assumption, leading to the use of invalid IVs. As alternatives, for each gene we tested the association in equation (1) with the null hypothesis H_0: $\beta_1 = \beta_2 = ... = \beta_m = 0$. Under H_0 (and the normality assumption for Y), the coefficient of determination R^2 follows an F-distribution with degrees of freedom $(m, n-m-1)$. Thus, we performed the F-test on each gene in Stage 1, and only retained the genes with p-values less than some cut-off (0.05 or 0.1) before applying the methods.

2.6 Simulation Set-ups

To further study the methods, we conducted simulation studies by using the ADNI data to mimic realistic scenarios. We randomly selected gene **PSPH** on chromosome 7 to study. We first fitted a LM in the first stage with 30 *SNP*s we selected:

$$Y_i = \beta_0 + \beta_1 \cdot SNP_{1,i} + ... + \beta_{30} \cdot SNP_{30,i} + U_i, \tag{13}$$

for $i = 1, ..., 712$, to obtain $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{30}$, $\hat{Y}$, $\hat{U}$ and $var(\hat{U})$. In the second stage, we first fitted a model like GLM 2SRI:

$$\text{logit}(p_i) = \alpha_0^{(1)} + \alpha_1^{(1)} \cdot Y_i + \alpha_2^{(1)} \cdot \hat{U}_i, \tag{14}$$

for $i = 1, ..., 712$, to obtain $\hat{\alpha_0}^{(1)}, \hat{\alpha_1}^{(1)}, \hat{\alpha_2}^{(1)}$ for a non-null model (with a causal effect $\hat{\alpha_1}^{(1)} \neq 0$. For a null model, we fitted a logistic regression model with only $\hat{U}$:

$$\text{logit}(p_i) = \alpha_0^{(2)} + \alpha_2^{(2)} \cdot \hat{U}_i \tag{15}$$

to obtain $\hat{\alpha_0}^{(2)}$ and $\hat{\alpha_2}^{(2)}$. In the following simulation set-ups, we used these estimated coefficients to generate simulated data.

In each simulation set-up, we generated simulated data in the following steps.
For $i = 1, ..., R \cdot 712$:

1. Generate U_i's as i.i.d. Normal(0, $var(\hat{U})$);
2. Generate $Y_i = \hat{\beta}_0 + \hat{\beta}_5 \cdot SNP_{5,i} + \hat{\beta}_{15} \cdot SNP_{15,i} + \hat{\beta}_{25} \cdot SNP_{25,i} + U_i$;
3. Generate $\text{logit}(p_i)$. If $S_Y = 0$, then $\text{logit}(p_i) = \hat{\alpha_0}^{(2)} + S_U \cdot \hat{\alpha_2}^{(2)} \cdot U_i$; otherwise, $\text{logit}(p_i) = \hat{\alpha_0}^{(1)} + S_Y \cdot \hat{\alpha_1}^{(1)} \cdot Y_i + S_U \cdot \hat{\alpha_2}^{(1)} \cdot U_i$;
4. Generate $Z_i \sim \text{Bernoulli}(p_i)$.

Here R controlled the sample size, S_Y the effect size of the gene expression Y, and S_U the effect size of the confounder U. The values of the *SNP*s were drawn from the ADNI data: for $i = k + l \cdot 712$, $k = 1, ...712$, and $l = 1, ..., R-1$, we defined $SNP_{j,i} = SNP_{j,k}$, meaning that

we replicated the *SNP* data R times to possibly increase the sample size. We tried various combinations of (R, S_Y, S_U):

$$\{(R, S_Y, S_U) | R = 1,3,5,7,9; S_Y = 0,1,2,3,4,5; S_U = 1,10,20,30,40,50\}.$$

Note that we chose 3 *SNP*s as the true IVs: SNP_5, SNP_{15}, and SNP_{25}. though we would use all 30 SNPs To mimic the real situation with valid IVs unknown, we would select 10 top SNPs as IVs from the 30 candidate SNPs as for the real data analysis. Then we fitted the LM in the first stage, and applied the four methods from formulas (4), (5), (6), and (7) in the second stage. Because we knew both Y and U with simulated data, we also considered the ideal (but not practical) *Oracle* method by fitting the true model (2) in the second stage.

For each simulated dataset, we applied the methods with both one-sample and two-sample approaches; for the former, we split a simulated dataset into half/half for the two stages respectively. For each simulation set-up, we generated 1000 independent simulated datasets.

2.7 Data Availability

The ADNI data are available to the approved user on the project web site `http://adni.loni.usc.edu/`. Some sample R code and data are available at `https://github.com/xue-hr/twas_methods`.

