Mendelian randomization (MR) has become a popular approach to study the effect of a modifiable exposure on an outcome by using genetic variants as instrumental variables. A challenge in MR is that each genetic variant explains a relatively small proportion of variance in the exposure and there are many such variants, a setting known as many weak instruments. To this end, we provide a full theoretical characterization of the statistical properties of the two popular estimators in MR, the inverse-variance weighted (IVW) estimator and the pre-screened IVW estimator with instruments selected from an independent selection dataset, under many weak instruments. We then propose a debiased IVW estimator, a simple modification of the IVW estimator, that is robust to many weak instruments and doesn’t require pre-screening. Additionally, we present two instrument selection methods to improve efficiency of the new estimator when a selection dataset is available. An extension of the debiased IVW estimator to handle balanced horizontal pleiotropy is also discussed. We conclude by demonstrating our results in simulated and real datasets.
First, when SNPs are *common genetic variants*, i.e., their minor allele frequencies (MAF) are greater than 0.05 [21, 16], they may have small effects on the exposure. Second, when SNPs are *rare variants*, i.e., their MAF are less than 0.05, they may have small or modest effects on the exposure, but their genetic variances are also small so that their estimated effects may be imprecise. Third, there are often many SNPs that have zero/null effects on the exposure.

In this article, we focus on a popular setup in MR known as two-sample summary-data MR, where two sets of summary statistics are obtained from two GWASs [29]. The first set consists of estimated marginal associations between $p$ SNPs and the exposure derived from one GWAS, denoted as $\hat{\gamma}_j$, $j = 1, \ldots, p$. The second set consists of estimated marginal associations between the same $p$ SNPs and the outcome derived from another GWAS, denoted as $\hat{\Gamma}_j$, $j = 1, \ldots, p$. In MR, the most popular estimator of the exposure effect on the outcome, denoted as $\beta_0$, is the inverse-variance weighted (IVW) estimator [10]

\[
\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{p} w_j \hat{\beta}_j}{\sum_{j=1}^{p} w_j}, \quad \hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}, \quad w_j = \frac{\hat{\gamma}_j^2}{\sigma_{Yj}^2}.
\]

Here, $\sigma_{Yj}$ is the standard error of $\hat{\Gamma}_j$, $j = 1, \ldots, p$. Despite its widespread usage, very little is known about the theoretical properties of this estimator, such as its consistency or asymptotic normality, in common MR setups where there are many weak IVs. It is known that if a SNP is a weak IV, the ratio estimator $\hat{\beta}_j$ is biased and its distribution is not well approximated by a normal distribution [31]. This raises serious questions on the validity of the IVW estimator $\hat{\beta}_{IVW}$, a weighted combination of $\hat{\beta}_j$, as well as other modern MR methods using $\hat{\beta}_j$'s [10, 6, 7, 22].

To reduce concerns from weak IVs, a widely adopted practice in MR is to only include SNPs that pass the genome-wide significance threshold (i.e., p-value associated with $\hat{\gamma}_j$ is less than $5 \times 10^{-8}$ [42]) in the IVW estimator, say

\[
\hat{\beta}_{\lambda,IVW} = \frac{\sum_{j \in S_\lambda} w_j \hat{\beta}_j}{\sum_{j \in S_\lambda} w_j}, \quad S_\lambda = \{k: \text{SNP } k \text{ is significant}\}.
\]

More details are given in Section 3.2. While the use of genome-wide significance threshold is well-justified in controlling false discovery rates from testing millions of SNPs [17, 28], there has not been an investigation on using the same strategy to address the issue of weak IVs in MR studies.
1.2. Prior Work and Our Contributions. Prior work on weak IVs in MR is vast, but mostly limited to numerical studies [12, 13, 14, 29, 10]. In econometrics, the issue of weak IVs has been studied, but the results are limited to one-sample individual-data settings; see Stock, Wright and Yogo [35] and Andrews and Stock [3] for surveys. Recent works by Zhao et al. [40, 41] and Bowden et al. [9] proposed new estimators robust to many weak IVs in two-sample summary-data settings. Also, work by Wang and Kang [38] proposed new tests when the number of instruments is fixed, but the instruments are arbitrarily weak. But, to the best of our knowledge, no work has addressed the theoretical properties of the IVW estimator in (1.1) or its pre-screened counterpart in (1.2), arguably the most popular estimators in MR, under a common setting where there are many weak IVs.

Our overarching goal is to characterize the properties of the IVW estimator and to propose some improvements over estimators in (1.1)-(1.2). The main contributions can be divided into three parts.

