Statistica Sinica Preprint No: SS-2021-0164	
Title	Mendelian Randomization Test of Causal Effect Using High-Dimensional Summary Data
Manuscript ID	SS-2021-0164
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0164
Complete List of Authors	Lu Deng, William Wheeler and Kai Yu
Corresponding Author	Kai Yu
E-mail	yuka@mail.nih.gov

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

MENDELIAN RANDOMIZATION TEST OF CAUSAL EFFECT USING HIGH-DIMENSIONAL SUMMARY DATA

Lu Deng[1], William Wheeler[2] and Kai Yu[3*]

[1]*School of Statistics and Data Science, Nankai University, Tianjin, P. R. China*

[2]*Information Management Services, Silver Spring, MD, U.S.A.*

[3]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, U.S.A.*

[*]*corresponding author yuka@mail.nih.gov*

Abstract: Mendelian randomization (MR) uses genetic variants as instrumental variables (IVs) to assess the causal effect of a risk factor on an outcome in the presence of unmeasured confounding. There is growing interest in conducting MR analyses using summary statistics on each IV's association with the risk factor and the outcome, which are generated from large-scale genome-wide association studies (GWAS). Most existing approaches use summary data on a set of IVs that have been established as being associated with the risk factor. They often have limited power because the set of identified IVs jointly explain only a small proportion of the variation in the measure of the risk factor. We propose a new MR testing procedure that takes full advantage of summary data on tens of thousands of genetic variants studied by GWAS. The test statistic is the maximum of a sequence of modified K-statistics defined by a range of thresholds. Compared with existing approaches, this new test gains power by collecting signals from many undetected IVs throughout the genome, and is robust to both balanced and unbalanced pleiotropy. We investigate the theoretic properties of the proposed procedure and demonstrate its advantages over existing ones using simulation studies and a real example.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Key words and phrases: Genome-wide association studies, Instrumental variables, Mendelian randomization, Pleiotropic effect, Summary statistics.

1. Introduction

Mendelian randomization (MR) analysis uses genetic variants as instrumental variables (IVs) to estimate the causal effect of a risk factor on an outcome based on observational studies (Lawlor et al., 2008; Smith and Ebrahim, 2003). MR is becoming an effective tool for studying the causal relationship between a risk exposure and a disease outcome, because many robust findings on the genetic basis underlying various common traits have been accumulated through large-scale genome-wide association studies (GWAS) over the past decade.

Most MR analyses are conducted under the two-sample setting by combining summary data from two separate GWAS, with one study providing summary statistics on the association between each IV (typically, a single nucleotide polymorphism, called an SNP) and the risk factor, and the other giving summary statistics on the IV and the outcome association (Burgess et al., 2015; Hemani et al., 2018). Because a two-sample MR analysis uses only SNP association summary statistics, its input data are easy to obtain and assemble. With more summary data becoming publicly available, two-sample MR studies are becoming more popular in the analysis of GWAS.

A typical MR analysis relies on identifying a set of IVs that are expected to meet the following three conditions (Didelez and Sheehan, 2007): 1. Relevance: each IV is associated with the risk factor X; 2. Exclusion restriction: each IV affects the outcome Y only through its influence on X; 3. Effective random assignment: all IVs are jointly independent of any unmeasured confounder that correlates with both X and Y. Most MR procedures satisfy the relevance assumption by using SNPs that either have been established by other studies to be associated with the risk factor, or have demonstrated genome-wide significant association with the risk factor (e.g., a p-value less than 5×10^{-8}) in the same study that provides the summary data (Bowden et al., 2016; Bowden et al., 2017; Burgess et al.,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

2016; Burgess et al., 2013; Burgess et al., 2020; Hartwig et al., 2017; Qi and Chatterjee, 2019).

Several approaches have been proposed for using invalid instruments that violate either the exclusive restriction or effective random assignment assumption. In the setting of MR analysis, these invalid IVs can be SNPs with pleiotropic effects, that is, SNPs that can affect both the risk factor and the outcome. Some of these approaches require using individual-level raw data (e.g., Guo et al., 2018; Kang et al., 2016; Kang et al., 2020; Tchetgen et al., 2017; Windmeijer et al., 2018). Most recently developed MR procedures use summary data generated from GWAS as input, and allow some of the chosen IVs (SNPs) to have pleiotropic effects (e.g., Bowden et al., 2015; Bowden et al., 2016; Burgess et al., 2020; Hartwig et al., 2017; Mprrison et al., 2020; Qi and Chatterjee, 2019; Xue et al., 2021). They usually choose genome-wide significant SNPs as IVs to satisfy the relevance condition, and require additional assumptions, such as plurality validity, to ensure the identifiability of the model parameters (Guo et al., 2018; Xue et al., 2021).

The strategy of using only genome-wide significant SNPs as IVs works well for risk factors that have been studied in a large-scale GWAS, such as BMI and blood pressure (Evangelou et al., 2018; Yengo et al., 2018), because these studies have adequate power for identifying hundreds of genome-wide significant SNPs to be selected as IVs. However, for the majority of other risk factors, the number of qualified SNPs is well below 100. MR analyses with these risk factors have limited power, because the set of chosen IVs might explain only a small proportion of the total variation of the risk factor (Deng et al., 2020).

Using selected SNPs as IVs has another potential limitation. When the same GWAS is used to select the IVs and to provide the summary statistics, there is a winner's curse effect on the summary data (Siegmund, 2002; Yu et al., 2007). For SNPs that barely pass the selection threshold, their levels of association with the risk factor tend to be over-estimated. Using biased

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

summary statistics in an MR study can lead to erroneous conclusions (Zhao et al., 2020). Zhao et al. (2019) proposed a three-sample genome-wide design that uses summary data from three independent GWAS, with one GWAS used to select the IVs showing evidence of association with the risk factor, and the other two GWAS for the two-sample MR analysis. Because the GWAS used to select the IVs is not used in the MR analysis, this three-sample design avoids the winner's curse effect, and allows the use of many IVs (around 1,000) for a more effective evaluation of the causal effect. However, it is not clear how to select an appropriate set of IVs to achieve optimal performance.

We propose a two-sample MR procedure to test whether there exists a causal effect of the risk factor on the outcome. The new approach uses summary data on all independent SNPs from a GWAS, instead of just a few that are genome-wide significant. The test statistic is the maximum of a sequence of modified K-statistics (Kleibergen, 2002), defined by a range of thresholds. Each K-statistic in the sequence is calculated using summary data on selected SNPs with correlations with the risk factor that are above a given threshold. We use this maximal thresholding statistic to optimize the power of detecting the causal effect manifested through an unspecified set of SNPs. A similar idea is used to identify sparse signals in simultaneous tests of a large number of unrelated hypotheses (Donoho and Jin, 2004; Zhong et al., 2013). The proposed approach gains its power by collecting signals from many weak IVs, but it focuses on testing instead of estimating the causal effect. It is a two-sample MR testing procedure, and does not require an additional GWAS for the selection of IVs. The new approach uses a data-driven threshold to identify the optimal set of SNPs for the test of the causal effect.