3 Results

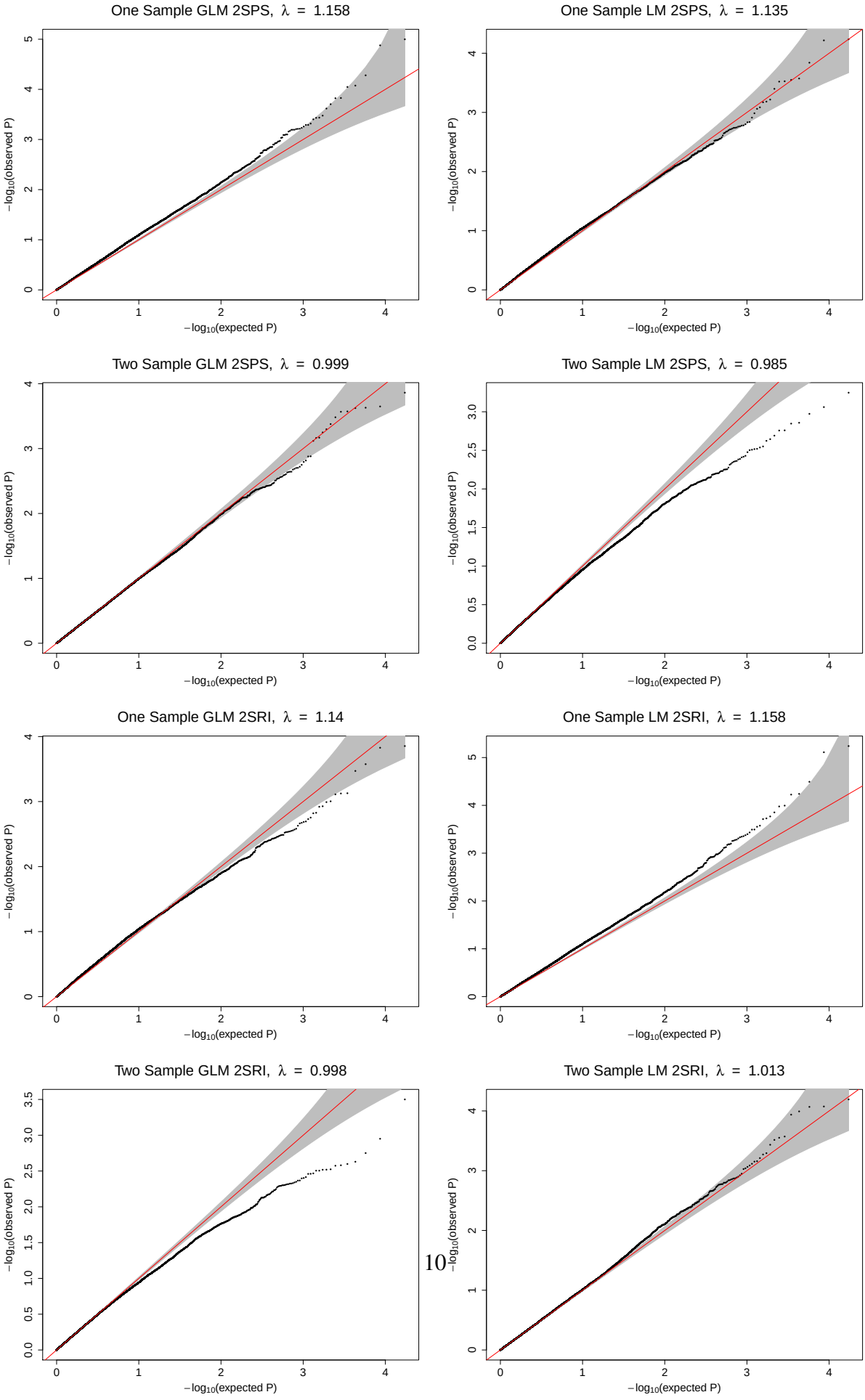
3.1 ADNI Data Analysis

We applied the methods to 17256 genes for 712 individuals in ADNI data. For each of the four models, we tried both one-sample and two-sample approaches, obtaining the p-values for α_1 for the 17256 genes, and draw Q-Q plots for these 17256 p-values in Figure 2.