1. We provide an asymptotic phase transition analysis of the IVW estimator (1.1) in terms of instruments’ average strength. We conduct a similar exercise for the pre-screened IVW estimator (1.2).

2. We propose a simple way to improve the IVW estimator (1.1) under many weak IVs, which we call the debiased IVW (dIVW) estimator. It is explicitly formulated as the IVW estimator multiplied by a bias correction factor; see equation (4.1). Unlike the IVW estimator, the dIVW estimator is robust to many weak IVs. In fact, even without pre-screening for strong IVs, the dIVW estimator is generally consistent and asymptotically normal. As such, the dIVW estimator does not need a third independent GWAS to select strong instruments while mitigating the “winner’s curse” bias [40, 10]. Finally, our dIVW estimator stands in contrast to recent optimization-based estimators (e.g., [41]) that are robust to many weak IVs, but are arguably more complex than the dIVW estimator. Specifically, these optimization-based estimators do not have explicit forms and may not have unique estimates in every data generating scenario.

3. To improve the efficiency of the dIVW estimator, we propose two data-driven methods to select “efficiency-increasing” SNPs for the dIVW estimator. The first one is straightforward and capable of eliminating IVs that have no effect on the exposure. The second one uses iterated thresholding and is able to construct the most efficient estimator in a given class.

The rest of this paper is organized as follow. Section 2 introduces notation,
setup, and assumptions. Section 3 characterizes the consistency and asymptotic normality of the IVW estimator (1.1) and its pre-screened counterpart (1.2). Section 4 proposes the dIVW estimator, gives its statistical properties, and Section 4.1 proposes data-driven IV selection methods. Section 5 studies an extension of dIVW estimator to balanced horizontal pleiotropy [33, 37, 23]. Results from simulation studies and a real data analysis are presented in Sections 6 and 7, respectively. The paper concludes with a discussion in Section 8. Technical proofs are in the supplementary material.

2. Notation, Setup, and Assumption. Following the two-sample summary-data MR literature [29, 8, 40], we consider $p$ SNPs, $Z_1, \ldots, Z_p$, where each SNP $Z_j$ takes on values 0, 1 or 2, and all SNPs are mutually independent from each other through linkage disequilibrium (LD) pruning or clumping done via software [24]. Let $X$ be the exposure and $Y$ be the continuous outcome. MR typically assumes the following models,

\begin{align}
X &= \sum_{j=1}^{p} \gamma_j Z_j + \eta_X U + E_X, \\
Y &= \beta_0 X + \eta_Y U + E_Y,
\end{align}

where $U$ is an unmeasured confounder independent of $Z_1, \ldots, Z_p$, $E_X$ and $E_Y$ are mutually independent random noises that are also independent of $(Z_1, \ldots, Z_p, U)$, and $(\gamma_j, \eta_X, \eta_Y, \beta_0)$ are unknown parameters. The goal in an MR analysis is to estimate the effect of the exposure $X$ on the outcome $Y$, which is represented by $\beta_0$. Since $U$ is not observed and possibly related with $X$, estimating $\beta_0$ using model (2.2) with ordinary least squares leads to biased estimates. Instead, an MR approach to estimating $\beta_0$ typically posits both models (2.1) and (2.2) and makes three core assumptions [19, 32, 5, 27, 25]. The first assumption is that instruments are associated with the exposure $X$, which amounts to $\gamma_j$’s in model (2.1) not simultaneously equaling to zero. We call the set of instruments where $\gamma_j \neq 0$ to be relevant or non-null instruments and the set of instruments where $\gamma_j = 0$ to be null instruments. The second assumption is that instruments are independent of the unmeasured confounder $U$; this is encoded by assuming $U$ is independent of $Z_1, \ldots, Z_p$ in (2.1)-(2.2). The third and last core assumption is that instruments affect the outcome $Y$ only through the exposure $X$; this is true under (2.1)-(2.2) since (2.2) does not involve $Z_j$’s. However, this last assumption may be violated in some studies; see Section 5 for details. For more detailed discussions on the core assumptions, models, and their implication in MR, see Didelez and Sheehan [20] and Bowden et al. [8].
In classic IV settings, estimation of $\beta_0$ is based on $n$ independent and identically distributed (i.i.d.) observations of $(Z_1, \ldots, Z_p, X, Y)$. In two-sample MR, estimation is based on $n_X$ i.i.d. observations of $(X, Z_1, \ldots, Z_p)$ from the exposure dataset, and $n_Y$ i.i.d. observations of $(Y, Z_1, \ldots, Z_p)$ from the outcome dataset. The two datasets are assumed to be independent of each other and we never jointly observe $Y$ and $X$. In two-sample summary-data MR, which is the most popular data setting in MR and the setting considered in the paper, only summary statistics from the exposure and outcome datasets are observed. Specifically, from the exposure dataset, we observe $\hat{\gamma}_j$, the ordinary least square estimate from a linear regression of $X$ on $Z_j$, and its standard error $\sigma_{Xj}$, $j = 1, \ldots, p$. From the outcome dataset, we observe $\hat{\Gamma}_j$, the ordinary least square estimate from a linear regression of $Y$ on $Z_j$, and its standard error $\sigma_{Yj}$, $j = 1, \ldots, p$. Note that models (2.1)-(2.2) and the independence of instruments immediately imply that $\hat{\gamma}_j$ consistently estimates $\gamma_j$ and $\hat{\Gamma}_j$ consistently estimates $\beta_0 \gamma_j$ for each $j$.