2. Method

2.1. Setup and notation

Let Y, X, and $\mathbf{Z}$ represent the outcome, risk factor, and vector of

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

genotypes on p independent SNPs, respectively. Here, we consider all independent SNPs extracted from a GWAS, with p possibly larger than 100,000. In the two-sample MR setting, we have summary data on the association between each SNP and the outcome from the GWAS of Y, and summary data on the association between each SNP and the risk factor from the GWAS of X. We denote the summary data from the outcome GWAS as $\{(\hat{\beta}_{Yk}, \sigma^2_{Yk}), k = 1, \ldots, p\}$, with $\hat{\beta}_{Yk}$ being the estimated regression coefficient on the association between the kth SNP and Y, and σ_{Yk} being the corresponding standard error. We denote the summary data from the risk factor GWAS as $\{(\hat{\beta}_{Xk}, \sigma^2_{Xk}), k = 1, \ldots, p\}$. Similarly to most two-sample MR procedures, we assume the two GWAS are conducted in the same population. If they come from different populations, our proposed method still works under the structural invariance assumption; that is, the underlying models for the risk factor and the outcome remain the same between the two populations (Zhao et al., 2019).

As in many two-sample MR procedures, we assume $\hat{\beta}_{Yk} \sim N(\beta_{Yk}, \sigma^2_{Yk})$ and $\hat{\beta}_{Xk} \sim N(\beta_{Xk}, \sigma^2_{Xk})$, with the unknown β_{Yk} and β_{Xk} representing the kth SNP's true marginal effect on Y and X, respectively. We consider β_{Xk}, for $k = 1, \ldots, p$, as independent and identically distributed (i.i.d.) random samples from a mixture distribution, that is, $\beta_{Xk} \sim \lambda f_X + (1 - \lambda)\delta$, with f_X being an arbitrary distribution, δ being the degenerated distribution taking a constant value zero, and λ being the mixture proportion. We let $d_k = 1$ if β_{Xk} takes a value from f_X, and $d_k = 2$ otherwise. Therefore, the kth SNP is associated with X if $d_k = 1$. An SNP can affect Y in several ways simultaneously. It can influence Y directly through its effect on X or other mediated factors. Here, β_{Yk} represents the kth SNP's true marginal effect on Y, summarizing all those effects. We treat σ_{Yk} and σ_{Xk} as known constant values.

Because we consider a set of independent SNPs in the setting of a two-sample MR study, the p random variables $\hat{\beta}_{Yk}, k = 1, \ldots, p$, are mutually independent. This is also true for $\hat{\beta}_{Xk}, k = 1, \ldots, p$. Zhao et al. (2020)

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

provide further justification. Furthermore, because the summary data come from two independent studies, we can assume that $(\hat{\beta}_{Yk}, \hat{\beta}_{Yk'}, \hat{\beta}_{Xk}, \hat{\beta}_{Xk'})$ are conditionally independent, given the true marginal effects. We call data satisfying these conditions two-sample independent summary data.

2.2. Method under the genome-wide InSIDE assumption

When a large set of SNPs is considered, it is inevitable that some of them have pleiotropic effects. Let α_k be the kth SNP's pleiotropic effect on Y, apart from the one mediated through X. Under the standard instrument strength independent of direct effect (InSIDE) assumption (Bowden et al., 2015), α_k and β_{Xk} are independent. The InSIDE assumption is typically made on the set of SNPs associated with the risk factor. We extend this assumption to SNPs throughout the genome, and call it the genome-wide InSIDE assumption, specified as the follows:

> **Genome-wide InSIDE Assumption**. For $k = 1, \ldots, p$, $\beta_{Yk} = \theta\beta_{Xk} + \alpha_k$, with $\alpha_k \perp \beta_{Xk}$ and α_k being i.i.d. random variables.

In this assumption, we regard $\alpha_k, k = 1, \ldots, p$, as i.i.d. random effects that are independent of β_{Xk} and have the same distribution as α, with $E(\alpha) = \mu$ and $\mathrm{Var}(\alpha) = \omega^2$. The distribution of α is unspecified. It can be a mixture distribution, similar to that of β_{Xk}. In addition, θ is a constant value representing the causal effect of the risk factor. The null hypothesis of the MR test is $H_0 : \theta = 0$.

We first present the test statistic assuming that the mean (μ) and variance (ω^2) of α are known. Then, we provide estimates of μ and ω^2, and show the property of the test statistic after plugging in the two estimates. We consider the following threshold K-statistic (Kleibergen, 2002; Wang and Kang, 2019):

$$Q(s) = \sum_{k=1}^{p} \left\{ \frac{(\hat{\beta}_{Yk} - \mu)}{\sqrt{\omega^2 + \sigma_{Yk}^2}} \frac{\hat{\beta}_{Xk}}{\sigma_{Xk}} \right\} I\left(\frac{\hat{\beta}_{Xk}^2}{\sigma_{Xk}^2} \geq 2s \log p \right), \tag{2.1}$$

where $I(\cdot)$ is an indicator function, and s is a chosen threshold parameter that takes a value within $[0, 1)$. The intuition for using a threshold on the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

K-statistic is to give more weight to SNPs that are likely to be associated with X. We consider the threshold at the scale of $2\log p$ because of the large deviations result (Petrov, 1995), which implies that if $\beta_{Xk} = 0$ for all $1 \le k \le p$, then $\Pr(\max_{1\le k\le p} \hat{\beta}_{Xk}^2/\sigma_{Xk}^2 \le 2\log p) \to 1$ as $p \to \infty$. We require s to be less than one to ensure that a sufficient number of SNPs (e.g., larger than 20) pass that threshold, because we require $Q(s)$ to be asymptotically normal in the proof (see the Supplementary Material S1).

One distinctive feature of $Q(s)$ defined in (2.1) is that the distribution of $\hat{\beta}_{Xk}^2/\sigma_{Xk}^2$ inside the indictor function is unknown under the null, because we do not know which SNP is associated with the outcome. This is different from the threshold statistics considered by Zhong et al. (2013). By assuming μ and ω^2 are known, based on the genome-wide InSIDE assumption, we can calculate the mean of $Q(s)$ as

$$
\begin{aligned}
E_\theta\{Q(s)\} &= \sum_{k=1}^{p} E_\theta\left\{\frac{(\hat{\beta}_{Yk}-\mu)}{\sqrt{\omega^2+\sigma_{Yk}^2}}\right\} E_\theta\left\{\frac{\hat{\beta}_{Xk}}{\sigma_{Xk}} I\left(\frac{\hat{\beta}_{Xk}^2}{\sigma_{Xk}^2} \ge 2s\log p\right)\right\} \\
&= \sum_{k=1}^{p} \frac{\theta\beta_{Xk}}{\sqrt{\omega^2+\sigma_{Yk}^2}}\left[\frac{\beta_{Xk}}{\sigma_{Xk}}\left\{\bar{\Phi}\left(\sqrt{2s\log p}+\frac{\beta_{Xk}}{\sigma_{Xk}}\right)+\bar{\Phi}\left(\sqrt{2s\log p}-\frac{\beta_{Xk}}{\sigma_{Xk}}\right)\right\}\right. \\
&\quad \left. +\,\phi\left(\sqrt{2s\log p}-\frac{\beta_{Xk}}{\sigma_{Xk}}\right)-\phi\left(\sqrt{2s\log p}+\frac{\beta_{Xk}}{\sigma_{Xk}}\right)\right] \equiv \theta\sum_{k=1}^{p} h(\beta_{Xk}), \quad (2.2)
\end{aligned}
$$

where $\phi(\cdot)$ and $\bar{\Phi}(\cdot)$ are the density function and the survival function, respectively, of the standard normal distribution. Here, and throughout this paper, the expectation is calculated over the summary data and the unobserved random effect α_k, conditioning on β_{Xk}. Because $h(\beta_{Xk}) = h(-\beta_{Xk})$, and $h(\beta_{Xk}) \ge 0$ for $\beta_{Xk} \ge 0$, $E_\theta\{Q_p(s)\}$ has the same sign as θ. Note that if all $\beta_{Xk} = 0$, then $E_\theta\{Q_p(s)\} = 0$, regardless of whether $\theta = 0$. Therefore, the proposed test has no power when no SNPs are associated with X.