First, it is clear that all four methods with the one-sample approach led to inflated type I errors with inflated genomic control factors $\hat{\lambda}$ (Devlin & Roeder, 1999), likely due to not accounting for SNP selection in Stage 1 with the same sample as that in Stage 2. Among the four methods, 1S-GLM-2SPS and 1S-LM-2SRI was most liberal with much inflated type I errors, while 1S-GLM-2SRI was conservative at the left tail of the p-value distribution (i.e. perhaps over-estimating the more significant/smaller p-values); in contrast, 1S-LM-2SPS performed almost ok. Second, while all four methods with the two-sample strategy did not yield inflated $\hat{\lambda}$, 2S-LM-2SPS and 2S-GLM-2SRI were too conservative, especially at the left tail of the p-value distribution; in contrast, 2S-GLM-2SPS performed almost ideally, followed by 2S-LM-2SRI.

We also conducted an F-test for a possible association between each gene's expression levels and its cis-SNPs in Stage 1; we only retained the 9102 (or 10564) genes with a p-value $<$ 0.05 (or 0.1) before applying the methods. The resulting Q-Q plots (Supplementary Materials) show the same patterns as discussed above; in particular, the inflation of the Type I error rates by the one-sample approaches was even more evident.

Figure 2: The ADNI data analysis: Q-Q plots of the obtained p-values of 17256 genes from each method versus the expected p-values under the null hypothesis of no association.

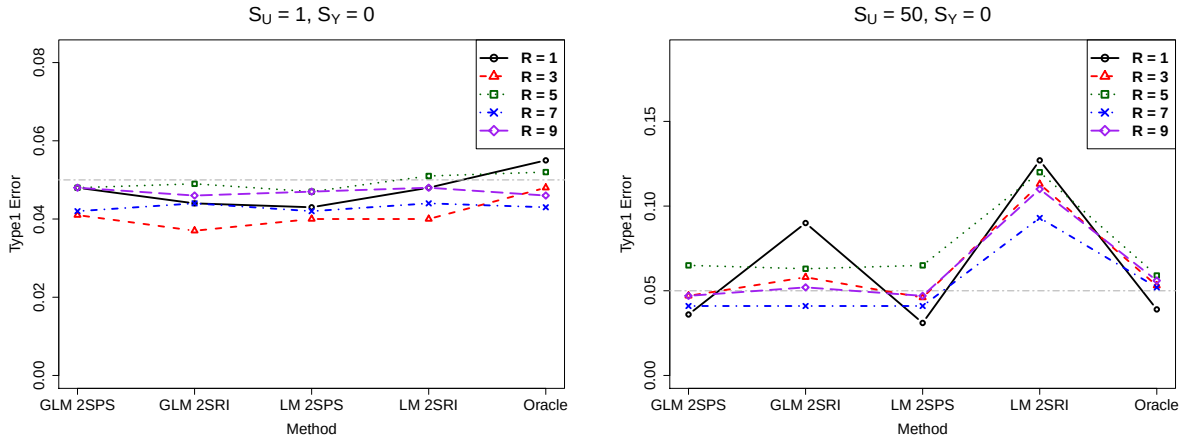


3.2 Simulation Results: Two-sample Approaches

3.2.1 Type I errors

For the null case with no causal effects (i.e. $\alpha_1 = 0$ or $S_Y = 0$), the empirical Type I error rates at the nominal significance level of 0.05 for the four methods and the Oracle are shown in Figure 3 based on 1000 simulations; for comparison, the nominal significance level is marked with a gray horizontal line (in each corresponding figure). If the confounding is not severe with $S_U = 1$, all the methods performed satisfactorily. However, with more severe confounding with $S_U = 50$, first, LM-2SRI consistently had inflated Type I error rates, while GLM-2SRI also had an inflated type I error for the small sample size (with $R = 1$), which however disappeared with increasing sample sizes. In contrast, GLM-2SPS and LM-2SPS always controlled their Type I error rates satisfactorily.

Figure 3: Simulations with the two-sample approaches: empirical Type I error rates of various methods.