We make the following assumptions about the summary statistics.

**Assumption 1.** For every $j = 1, \ldots, p$, $\sigma^2_{Xj}$ and $\sigma^2_{Yj}$ are known, $\hat{\gamma}_j \sim N(\gamma_j, \sigma^2_{Xj})$, $\hat{\Gamma}_j \sim N(\beta_0 \gamma_j, \sigma^2_{Yj})$, and $\{\hat{\gamma}_j, \hat{\Gamma}_j, j = 1, \ldots, p\}$ are mutually independent.

We briefly assess the plausibility of Assumption 1 in two-sample summary-data MR; see Zhao et al. [40] who also make the same assumption. Modern GWAS usually encompasses tens of thousands of participants and $n_X$ and $n_Y$ are on the order of $10^4$. Hence, the normality of $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ is plausible. For the same reason, the standard errors of these estimates are precisely estimated and assuming them to be known is a reasonable approximation. The two-sample MR data structure guarantees the independence of $\hat{\gamma}_j$’s and $\hat{\Gamma}_j$’s. Also, two-sample MR prunes/clumps SNPs to be far apart in genetic distance and each SNP only explains a very small proportion of the total variance in the exposure variable, making independence between $\hat{\gamma}_j$’s and independence between $\hat{\Gamma}_j$’s likely. Furthermore, if $Y$ is binary, Assumption 1 is a first-order local approximation of a logistic outcome model [40, 41]. In Section 5, we relax Assumption 1 to allow for balanced horizontal pleiotropy.

Next, we make the following assumption about the standard errors of the summary statistics.

**Assumption 2.** Let $v_j = \sigma_{Xj}/\sigma_{Yj}$. There are positive constants $c_\sigma$ and $c'_\sigma$ such that $c_\sigma \leq v_j \leq c'_\sigma$ for all $j = 1, \ldots, p$.

To explain Assumption 2, consider the formulas for the standard errors
of $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ from ordinary least squares, 

$$
\sigma^2_{Xj} = \frac{\text{Var}(X) - \gamma_j^2 \text{Var}(Z_j)}{n_X \text{Var}(Z_j)} \quad \text{and} \quad \sigma^2_{Yj} = \frac{\text{Var}(Y) - \beta_0^2 \gamma_j^2 \text{Var}(Z_j)}{n_Y \text{Var}(Z_j)}.
$$

The ratio of these two standard errors is 

$$
\nu^2_j = \frac{\sigma^2_{Xj}}{\sigma^2_{Yj}} = \frac{\text{Var}(X) - \gamma_j^2 \text{Var}(Z_j)}{\text{Var}(Y) - \beta_0^2 \gamma_j^2 \text{Var}(Z_j)} \frac{n_Y}{n_X}.
$$

If SNP $j$ is a rare variant, then $\text{Var}(Z_j)$ is small and hence, this ratio is about $n_Y n_X^{-1} \text{Var}(X)/\text{Var}(Y)$. When SNP $j$ is a common variant, the terms $\gamma_j^2 \text{Var}(Z_j)$ and $\beta_0^2 \gamma_j^2 \text{Var}(Z_j)$ are still small compared to $\text{Var}(X)$ and $\text{Var}(Y)$ because an individual SNP typically explains a small proportion of the total variance in the exposure and outcome variables. Thus, if the sample sizes $n_X$ and $n_Y$ are of the same order, Assumption 2 will hold in both cases.