Under the null, we have $E_{\theta=0}\{Q(s)\} = 0$. The variance of $Q(s)$ under

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

the null can be written as

$$V^2(s) \equiv \mathrm{Var}_{\theta=0}\{Q(s)\} = \sum_{k=1}^{p} E_{\theta=0}\left\{\frac{\hat{\beta}_{Xk}^2}{\sigma_{Xk}^2} I\left(\frac{\hat{\beta}_{Xk}^2}{\sigma_{Xk}^2} \geq 2s\log p\right)\right\}. \quad (2.3)$$

In practice, we can estimate $V^2(s)$ as

$$\hat{V}^2(s) = \sum_{k=1}^{p} \frac{\hat{\beta}_{Xk}^2}{\sigma_{Xk}^2} I\left(\frac{\hat{\beta}_{Xk}^2}{\sigma_{Xk}^2} \geq 2s\log p\right). \quad (2.4)$$

The following result establishes the asymptotic normality of $Q_p(s)$; the proof is given in the Supplementary Material S1.

Theorem 1. *Given two-sample independent summary data* $(\hat{\beta}_{Yk}, \sigma_{Yk}^2, \hat{\beta}_{Xk}, \sigma_{Xk}^2)$, $k = 1, \ldots, p$, *under the genome-wide InSIDE assumption, we have for any fixed* $s \in (0, 1)$, *as* $p \to \infty$, $V^{-1}(s)Q(s) \xrightarrow{D} N(0, 1)$ *and* $\hat{V}^{-1}(s)Q(s) \xrightarrow{D} N(0, 1)$ *under* $H_0 : \theta = 0$.

We can use this result to construct the standardized test statistic $\hat{V}^{-1}(s)$ $Q(s)$ to test $H_0 : \theta = 0$. Note that the calculation of $\hat{V}^2(s)$ in (2.4) does not need $\hat{\beta}_{Yk}$ or σ_{Yk}^2. The above test statistic (2.1) relies on the choice of a threshold s. A more effective approach is to consider all possible thresholds within a given range, and then to choose the one that leads to an optimal test adaptive to the signal-to-noise ratio in the data. Therefore, we propose the following maximal thresholding statistic, called the two-sided MaxK test statistic:

$$T = \max_{s \in \mathcal{S}} \hat{V}^{-1}|Q(s)|, \quad (2.5)$$

where $\mathcal{S} = [s_a, s_b]$. To search for as wide a range as possible, we can let $s_a = 0$ and s_b be a value close to one, such as $s_b = 0.98$. We use this range in the simulation study and the real-data application.

Because both $Q(s)$ and $\hat{V}(s)$ are step functions of s, we can obtain T given by (2.5) exactly by checking $\hat{V}^{-1}(s)|Q(s)|$ at a finite number (at most p) of values for s. We derive the asymptotic distribution of T by showing that $\hat{V}^{-1}(s)Q(s)$ follows a Gaussian process, and establish the asymptotic

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

distribution of T at the tail end (see the Supplementary Material S2 for the proof).

Theorem 2. *Under the conditions given in Theorem 1, we have under* H_0,

$$\lim_{x\to+\infty}\frac{1}{x\phi(x)}\Pr(T>x)-2\tau=0,$$

where $\tau=2^{-1}\log\{\hat{V}^2(s_a)/\hat{V}^2(s_b)\}$.

This result provides an approximation formula to calculate the p-value for the two-sided MaxK test. When T is relatively large, we can calculate its p-value as $2T\phi(T)\tau$. In our numeric experiments, we find this formula works very well for approximating a relatively small p-value (e.g., less than 0.1).

According to (2.2), $E_\theta\{Q(s)\}$ has the same sign as θ. Thus, our procedure can be extended easily to test the direction of the causal effect. For example, to target the one-sided alternative hypothesis $\theta>0$, we can modify the two-sided MaxK statistic as $T_1=\max_{s\in\mathcal{S}}\hat{V}^{-1}(s)Q(s)$. Its p-value can be calculated as $|T_1|\phi(|T_1)\tau$. Similarly, for the alternative $\theta<0$, the one-sided test and its p-value can be $T_2=\max_{s\in\mathcal{S}}\{-\hat{V}^{-1}(s)Q(s)\}$ and $|T_2|\phi(|T_2)\tau$, respectively.

Thus far, we have assumed that the mean (μ) and the variance (ω^2) of the random effect are known. In practice, we can estimate them using summary data on a set of SNPs that are not associated with X. Define $z_{Xk}=\hat{\beta}_{Xk}/\sigma_{Xk}$, and let $\Omega=\{k:|z_{Xk}|<z^*\}$ be the set of SNPs with z-scores for their association with X that are below a certain threshold z^* (e.g., $z^*=1.28$, corresponding to an SNP and X association p-value of 0.20). Given that most GWAS SNPs are not associated with X, it is reasonable to claim that we have $\beta_{Xk}=0$ for $k\in\Omega$. Together with the genome-wide InSIDE assumption, we can estimate μ and ω^2 as

$$\hat{\mu}=\frac{1}{|\Omega|}\sum_{k\in\Omega}\hat{\beta}_{Yk}, \tag{2.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

$$\hat{\omega}^2 = \frac{1}{|\Omega|}\sum_{k\in\Omega}\left\{(\hat{\beta}_{Yk} - \hat{\mu})^2 - \sigma_{Yk}^2\right\}, \tag{2.7}$$

where $|\Omega|$ is the size of Ω. These two estimates are consistent with any θ, as long as $\beta_{Xk} = 0$, for $k \in \Omega$.

We can replace μ and ω^2 with their estimates in the calculation of $Q(s)$, and define it as $\hat{Q}(s)$. In practice, we can conduct the MaxK test with the following test statistic:

$$T_{GW} = \max_{s\in\mathcal{S}} \hat{V}^{-1}(s)|\hat{Q}(s)|. \tag{2.8}$$

We call this version the MaxK-1 test. Let m be the number of SNPs with $d_k = 1$. We show the following result in the Supplementary Material S3.

Corollary 1. *Under the conditions given in Theorem 1, if* $\hat{\mu}-\mu = o_p(m^{-1/2})$ *and* $\hat{\omega}^2 - \omega^2 = o_p(1)$*, then* T *and* T_{GW} *share the same asymptotic distribution.*

According to Corollary 1, as long as we have reasonably precise estimates of μ and ω^2, we can apply the formula given by Theorem 2 to approximate the p-value of the MaxK-1 test. Because we can use most of the SNPs to estimate μ and ω^2 according to (2.6) and (2.7) under the genome-wide InSIDE assumption, we have $|\Omega| = O_p(p)$. Therefore, the conditions listed in Corollary 1 are clearly met.