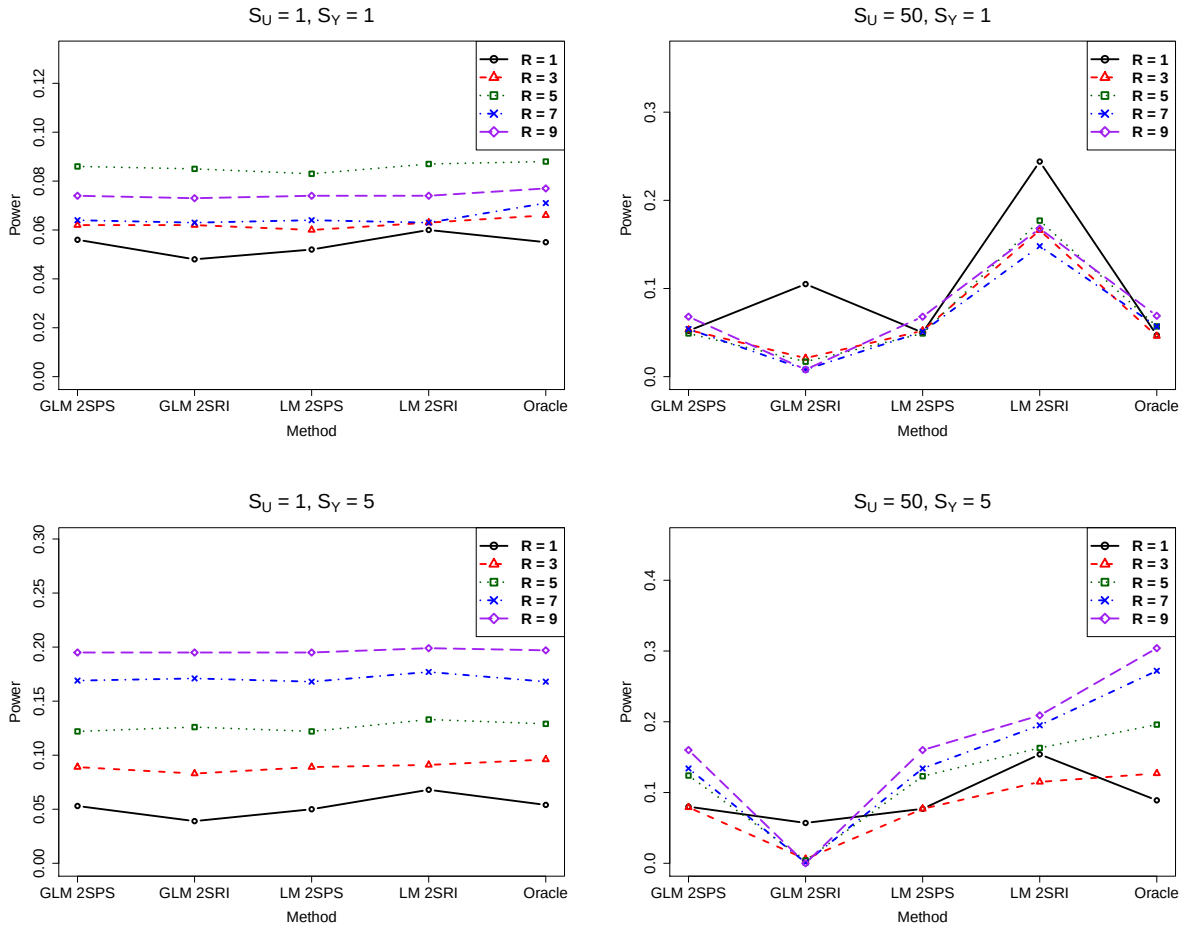
3.2.2 Power

In the presence of causal effects, as shown in Figure 4, with small confounding (with $S_U = 1$), all the methods performed similarly. However, with severe confounding (with $S_U = 50$), GLM-2SRI was low-powered, while GLM-2SPS and LM-2SPS performed similarly. Note that the high power of LM-2SRI was likely due to its inflated Type I error rates.