We conclude this section by defining the average strength of $p$ IVs

$$
(2.3) \quad \kappa = \frac{1}{p} \sum_{j=1}^{p} \mu_j^2, \quad \mu_j = \frac{\gamma_j}{\sigma_{Xj}}.
$$

Note that $\mu_j$ can be viewed as a z-score of SNP $j$’s effect on $X$. If $\kappa$ is small, SNPs are, on average, weakly associated with the exposure or there are many rare variants. If $\kappa$ is large, SNPs are, on average, strongly associated with the exposure and there are many common variants. An unbiased estimator of $\kappa$ is the average of F-statistics minus 1.

$$
(2.4) \quad \hat{\kappa} = \frac{1}{p} \sum_{j=1}^{p} \frac{\hat{\gamma}_j^2}{\sigma_{Xj}^2} - 1.
$$

As we will see in later sections, the limiting values of $\kappa$ in (2.3) plays a key role in characterizing the properties of the IVW estimator and as such, the unbiased estimator of $\kappa$ can serve as a guideline in empirical applications to check the theoretical conditions underlying the properties of the IVW estimator.

3. Properties of the IVW Estimator.

3.1. IVW Estimator (1.1). We first study the consistency and asymptotic normality of the IVW estimator $\hat{\beta}_{IVW}$ in (1.1) under different limiting values of $\kappa$ defined in (2.3). In what follows, $\xrightarrow{P}$ denotes convergence in probability and $\xrightarrow{D}$ denotes convergence in distribution.
Theorem 3.1. Suppose Assumptions 1 and 2 hold. The following conclusions can be made about the IVW estimator.

(a) If \( \kappa/p \to \infty \), \( p \to \infty \) and \( \mu_j^2/(\kappa p) \to 0 \) for every \( j \), then the IVW estimator \( \hat{\beta}_{IVW} \) is consistent and asymptotically normal, i.e., \( \hat{\beta}_{IVW} \xrightarrow{p} \beta_0 \) and

\[
V_{IVW}^{-1/2} \left( \hat{\beta}_{IVW} - \beta_0 \right) \xrightarrow{D} N(0, 1),
\]

where

\[
V_{IVW} = \frac{\sum_{j=1}^{p} \left[ v_j^2 (\mu_j^2 + 1) + \beta_0^2 v_j^4 (\mu_j^2 + 2) \right]}{\left[ \sum_{j=1}^{p} v_j^2 (\mu_j^2 + 1) \right]^2}, \quad v_j = \frac{\sigma_X_j}{\sigma_Y_j}.
\]

The result still holds if we replace \( V_{IVW} \) by a plug-in consistent estimator

\[
\hat{V}_{IVW} = \frac{\sum_{j=1}^{p} \left[ v_j^2 \hat{\mu}_j^2 + \beta_{IVW}^2 v_j^4 (\hat{\mu}_j^2 + 1) \right]}{\left[ \sum_{j=1}^{p} v_j^2 \hat{\mu}_j^2 \right]^2}, \quad \hat{\mu}_j = \frac{\hat{\gamma}_j}{\sigma_X_j}.
\]

(b) If \( \kappa \to \infty \), then \( \hat{\beta}_{IVW} \xrightarrow{p} \beta_0 \).

(c) If \( \kappa \to c > 0 \) and \( p \to \infty \), then

\[
\hat{\beta}_{IVW} - \beta_0 \xrightarrow{p} 0, \quad \frac{\sum_{j=1}^{p} v_j^2 \mu_j^2}{\sum_{j=1}^{p} v_j^2 (\mu_j^2 + 1)} \xrightarrow{p} 0.
\]

(d) If \( \kappa \to 0 \) and \( p \to \infty \), then \( \hat{\beta}_{IVW} \xrightarrow{p} 0 \).