Next, we study the consistency of the MaxK test, assuming that μ and ω^2 are known. For the purpose of illustration, we consider the following simplified model. Suppose that among the p considered SNPs, there are $m = p^{1-\kappa}$ SNPs associated with X, with $\beta_{Xk}/\sigma_{Xk} = \sqrt{2r\log p}$, where κ is the parameter controlling the proportion of SNPs associated with the risk factor, and r can be viewed as the instrument strength, specifying the SNP's effect size on the risk factor. We further assume $\sigma_{Xk}^2 = \sigma_{Yk}^2 = \sigma^2$, for $k = 1, \ldots, p$. Then, under an alternative hypothesis $H_1 : \theta \neq 0$, the mean

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

of $Q(s)$ can be calculated as

$$\begin{aligned}&E_\theta(S) \equiv E_\theta\{Q(s)\}\\ =&\frac{\theta p^{1-\kappa}\sigma\sqrt{2r\log p}}{\sqrt{\sigma^2+\omega^2}}E\left\{\frac{\hat\beta_{Xk}}{\sigma_{Xk}}I\left(\frac{\hat\beta_{Xk}^2}{\sigma_{Xk}^2}\geq 2s\log p\right)\right\} = \frac{\theta p^{1-\kappa}\sigma\sqrt{2r\log p}}{\sqrt{\sigma^2+\omega^2}}h_s(r),\end{aligned}$$

with

$$\begin{aligned}h_s(r) = \sqrt{2s\log p}\{&\bar\Phi(\sqrt{2s\log p}+\sqrt{2r\log p})+\bar\Phi(\sqrt{2s\log p}-\sqrt{2r\log p})\}\\ &+\phi(\sqrt{2s\log p}-\sqrt{2r\log p})-\phi(\sqrt{2s\log p}+\sqrt{2r\log p}).\end{aligned}$$

The variance of $Q(s)$ can be calculated as

$$\begin{aligned}&V_\theta^2(s) \equiv \mathrm{Var}_\theta\{Q(s)\}\\ =&V^2(s)+\frac{2\theta^2 p^{1-\kappa}\sigma^2 r\log p}{\sigma^2+\omega^2}\mathrm{Var}\left\{\frac{\hat\beta_{Xk}}{\sigma_{Xk}}I\left(\frac{\hat\beta_{Xk}^2}{\sigma_{Xk}^2}\geq 2s\log p\right)\right\}\\ =&V^2(s)+\frac{2\theta^2 p^{1-\kappa}\sigma^2 r\log p}{\sigma^2+\omega^2}\{g_s(r)-h_s(r)\},\end{aligned}$$

with

$$\begin{aligned}g_s(r) = (1+2r\log p)\{&\bar\Phi(\sqrt{2s\log p}+\sqrt{2r\log p})+\bar\Phi(\sqrt{2s\log p}-\sqrt{2r\log p})\}\\ &+(\sqrt{2s\log p}+\sqrt{2r\log p})\phi(\sqrt{2s\log p}-\sqrt{2r\log p})\\ &+(\sqrt{2s\log p}-\sqrt{2r\log p})\phi(\sqrt{2s\log p}+\sqrt{2r\log p}),\end{aligned}$$

and $V^2(s)$ given by (2.3).

Similarly to the proof of Theorem 1, we can establish the asymptotic normality of $Q(s)$ under H_1 as $V_\theta^{-1}(s)\{Q(s)-E_\theta(s)\} \xrightarrow{D} N(0,1)$. Furthermore, we have the following result on the consistency of the MaxK test as $p\to\infty$.

Theorem 3. *Under the above considered model and $H_1: \theta\neq 0$, (i) if $r>\rho^*(\kappa)$, the power of the MaxK test converges to one with the nominal size $\alpha(p) = O\{(\log\log p)^{1/2}(\log p)^{-1}\}$ as $p\to\infty$; (ii) if $r<\rho^*(\kappa)$, the power of the MaxK test converges to zero when the nominal size $\alpha(p)\to 0$ as $p\to\infty$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

The definition of $\rho^*(\kappa)$ and the proof of Theorem 3 are given in the Supplementary Material S4. According to this result, given the proportion of X associated SNPs, the MaxK test is consistent if the instrument strength (r) of those SNPs is stronger than $\rho^*(\kappa)$. Because $\rho^*(\kappa)$ is a monotone increasing function of κ, the instrument strength required to ensure the test consistency becomes higher as the proportion of risk factor associated SNPs decreases.

2.3. Further relaxation of the genome-wide InSIDE assumption

Because of the wide spread of SNPs with pleiotropic effects, it can be argued that if an SNP has an effect on X, it might have a higher chance of affecting Y, as compared with an SNP randomly picked from the genome. Thus, the genome-wide InSIDE assumption might not be appropriate. We can relax this assumption to allow it to hold, conditioning on whether the SNP is associated with the risk factor.

> **Conditional InSIDE Assumption**. For $k = 1, \ldots, p$, $\beta_{Yk} = \theta\beta_{Xk} + \alpha_k$. β_{Xk} and α_k are conditionally independent given d_k. Furthermore, α_k are i.i.d. with $\alpha_k \sim \alpha^{(1)}$ among SNPs having $d_k = 1$, and α_k are i.i.d. with $\alpha_k \sim \alpha^{(2)}$ among SNPs having $d_k = 2$.

The conditional InSIDE assumption is the same as the standard InSIDE assumption on the set of SNPs associated with X. For SNPs not associated with X (i.e., $d_k = 2$), the requirement of $\alpha_k \perp \beta_{Xk}$ is always met, because β_{Xk} is constant zero. In that sense, we can regard the conditional InSIDE assumption as being the same as the standard InSIDE assumption. Both assumptions are reasonable if the considered SNPs are not related to any genetic pathway or confounder that affects both X and Y.

In the above assumption, $\alpha^{(1)}$, the random effect on Y of an SNP associated with X, is allowed to have a different distribution from $\alpha^{(2)}$, which is the random effect from an SNP not associated with X. The distributions of $\alpha^{(1)}$ and $\alpha^{(2)}$ can be arbitrary. When the two have the same distribution, the conditional InSIDE assumption reduces to the genome-wide InSIDE

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

assumption. To illustrate the difference between the two, we consider the following four-component mixture model suggested by Qi and Chatterjee (2021):

$$\begin{pmatrix} \beta_{Xk} \\ \alpha_k \end{pmatrix} \sim \pi_1 \begin{pmatrix} N(0, \sigma_X^2) \\ \delta \end{pmatrix} + \pi_2 \begin{pmatrix} N(0, \sigma_X^2) \\ N(\mu_Y, \sigma_Y^2) \end{pmatrix} + \pi_3 \begin{pmatrix} \delta \\ N(\mu_Y, \sigma_Y^2) \end{pmatrix} + \pi_4 \begin{pmatrix} \delta \\ \delta \end{pmatrix}. \tag{2.9}$$

Under this model, we know $\alpha^{(1)} \sim \frac{\pi_2}{\pi_1+\pi_2} N(\mu_Y, \sigma_Y^2) + \frac{\pi_1}{\pi_1+\pi_2}\delta$, and $\alpha^{(2)} \sim \frac{\pi_3}{\pi_3+\pi_4} N(\mu_Y, \sigma_Y^2) + \frac{\pi_4}{\pi_3+\pi_4}\delta$. This model clearly satisfies the conditional InSIDE assumption. It satisfies the genome-wide InSIDE condition if $\pi_1\pi_3 = \pi_2\pi_4$.