3.2.3 Biases

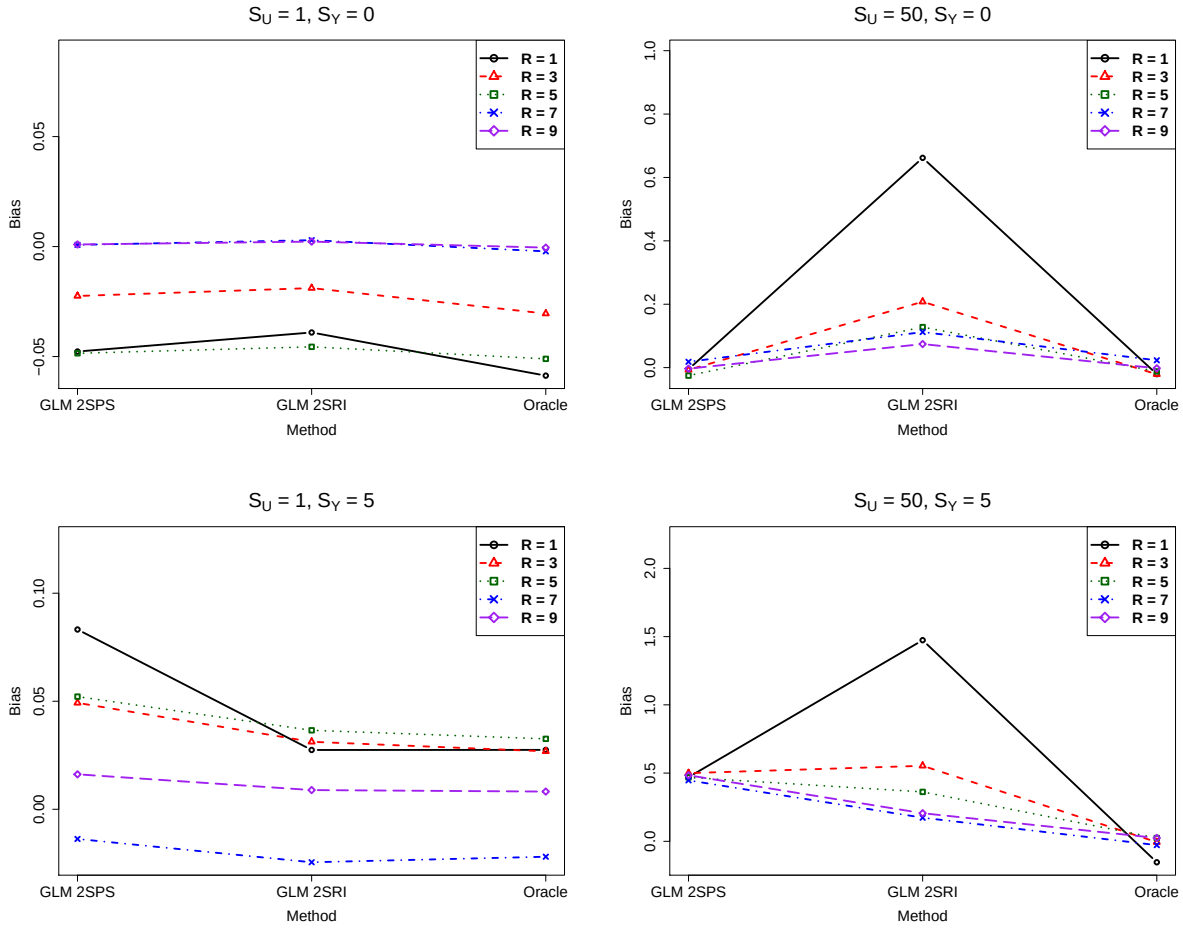
We could compare the true causal effect size α_1 with its estimate $\hat{\alpha}_1$ from each method to assess the extent of the bias, if any, by each method. Since the true model was a GLM in Stage 2, we only compared the GLM-based methods. For a specific simulation setting, based on 1000 simulations, we estimated the bias as $\sum_{i=1}^{1000}(\hat{\alpha}_1^i - \alpha_1)/1000$, where $\hat{\alpha}_1^i$ was the estimate of the true α_1 in the i^{th} simulation. As shown in Figure 5, first, in the null case (i.e. $S_Y = 0$), with small confounding (i.e. small S_U), the three methods performed similarly. However, with a large S_U, at a small sample size GLM-2SPS and GLM-Oracle, but not GLM-2SRI, performed

Figure 4: Simulations with the two-sample approaches: empirical power of various methods.



well; as the sample size increases, GLM-2SRI became less biased. Second, in the non-null case (i.e. $S_Y \neq 0$), GLM-2SRI was less biased than GLM-2SPS in most cases, but not for $S_U = 50$, $S_Y = 5$ and $R = 1$ (i.e. a small sample size), in which GLM-2SRI was more biased than GLM-2SPS.