Let the mean and the variance of the random effect $\alpha^{(i)}$ be μ_i and ω_i^2, respectively, for $i = 1, 2$. To adopt the MaxK test under the conditional InSIDE assumption, we can modify the definition of $Q(s)$ as follows:

$$Q^*(s) = \sum_{k=1}^{p} \left\{ \frac{(\hat{\beta}_{Yk} - \mu_{d_k})}{\sqrt{\omega_{d_k}^2 + \sigma_{Yk}^2}} \frac{\hat{\beta}_{Xk}}{\sigma_{Xk}} \right\} I\left(\frac{\hat{\beta}_{Xk}^2}{\sigma_{Xk}^2} \geq 2s \log p \right).$$

In $Q^*(s)$, each SNP's contribution is adjusted by either (μ_1, ω_1^2) or (μ_2, ω_2^2), depending on whether $d_k = 1$ or 2. In $Q(s)$, the same (μ, ω^2) is applied to all SNPs. If we know d_k for each SNP, we can estimate (μ_1, ω_1^2) using summary data on the SNPs belonging to $\mathcal{M} = \{k : d_k = 1\}$ as

$$\hat{\omega}_1^2 = \frac{1}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} \left\{ (\hat{\beta}_{Yk} - \hat{\mu}_1)^2 - \hat{\theta}^2 \hat{\sigma}_{Xk}^2 - \sigma_{Yk}^2 \right\}, \tag{2.10}$$

where $(\hat{\mu}_1, \hat{\theta})$ are the estimated coefficients from the linear regression model $\hat{\beta}_{Yk} = \mu_1 + \theta \hat{\beta}_{Xk} + \epsilon_k$, $k \in \mathcal{M}$, with ϵ_k being the error term. The same procedure can be used to estimate (μ_2, ω_2^2) using the SNPs in $\bar{\mathcal{M}} = \{k : d_k = 2\}$. Because we do not know d_k in practice, we propose the following strategy for calculating $Q^*(s)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Because an SNP in $\mathcal{M}$ tends to have a larger $|z_{Xk}|$ than does an SNP from $\bar{\mathcal{M}}$, the order of $|z_{Xk}|$ is related to the membership of $\mathcal{M}$. Furthermore, when all independent SNPs from a GWAS are considered, it is reasonable to assume $|\bar{\mathcal{M}}|$ is much larger than $|\mathcal{M}|$. Given these two observations, we use the following strategy. First, we arrange all SNPs according to $|z_{Xk}|$ in descending order. Then, we divide the p SNPs evenly into L groups, so that each group consists of about $\lceil p/L \rceil$ consecutive SNPs. In the following numerical experiments and real-data application, we let $\lceil p/L \rceil = 100$. Let G_l be the set of SNPs belonging to group l, $1 \leq l \leq L$. For all SNPs in G_l, we can assume their random effects on Y share a common distribution, and estimate its mean and variance as $\hat{\mu}_l$ and $\hat{\omega}_l^2$, respectively, according to (2.10) using the summary data on the SNPs in G_l. Finally, we can define the following statistic to approximate $Q^*(s)$:

$$\hat{Q}^*(s) = \sum_{l=1}^{L} \sum_{k \in G_l} \left\{ \frac{(\hat{\beta}_{Yk} - \hat{\mu}_l)}{\sqrt{\hat{\omega}_l^2 + \sigma_{Yk}^2}} \frac{\hat{\beta}_{Xk}}{\sigma_{Xk}} \right\} I\left(\frac{\hat{\beta}_{Xk}^2}{\sigma_{Xk}^2} \geq 2s \log p \right).$$

Noting that if $G_l \subseteq \mathcal{M}$, $\hat{\mu}_l$ and $\hat{\omega}_l^2$ should be valid estimates of μ_1 and ω_1^2, respectively. As a result, for SNPs in G_l, their contributions to $\hat{Q}^*(s)$ are approximately the same as those to $Q^*(s)$. The same is true for $G_l \subseteq \bar{\mathcal{M}}$. The variance of $\hat{Q}^*(s)$ can still be estimated with (2.4). We define the new MaxK test statistic as

$$T_{Cond} = \max_{s \in \mathcal{S}} \hat{V}^{-1}(s) |\hat{Q}^*(s)|. \tag{2.11}$$

We call this version the MaxK-2 test, and prove the following result in the Supplementary Material S5.

Corollary 2. *Given two-sample independent summary data, under the conditional InSIDE assumption, if $G_l \subseteq \mathcal{M}$ or $G_l \subseteq \bar{\mathcal{M}}$, $|G_l| \to \infty$ as $p \to \infty$, for $1 \leq l \leq L$, and if β_{Xk}/σ_{Xk}, $1 \leq k \leq p$, follows a symmetrical distribution around zero, then under H_0, we have*

$$\lim_{x \to +\infty} \frac{1}{x\phi(x)} \Pr(T_{Cond} > x) - 2\tau = 0,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

where $\tau = 2^{-1}\log\{\hat{V}^2(s_a)/\hat{V}^2(s_b)\}$.

In real applications, we cannot ensure all G_l, $1 \le l \le L$, are subsets of $\mathcal{M}$ or $\bar{\mathcal{M}}$. However, when L is relatively large (e.g., $L > 100$), because of the way all groups are formed and the fact that $|\bar{\mathcal{M}}|$ is much larger than $|\mathcal{M}|$, $G_l \subseteq \mathcal{M}$ or $G_l \subseteq \bar{\mathcal{M}}$ should be true (or almost true) for most of the L groups, with only a few groups having a mixed bag of SNPs from $\mathcal{M}$ and $\bar{\mathcal{M}}$. Therefore, we can still use Corollary 2 to approximate the p-value of the MaxK-2 test.

Corollary 2 requires β_{Xk}/σ_{Xk} to be symmetric around zero. A similar condition is also needed by the procedure of Zhao et al. (2019). Because the sign of β_{Xk} depends on the genotype coding at the SNP, we can adopt a coding scheme to ensure this. We first rearrange the SNPs according to $|z_{Xk}|$ in descending order, and then choose the genotype coding in such a way that the sign of $\hat{\beta}_{Xk}$ is alternated along the ordered sequence. Given that we usually deal with a large number of SNPs (e.g., $p > 50,000$), this coding scheme should ensure β_{Xk}/σ_{Xk} is nearly symmetric. Numerical experiments described later confirm that p-values can be estimated accurately by the formula given by Corollary 2 using this strategy.

Corollary 2 can be valid under other conditions. For example, if we assume $\mu_i = 0$, $i = 1, 2$, we can set $\hat{\mu}_l = 0$ in the calculation of T_{Cond}. Under this situation, we can show that the test based on T_{Cond} is still valid, regardless of the distributional property of β_{Xk}/σ_{Xk}.

2.4. MaxK test with weakly dependent SNPs

So far, we have described MR methods using summary data from independent SNPs. We show how to ensure this independence in our application to real data by removing dependent SNPs. Next, we consider the MaxK test with a set of weakly dependent SNPs.

We rearrange the SNPs according to their locations on each chromosome. Let $Z_k, k = 1, \ldots, p$, be a vector of genotypes on the kth SNP from the two GWAS. Instead of requiring that they are independent, we can

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

allow $\mathbf{Z} = \{Z_k\}_{k=1}^p$ to have a weak correlation structure, called ρ-mixing, such that their ρ-mixing coefficients $\rho_Z(k) \leq Cv^k, k = 1, \ldots, p-1$, where v and C are constant values satisfying $0 < v < 1$, and $C > 0$. The ρ-mixing coefficient is defined as

$$\rho_Z(k) = \sup_{1 \leq l \leq p-1, \xi \in L^2(\mathcal{F}_1^l), \eta \in L^2(\mathcal{F}_{l+k}^p)} |\text{Corr}(\xi, \eta)|,$$

with $\mathcal{F}_m^n$ being the σ-algebra generated by $\{Z_k\}_{k=m}^n$. Further discussion on the concept of ρ-mixing can be found in Doukhan (1994). The ρ-mixing dependent structure implies that the genotype correlation between two SNPs decreases exponentially over their distance, which, in general, makes sense in a human genome, especially after highly correlated SNPs are pruned away. An example of this structure is the autocorrelation structure with the correlation coefficient of Z_i and Z_j being $\rho_0^{|i-j|}$, for some constant value $0 < \rho_0 < 1$.

We can show that all results obtained with independent SNPs still hold with ρ-mixing dependent SNPs. For example, in the Supplementary Material S6, we prove the following conclusion.

Theorem 4. *For summary data from two separate GWAS on SNPs with a ρ-mixing dependent structure, under the genome-wide InSIDE assumption, the results stated in Theorems 1 and 2 still hold on SNPs with a ρ-mixing correlation structure.*

3. Simulation study

3.1. Under the InSIDE assumption

We conducted simulation studies to evaluate the performance of the two MaxK tests. We adopted a similar simulation model setup to that described in Qi and Chatterjee (2021). We assumed that summary data on a set of $p = 200,000$ independent SNPs were generated from a risk factor GWAS and an outcome GWAS, where each GWAS had N subjects. Using the same notation as before, we assumed β_{Xk} is the kth SNP's effect

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

on X, and α_k is its random effect on Y. They follow the four-component mixture model given by (2.9). The kth SNP's true marginal effect on Y is defined as $\beta_{Yk} = \theta\beta_{Xk} + \alpha_k$. We considered two scenarios based on model (2.9): Scenario I under the genome-wide InSIDE assumption, and Scenario II under the conditional InSIDE assumption. In both scenarios, we fixed $\sigma_X^2 = \sigma_Y^2 = 10^{-5}$. In Scenario I, we set $\pi_1 = 1.96\%$, $\pi_2 = 0.04\%$, $\pi_3 = 1.96\%$, and $\pi_4 = 96.04\%$. Because $\pi_1\pi_3 = \pi_2\pi_4$ under this setting, β_{Xk} and α_k are generated independently. We further assumed $\mu_Y = 0$ or 0.005 in (2.9), which correspond to the balanced pleiotropy and unbalanced pleiotropy setting, respectively. For Scenario II, we set $\pi_1 = 1\%$, $\pi_2 = 1\%$, $\pi_3 = 1\%$, and $\pi_4 = 97\%$. Under this setting, 50% of the SNPs associated with X had pleiotropic effects on Y. Similarly to Scenario I, we chose $\mu_Y = 0$ or 0.005 to generate balanced or unbalanced pleiotropic effects. For each given mixture model (2.9), we chose the causal effect θ within the interval $[0, 0.1]$, with the sample size for each GWAS (N) falling between 300,000 and 500,000.

Given the causal effect θ, parameters in model (2.9), and sample size N, we generated (β_{Yk}, β_{Xk}) for each SNP independently, and then simulated summary data as $\hat{\beta}_{Yk} \sim N(\beta_{Yk}, \frac{1}{N})$ and $\hat{\beta}_{Xk} \sim N(\beta_{Xk}, \frac{1}{N})$, for $k = 1, \ldots, p$. We replicated the above steps to create 2,000 summary data under each setting to evaluate the performances of the considered tests, which included the inverse-variance weighted method with multiplicative random effects (IVW) (Burgess et al., 2013), weighted median estimate (W-Median) (Bowden et al., 2016), weighted mode estimate (W-Mode) (Hartwig et al., 2017), MR-Egger (Bowden et al., 2015), MR-Robust (Burgess et al., 2016), contamination mixture (Con-mix) (Burgess et al., 2020), MRMix (Qi and Chatterjee, 2019), and two versions of the MaxK test, given by (2.8) and (2.11), respectively. Except for the two MaxK tests, all other tests used summary statistics on genome-wide significant SNPs (i.e., SNPs having an X association p-value of less than 5×10^{-8}). Clearly, additional significant SNPs become available as N increases. In our considered settings, as N in-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

creases from 300k to 500k, the average number of significant SNPs changes from 26 to 104.

First, we compare all tests under the genome-wide InSIDE assumption (Scenario I). Table 1 summarizes the empirical type-I errors for all considered tests when $\mu_Y = 0$ (i.e., balanced pleiotropy). Table 1 shows that five tests, namely MaxK-1, MaxK-2, IVW, MR-Egger, and MR-Robust, maintain their type-I errors properly. W-Median, W-Mode, and MRMix are over conservative, especially W-Mode. The Con-mix test has an inflated type-I error. Similar conclusions on the type-I error evaluation are reached under unbalanced pleiotropy (Supplementary Table 1). Figure 1 shows the power comparison under balanced pleiotropy with N = 300k, 400k, 450k, and 500k. We did not consider Con-Mix in the power comparison, because it has an inflated type-I error rate. Figure 1 shows that both versions of the MaxK test have a clear power advantage over other tests. MaxK-1 and MaxK-2 are almost indistinguishable, especially when $N \geq$ 400k. This suggests that using a locally estimated mean and variance of the random effect in MaxK-2 does not lead to any noticeable loss of efficiency. Similar conclusions can be made on the power comparison in the simulations under unbalanced pleiotropy (Supplementary Figure 1).

Next, we compare all tests under the conditional InSIDE assumption (Scenario II). Table 2 and Supplementary Table 2 provide the empirical type-I error rates under balanced and unbalanced pleiotropy, respectively. We can see from both tables that MaxK-2, IVW, MR-Egger, and MR-Robust properly maintain their type-I errors under all considered sample sizes. The performance of MRMix depends on the number of significant SNPs. It can maintain its type-I error appropriately only with a relatively large number of genome-wide significant SNPs. The other tests (W-Median, W-Mode, Con-mix, and MaxK-1) are either too conservative or too liberal. The MaxK-1 test cannot control its type-I error, because it estimates the mean and variance of the random effect under the independent assumption, which is not valid under Scenario II. We exclude Con-mix and MaxK-1 from

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

the power comparison because of their highly inflated type-I errors. Figure 2 and Supplementary Figure 2 show the power comparison under balanced and unbalanced pleiotropy with various sample sizes. According to both figures, MaxK-2 appears to be the clear winner.

Finally, to demonstrate the advantage of the proposed MaxK-2 procedure over the testing procedure based on the original K-statistic, we compared its power with that of tests based on $\hat{V}^{-1}(s)|\hat{Q}^*(s)|$, with a fixed s threshold. We considered $s = 0$, $s = 1.64/(2\log p)$, and $s = 1.96/(2\log p)$, and denoted the corresponding tests as K(0.0), K(1.64), and K(1.96), respectively. K(0.0) is equivalent to the orginal K-statistic. We found that the MaxK-2 procedure has a noticeable power advantage over the tests with a fixed threshold. For example, using the same simulation setup under the conditional InSIDE assumption, we compared their power under various sample sizes and causal effect sizes, shown in Supplementary Figures 3 and 4 for balanced and unbalanced pleiotropic effects, respectively. Both figures illustrate the robust performance of the MaxK-2 procedure.

3.2. Under correlated pleiotropic effects

We evaluated the robustness of MaxK-2 when the InSIDE assumption is not met (i.e., with correlated pleiotropic effects). We considered the InSIDE assumption violated pleiotropy model described in Qi and Chatterjee (2021), where some SNPs have correlated effects (with a 10% correlation coefficient) on the outcome and the risk factor, owing to their collections with a common mediation factor. We found that MaxK-2 can control its type-I error reasonably well when less than 10% of the risk factor-associated SNPs have a correlated effect on the outcome (Supplementary Table 3). However, it tends to have an inflated type-I error when the percentage of SNPs with correlated effects becomes large (Supplementary Table 3). This is expected because MaxK-2 is derived under the conditional InSIDE assumption.

4. Real application

Jones et al. (2019) recently studied the genetic basis underlying vari-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

ous human sleep behaviors. In their study, they conducted MR analyses to identify risk factors with causal effects on sleep behaviors. In particular, they considered eight sleep behaviors, quantified by accelerometer-derived measures, and used an MR to assess whether the waist-hip ratio (WHR) causally affects them. Here, we apply our new test to re-assess those relationships. The summary data on the SNP's association with the WHR (after adjusting the BMI) are obtained from Shungin et al. (2015). Summary data on the SNP's association with each sleep behavior are taken from Jones et al. (2019). The eight considered sleep behaviors (inverse-normalized) are listed in Table 3.