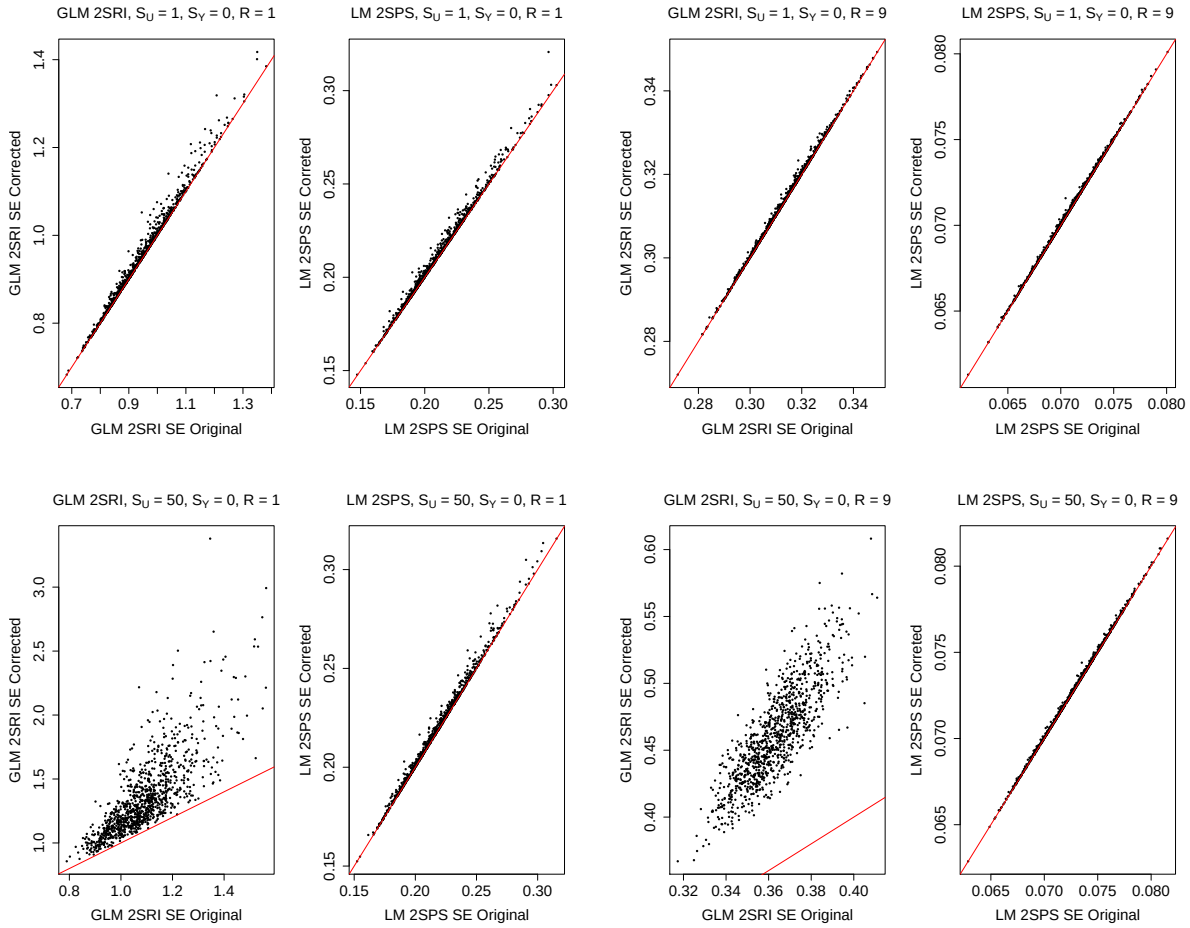
Figure 5: Simulations with the two-sample approaches: biases of various methods.



3.2.4 Comparison of the original and corrected standard error estimates

For GLM-2SRI and LM-2SPS, we compared their original and corrected standard error estimates. Figure 6 shows that, first, with a small S_U, the corrected SEs were slightly larger than the original SEs, but the difference became much smaller as the sample size R increased. On the other hand, under more severe confounding with a larger S_U, the corrected SEs were always much larger than the original one for GLM-2SRI, regardless of the sample size; in comparison, the differences between the two for LM-2SPS were much smaller, and tended to disappear as the sample size increased, across all the simulation settings (also see Supplementary figures). These results suggest that it is perhaps necessary to use the corrected SE for GLM-2SRI, but not so for LM-2SPS (especially for large sample sizes).

Figure 6: Simulations with the two-sample approaches: comparison of the original and corrected standard error estimates.



3.2.5 Results For Other Settings

We have only shown some representative results for the simulation set-ups with $S_Y = 0$ or 5, and $S_U = 1$ or 50. The results for other set-ups are shown in the Supplementary Materials.

3.3 Simulation Results: One-sample Approaches

We reached similar conclusions on the relative performance of the four methods for the one-sample case as for the two-sample case shown earlier, except that GLM-2SRI performed better and were more in par with GLM-2SPR and LM-2SPR. Again LM-2SRI yielded inflated Type I errors. The detailed results are shown in the Supplementary Materials. It is noted that the three one-sample approaches performed well was presumably due to our "weak" selection of SNPs: we only selected 10 SNPs from 30 candidate SNPs as IVs; if more candidate SNPs were included while more or less SNPs were allowed to be selected as IVs, the effects of the selection bias would be larger as shown for the real data analysis.

4 Conclusions and Discussion

In summary, presumably due to selection bias, one-sample approaches may lead to inflated type I errors and thus are not recommended. Among the two-sample approaches, GLM-2SPS, followed by LM-2SPS, performed best; note that these two methods are used as default in practice for TWAS. Two-sample GLM-2SRI performed well if the sample size was large enough; otherwise it could be conservative, and even with large biases. In contrast, two-sample LM-2SRI did not perform well across all simulations, and should not be used.

Why did 2SPS perform well, even better than 2SRI for non-linear logistic regression model in Stage 2 in our study? Does this contradict the general theory of 2SRI? A quick answer to the second question is no. In retrospect, the reason for the first question is simple: it is related to the currently well accepted practice of applying a linear model to a binary trait in GWAS, because a linear model can approximate well the corresponding logistic (or other non-linear) regression model due to the small effect sizes of SNPs (Zhao, Wang, Hemani, Bowden & Small, 2019). Furthermore, in 2SRI, with the high correlation between the observed gene expression (Y_i in our notation) and the residual/estimated confounding ($\hat{U}_i = Y_i - \hat{Y}_i$) due to the often low predictivity of a gene's expression level by its cis-SNPs, fitting the Stage-2 model in 2SRI requires a larger sample size for it to perform well. Finally, it is not feasible to even apply 2SRI with two separate samples of eQTL and GWAS data, because of no observed gene expression levels (Y_i's) in the typical GWAS data.

We also note that in practice of using TWAS, the statistical uncertainty (i.e. estimation error) in imputing gene expression in Stage 1 is ignored. Although this uncertainty can be taken account using the corrected SE estimator, our numerical study suggested its negligible effects. Hence again the usual practice with TWAS of no correction appears to be fine.

We emphasize that our main conclusion (that the standard TWAS performs well) holds only under the conditions with the large sample size and small effect sizes of genetic variants on complex traits and common diseases. Otherwise, for example, in extensions of TWAS to molecular traits or other endophenotypes (Xu, Wu, Wei, Pan & Alzheimer's Disease Neuroimaging Initiative, 2017b; Wu et al., 2018), on which genetic variants (or other IVs) may have

much larger effect sizes, cautions should be taken: 2SPS as adopted in the standard TWAS may not be even consistent for a non-linear model in Stage 2.

There are other limitations with the current study, including the following two important and challenging issues. First, instead of OLS estimation, penalized regression methods, such as Lasso or elastic net (Zou & Hastie, 2005), or Bayesian methods, are often used in the first stage for TWAS in practice (Gamazon et al., 2015; Gusev et al., 2016). The benefits include selecting relevant SNPs as valid IVs, avoiding biases of weak IVs, and obtaining better estimates to impute gene expression better, which presumably would lead to better inference (e.g. more precise estimates and higher power) for the causal parameter in the second stage. Some large sample properties, e.g. square-root-n consistency, of such Lasso or post-Lasso procedures, have been established for sparse models under suitable conditions (Belloni et al., 2012). However, for finite (especially small) samples, these methods may *not* yield imputed gene expression leve;s orthogonal to or uncorrelated with the confounders, leading to possibly biased inference on the causal parameter with 2SPS in Stage 2. An alternative approach is jackknife instrumental variable estimation, which predicts/imputes each observation i by fitting a Stage 1 model with other observations after excluding observation i, but it is computationally intensive (Angrist, Imbens, & Krueger, 1999; Hansen & Kozbur, 2014). Furthermore, given the relatively large sample size in TWAS, it is not clear how biased the causal parameter inference would be if no correction is applied. Second, in our simulations, we only considered "weak selection" of IVs: we selected 10 SNPs as IVs from 30 candidate ones containing 3 causal SNPs (i.e. 3 valid IVs), which was expected to select at least one valid IV with a high probability while having a relatively small selection bias. The more difficult cases with few or no valid IVs, or more practically as in TWAS with a larger set of candidate SNPs/IVs, may render it necessary to use penalized regression or other more sophisticated methods in Stage 1, introducing some challenges as discussed earlier. More work is needed.

Supporting Information

In the Supplementary Materials, we provide more analysis results for the ADNI data (Figures S1-S2), simulations for the two-sample approaches (Figures S3-S44), and simulations for the one-sample approaches (Figures S45–S86).

Acknowledgements

We thank the reviewers for many helpful comments and suggestions. WP would like to thank Dr. Todd MacKenzie for first introducing 2SRI to him. This work was supported by NIH grants R21AG057038, R01HL116720, R01GM113250, R01GM126002 and R01HL105397, and by the Minnesota Supercomputing Institute at the University of Minnesota.