We preprocessed the summary data using the following criteria. We restricted the SNPs to those sharing a WHR and sleep behavior GWAS and with minor allele frequencies (MAFs) larger than 2%. MAFs were estimated using European reference genomes from the 1000 Genomes (Genomes et al., 2015). Then, we used the clumping function of PLINK (Purcell et al., 2007) with $r^2 = 0.1$ within a window size of 1000 kb as the linkage disequilibrium threshold to select a set of independent SNPs. When we applied the clumping procedure, we randomly picked index SNPs without referring to their levels of association with the WHR and sleep behaviors, in order to ensure there was no selection bias. In the end, we had 95,819 SNPs for the MaxK test. For the MR analysis with other considered procedures, we used a set of 56 independent SNPs that were genome-wide significantly associated with the WHR (i.e., with a p-value less than 5×10^{-8}) in the WHR GWAS.

The results are summarized in Table 3. We do not present the results from Con-mix and MaxK-1, because they had inflated type-I errors, according to our simulation results. From Table 3, we can see that the MaxK-2 test detects most signals among all considered tests, with five out of eight outcomes having MaxK-2 test p-values less than 0.05, and three having MaxK-2 test p-values less than the Bonferroni threshold (i.e., $0.05/8 \approx 0.006$). Among all considered tests, the MaxK-2 test has the most significant result

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

on four outcomes (i.e., L5 timing, sleep duration variability, sleep efficiency, and sleep midpoint timing), and MRMix and MR-robust each have one.

In this real-data example, except for the proposed MaxK test, the other seven tests used summary data on 56 SNPs that were genome-wide significantly associated with the WHR. As pointed out earlier, the requirement of using only genome-wide significant SNPs is a major limitation of these approaches. By combining signals throughout the genome, instead of relying on a few SNPs, the Max-K test detects the casual effect of the WHR on more sleep habits.

5. Discussion

We propose a new MR test (called the MaxK test) that takes full advantage of information generated from GWAS. Unlike most existing procedures that rely on a few SNPs that demonstrate strong evidence for their association with the risk factor, the MaxK test synthesizes evidence of a causal effect from tens of thousands of SNPs studied by GWAS. This test can properly control its type-I error under the InSIDE assumption with balanced or unbalanced pleiotropy. It is more powerful than existing approaches, even when there is only a small proportion (e.g., 1 or 2%) of SNPs carrying the signal.

It is challenging to develop MR procedures with a properly controlled type-I error when some IVs have correlated pleiotropic effects. The proposed method has difficulty maintaining its type-I error when a large proportion of considered SNPs have correlated pleiotropic effects. Another limitation is that our method does not estimate the causal effect. Given its promising performance under the InSIDE assumption, it would be worthwhile improving the MaxK procedure to have more robust performance when SNPs with correlated pleiotropic effects are used as IVs. Finally, the proposed procedure focuses on testing the null hypothesis that the risk factor has no casual effect on the outcome. It does not provide an estimate of the causal effect. Further investigation is needed to expand the procedure to evaluate the magnitude of the casual effect.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Supplementary Material

All technical details and additional numeric results are relegated to the online Supplementary Material.

Acknowledgments

The study used the computational resource of the NIH Biowulf cluster (https://hpc.nih.gov/). The research of Dr. Lu Deng was partially supported by the National Natural Science Foundation of China, grant #12101331. The authors would like to thank the associate editor and two referees for their insightful comments.

References

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-525.

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol* **40**, 304-314.

Bowden, J., Del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med* **36**, 1783-1802.

Burgess, S., Bowden, J., Dudbridge, F., and Thompson, S. G. (2016). Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *arXiv 1606.03279.*

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37**, 658-665.

Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. M. (2020). A robust

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat Commun* **11**, 376.

Burgess, S., Scott, R. A., Timpson, N. J., Davey Smith, G., Thompson, S. G., and Consortium, E.-I. (2015). Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* **30**, 543-552.

Deng, L., Zhang, H., and Yu, K. (2020). Power calculation for the general two-sample Mendelian randomization analysis. *Genet Epidemiol* **44**, 290-299.

Didelez, V., and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res* **16**, 309-330.

Donoho, D. L., and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* **32**, 962-994.

Doukhan, P. (1994). Mixing: Properties and Examples. In Lecture Notes in Statistics. *New York: Springer.*

Evangelou, E., Warren, H. R., Mosen-Ansorena, D., et al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet* **50**, 1412-1425.

Genomes Project, C., Auton, A., Brooks, L. D., et al. (2015). A global reference for human genetic variation. *Nature* **526**, 68-74.

Guo, Z., Kang, H., Cai, T. T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *J R Statist Soc B* **80**, 793-815.

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol* **46**, 1985-1998.

Hemani, G., Zheng, J., Elsworth, B., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Jones, S. E., van Hees, V. T., Mazzotti, D. R., et al. (2019). Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. *Nat Commun* **10**, 1585.

Kang, H., Lee, Y., Cai, T. T., and Small, D. S. (2020). Two robust tools for inference about causal effects with invalid instruments. *Biometrics* doi: 10.1111/biom.13415. Epub ahead of print. PMID: 33616910.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016) Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *J Am Stat Assoc* **111**, 132-144,

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* **70**, 1781-1803.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**, 1133-1163.

Morrison, J., Knoblauch, N,, Marcus, J. H., Stephens, M. and He, X. (2020) Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat Genet* **52**, 740-747.

Petrov, V. V. (1995). Limit theorems of probability theory. Sequences of independent random variables. *New York: Oxford University Press.*

Purcell, S., Neale, B., Todd-Brown, K., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575.

Qi, G., and Chatterjee, N. (2019). Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nat Commun* **10**, 1941.

Qi, G., and Chatterjee, N. (2019b). A comprehensive evaluation of methods for Mendelian randomization using realistic simulations and an analysis of 38 biomarkers for risk of type-2 diabetes. *Int J Epidemiol* **50**, 1335-1349.

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., et al. (2015). New genetic loci link adipose

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

and insulin biology to body fat distribution. *Nature* **518**, 187-196.

Siegmund, D. (2002). Upward bias in estimation of genetic effects. *Am J Hum Genet 71*, 1183-1188.

Smith, G. D., and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1-22.

Tchetgen, E. J. T., Sun, B., and Walter, S. (2017). The GENIUS approach to robust Mendelian randomization inference. *Statist Sci* **36**, 443 - 464,

Wang, S., and Kang, H. (2019). Weak-Instrument robust tests in two-sample summary-data Mendelian randomization. *arXiv 1909.06950*

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2018). On the use of the lasso for instrumental variables estimation with some invalid instruments. *J Am Stat Assoc* **114**, 1339-1350.

Xue, H., Shen, X., and Pan, W. (2021). Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *Am J Hum Genet* **108**, 1251-1269.

Yengo, L., Sidorenko, J., Kemper, K. E., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649.

Yu, K., Chatterjee, N., Wheeler, W., et al. (2007). Flexible design for following up positive findings. *Am J Hum Genet* **81**, 540-551.

Zhao, Q., Chen, Y., Wang, J., and Small, D. S. (2019). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *Int J Epidemiol* **48**, 1478-1492.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Annals of Statistics **48**, 1742-1769.

Zhong, P., Chen, S. X., and Xu, M. (2013). Tests alternative to Higher criticism for high dimensional means under sparsity and column-wise dependence. *Annals of Statistics* **41**, 2820-2851.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Table 1: Simulation results on type-I errors under the genome-wide InSIDE assumption with balanced pleiotropy. The results are summarized based on the performance over 2000 data sets generated from two GWAS of equal sample size (N). Each simulated data set consists of summary statistics on 200,000 independent SNPs.