References

[1] Angrist, J. D., & Krueger, A. B.(1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics, 106*(4), 979-1014. https://doi.org/10.2307/2937954

[2] Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics, 14*(1), 57-67.

[3] Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics : An empiricist's companion*. Princeton: Princeton University Press.

[4] Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine, 33*(13), 2297-2340. https://doi.org/10.1002/sim.6128

[5] Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica, 80*, 2369-2429.

[6] Dai, J. Y., & Zhang, X. C. (2015). Mendelian Randomization Studies for a Continuous Exposure Under Case-Control Sampling. *American Journal of Epidemiology, 181*—(6), 440-449. https://doi.org/10.1093/aje/kwu291

[7] Devlin, B., & Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics, 55*(4), 997-1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x

[8] Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., . . . Im., H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics, 47*(9), 1091-8. https://doi.org/10.1038/ng.3367

[9] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., . . . Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics, 48*(3), 245-24552. https://doi.org/10.1038/ng.3506

[10] Hansen, C., & Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics, 182*(2), 290-308. https://doi.org/10.1016/j.jeconom.2014.04.022

[11] He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., & Li, H. (2013). Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. *The American Journal of Human Genetics, 92*(5), 667-680. https://doi.org/10.1016/j.ajhg.2013.03.022

[12] Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., . . . Zhao, H. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics, 51*(3), 568-576. https://doi.org/10.1038/s41588-019-0345-7

[13] Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica, 62*(2), 467-475. http://doi.org/10.2307/2951620

[14] Inoue, Atsushi, & Solon, Gary. (2010). Two-Sample Instrumental Variables Estimators. The Review of Economics and Statistics, 92(3), 557-561.

[15] Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., & Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics, 51*(4), 675-2,682A-682B. https://doi.org/10.1038/s41588-019-0367-1

[16] MacKenzie, T., Tosteson, A., Morden, T., Stukel, D., & O'Malley, N. (2014). Using instrumental variables to estimate a Cox?s proportional hazards regression subject to additive confounding. Health Services and Outcomes Research Methodology, 14(1-2), 54-68.

[17] Palmer, T. M., Thompson, J. R., Tobin, M. D., Sheehan, N. A., & Burton, P. R.(2008). Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *International Journal of Epidemiology, 37*(5), 1161-1168. https://doi.org/10.1093/ije/dyn080

[18] Palmer, T. M., Sterne, J. A. C., Harbord, R. M., Lawlor, D. A., Sheehan, N. A., Meng, S., . . . Didelez, V. (2011). Instrumental Variable Estimation of Causal Risk Ratios and Causal Odds Ratios in Mendelian Randomization Analyses. *American Journal of Epidemiology, 173*(12), 1392-1403. https://doi.org/10.1093/aje/kwr026

[19] Shen, L., Thompson, P. M., Potkin, S. G., Bertram, L., Farrer, L. A., Foroud, T. M., . . . Saykin, A. J., & Alzheimer's Disease Neuroimaging Initiative. (2014). Genetic analysis of quantitative phenotypes in AD and MCI: Imaging, cognition and biomarkers. *Brain Imaging and Behavior, 8*(2), 183-207. https://doi.org/10.1007/s11682-013-9262-z

[20] Terza, J. V., Basu, A., & Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics, 27*(3), 531-543. https://doi.org/10.1016/j.jhealeco.2007.09.009

[21] Terza, J. V.(2018). Two-Stage Residual Inclusion Estimation in Health Services Research and Health Economics. *Health Services Research, 53*(3), 1890-1899. https://doi.org/10.1111/1475-6773.12714

[22] Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., . . . Kundaje, A. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics, 51*(4), 592-599. https://doi.org/10.1038/s41588-019-0385-z

[23] Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., . . . Yang, J.(2018). Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Communications, 9*(1), 918. https://doi.org/10.1038/s41467-018-03371-0

[24] Xu, Z., Wu, C., Wei, P., & Pan, W. (2017a). A Powerful Framework for Integrating eQTL and GWAS Summary Data. *Genetics, 207*(3), 893-902. https://doi.org/10.1534/genetics.117.300270

[25] Xu, Z., Wu, C., Pan, W., & Alzheimer's Disease Neuroimaging Initiative. (2017b). Imaging-wide association study: Integrating imaging endophenotypes in GWAS. *NeuroImage, 159*, 159-169. https://doi.org/10.1016/j.neuroimage.2017.07.036

[26] Zhao, Q., Wang, J., Hemani, G., Bowden, J., Small, D. S. (2019). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. arXiv:1801.09652.

[27] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society Series B-Statistical Methodology, 67*, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x