	Sample size N (average number of significant SNPs)[a]				
MR method[b]	300k(26)	350k(41)	400k(59)	450k(80)	500k(104)
MaxK-1	0.044	0.048	0.046	0.048	0.049
MaxK-2	0.038	0.044	0.042	0.045	0.044
IVW	0.045	0.043	0.045	0.042	0.048
W-Median	0.026	0.023	0.021	0.025	0.019
W-Mode	0.001	0.001	0.001	0.001	0.002
IVW-Robust	0.040	0.041	0.042	0.042	0.046
MR-Egger	0.041	0.048	0.045	0.056	0.045
Con-mix	0.122	0.155	0.170	0.192	0.208
MRMix	0.031	0.028	0.024	0.021	0.016

a. Summary data are generated from two GWAS of equal sample size N. The average number of significant SNPs is the number of SNPs with risk factor association p-values less than 5×10^{-8}, averaged over 2000 simulated data sets.

b. Both MaxK-1 and MaxK-2 use summary statistics on 200,000 independent SNPs. All other tests use summary statistics on SNPs that are genome-wide significantly associated with the risk factor.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Table 2: Simulation results on type-I errors under the conditional InSIDE assumption with balanced pleiotropy. The results are summarized based on the performance over 2000 data sets generated from two GWAS of equal sample size (N). Each simulated data set consists of summary statistics on 200,000 independent SNPs.

	Sample size N (average number of significant SNPs)[a]				
MR method[b]	300k(26)	350k(41)	400k(59)	450k(80)	500k(104)
MaxK-1	0.254	0.291	0.325	0.347	0.365
MaxK-2	0.049	0.052	0.057	0.058	0.058
IVW	0.064	0.073	0.057	0.056	0.051
W-Median	0.081	0.079	0.076	0.073	0.071
W-Mode	0.012	0.004	0.004	0.004	0.004
IVW-Robust	0.061	0.064	0.062	0.055	0.060
MR-Egger	0.058	0.053	0.048	0.048	0.058
Con-mix	0.196	0.215	0.206	0.210	0.218
MRMix	0.089	0.069	0.059	0.058	0.052

a. Summary data are generated from two GWAS of equal sample size N. The average number of significant SNPs is the number of SNPs with risk factor association p-values less than 5×10^{-8}, averaged over 2000 simulated data sets.

b. Both MaxK-1 and MaxK-2 use summary statistics on 200,000 independent SNPs. All other tests use summary statistics on SNPs that are genome-wide significantly associated with the risk factor.

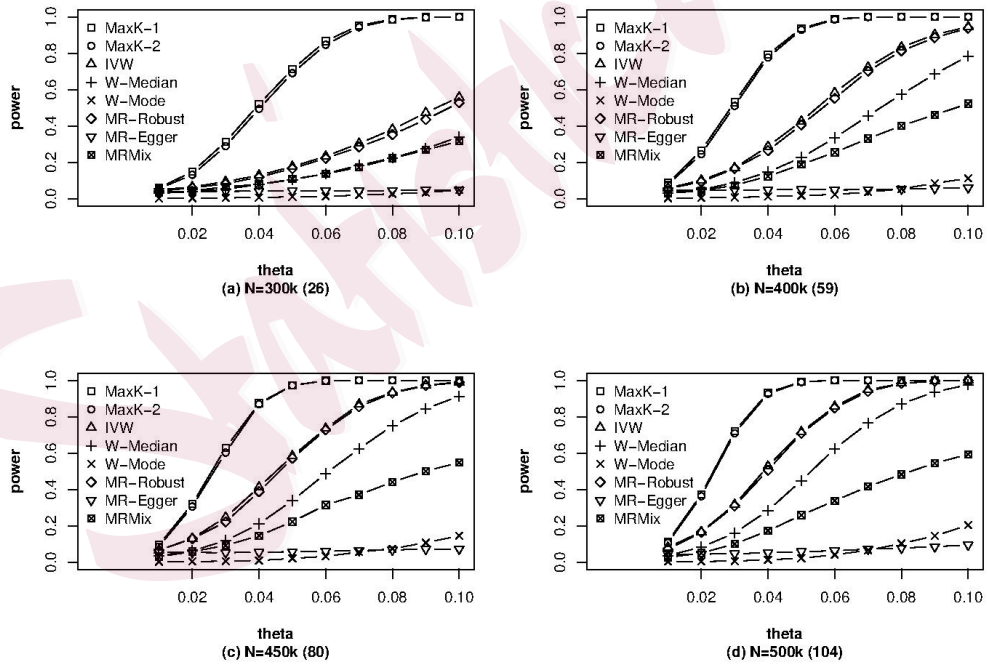
Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Table 3: MR testing results (p-values) on the casual effect of the waist-hip ratio on eight sleep traits.

	MR method			
Outcome	Maxk-2	IVW	W-Median	W-Mode
Diurnal inactivity	9.84E-01	1.20E-01	9.12E-02	6.14E-01
L5 timing	2.44E-02	5.83E-01	5.40E-01	3.27E-01
M10 timing	8.18E-01	4.76E-01	1.64E-01	3.04E-01
Num. nocturnal sleep episodes	4.67E-01	9.75E-01	3.60E-01	5.39E-01
Sleep duration	1.69E-04	7.56E-05	1.33E-02	5.78E-01
Sleep duration variability	1.07e-03	5.98E-02	1.27E-02	1.29E-01
Sleep efficiency	2.26E-04	2.40E-03	2.85E-03	1.51E-01
Sleep midpoint timing	3.45e-02	1.34E-01	4.40E-01	7.64E-01
	IVW-Robust	MR-Egger	MRMix	
Diurnal inactivity	7.88E-02	2.84E-01	2.26E-01	
L5 timing	9.29E-01	8.07E-01	5.79E-01	
M10 timing	4.16E-01	2.62E-01	3.39E-04	
Num. nocturnal sleep episodes	6.37E-01	2.44E-02	8.26E-01	
Sleep duration	5.53E-05	2.25E-01	4.67E-01	
Sleep duration variability	5.32E-02	3.45E-02	4.31E-01	
Sleep efficiency	1.01E-03	8.50E-01	7.30E-01	
Sleep midpoint timing	9.30E-02	6.47E-01	7.02E-02	

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Figure 1: **Simulation results on power comparisons under the genome-wide InSIDE assumption with balanced pleiotropy**. The results are summarized based on the performance over 2000 simulated data sets under a given causal effect (theta) and sample size (N). Each simulated data set consists of summary statistics on 200,000 independent SNPs generated from two GWAS of equal sample size N. (a) N = 300k and $J = 26$, with J being the average number of SNPs significantly associated with the risk factor; (b) N = 400k and $J = 59$; (c) N = 450k and $J = 80$; and (d) N = 500k and $J = 104$.



Statistica Sinica: Preprint
doi:10.5705/ss.202021.0164

Figure 2: **Simulation results on power comparisons under the conditional InSIDE assumption with balanced pleiotropy**. The results are summarized based on the performance over 2000 simulated data sets under a given causal effect (theta) and sample size (N). Each simulated data set consists of summary statistics on 200,000 independent SNPs generated from two GWAS of equal sample size N. (a) $N = 300$k and $J = 26$, with J being the average number of SNPs significantly associated with the risk factor; (b) $N = 400$k and $J = 59$; (c) $N = 450$k and $J = 80$; and (d) $N = 500$k and $J = 104$.