Consider part (a) of Theorem 3.1, which is the only regime where the IVW estimator is both consistent and asymptotically normal. The condition \( \mu_j^2/(\kappa p) \to 0 \) for every \( j \) says that a single SNP’s strength is small compared to the total IV strength \( \kappa p \), a reasonable assumption in MR studies. However, the condition \( \kappa/p \to \infty \) requires that the average IV strength \( \kappa \) diverges to infinity at a faster rate than \( p \). This is extremely unlikely in MR studies unless every SNP is very strong. For example, even if the magnitude of squared z-score \( \mu_j^2 \) grows linearly in \( p \) for each SNP \( j \), \( \kappa/p \) still converges to a constant and the condition fails. However, if \( \mu_j^2 \) grows quadratically in \( p \), the condition \( \kappa/p \to \infty \) holds. Unless this latter condition holds for every SNP, the current practice of inference using normal approximation of \( \hat{\beta}_{IVW} \) is questionable.
Part (b) of Theorem 3.1 shows that the IVW estimator is consistent when the average IV strength goes to infinity, \( \kappa \to \infty \); we remark that unlike other conditions, condition (b) does not necessarily require \( p \to \infty \) and as such, can also approximate non-growing \( p \) regimes. Using the relationship between F-statistics and \( \kappa \) at the end of Section 2, condition (b) indicates that the average of \( p \) F-statistics must be large for the IVW estimator to be consistent. We believe a typical MR study would fail this condition as SNPs in MR are not simultaneously strong and have high average F-statistics. Combining parts (a) and (b) stresses that the IVW estimator’s consistency and asymptotic normality strongly rely on having strong IVs.

Part (c) of Theorem 3.1 states that when the average IV strength \( \kappa \) converges to a nonzero constant, the IVW estimator is biased towards 0. This regime reflects MR studies where a majority of instruments are weak with small, but non-zero \( \mu_j \)’s. In part (d) of Theorem 3.1, when the average IV strength \( \kappa \) converges to zero, the IVW estimator converges to 0 in probability. This regime reflects MR studies where there are many weak and null IVs. We believe a typical MR study likely falls in regimes (c) and (d), suggesting that using the IVW estimator likely leads to a biased estimate of the exposure effect \( \beta_0 \).

We conclude the section by making two points. First, Theorem 3.1 provides a theoretical justification of the numerical observations in the literature concerning the property of the IVW estimator with weak instruments. In particular, we show that the IVW estimator requires stringent conditions for asymptotic normality that will typically not be possible in most MR studies. Having said that, it is not common in MR for an investigator to directly use the IVW estimator with thousands of SNPs. Instead, many MR investigators pick the strongest instruments through a pre-screening procedure and run the IVW estimator among the pre-screened instruments. The next section illustrates what happens to the IVW estimator when pre-screening is conducted beforehand.

3.2. IVW Estimator (1.2) with Pre-screening. It is common in MR studies to pre-screen for strong IVs and to use only the selected IVs in the IVW estimator. Typically, the pre-screening process selects IVs whose p-values associated with the effect estimates \( \hat{\gamma}_j \)'s are below the GWAS p-value threshold of \( 5 \times 10^{-8} \). This pre-screening procedure is equivalent to selecting IVs whose estimated z-scores associated with the effect estimates \( \hat{\gamma}_j \)'s are above the z-score threshold \( \lambda \approx 5.45 \); this is based on finding the critical value of a standard normal distribution where its two-sided tail probability equals to \( 5 \times 10^{-8} \). To avoid selection bias or the “winner’s curse”, it is usually rec-
ommended to use a third independent dataset, called the selection dataset, solely for pre-screening instruments \[10, 40\]. This section studies the properties of the IVW estimator (1.2) with pre-screening from a selection dataset.

Formally, the IVW estimator (1.2) with pre-screening can be expressed as a hard-thresholding estimator with z-score threshold \( \lambda \geq 0 \),

\[
\hat{\beta}_{\lambda,\text{IVW}} = \frac{\sum_{j=1}^{p} \hat{\Gamma}_j \hat{\gamma}_j \sigma_{\hat{\gamma}_j}^{-2} I(|\hat{\gamma}_j^*| > \lambda \sigma_{\hat{\gamma}_j}^*)}{\sum_{j=1}^{p} \hat{\gamma}_j^2 \sigma_Y^{-2} I(|\hat{\gamma}_j^*| > \lambda \sigma_{\hat{\gamma}_j}^*)}.
\]

Here, the asterisks denote estimates coming from the selection dataset. For example, \( \hat{\gamma}_j^* \) and \( \sigma_{\hat{\gamma}_j}^* \) are estimates derived from the selection dataset with \( \hat{\gamma}_j^* \sim N(\gamma_j, \sigma_{\hat{\gamma}_j}^2) \). If \( \lambda = 0 \), \( \hat{\beta}_{\lambda,\text{IVW}} \) becomes the original IVW estimator \( \hat{\beta}_{\text{IVW}} \) in equation (1.1).

To incorporate the selection dataset into our analysis, we make assumptions similar to Assumptions 1 and 2 that include the selection dataset.

**Assumption 1’.** For every \( j = 1, \ldots, p \), \( \sigma_{\hat{\gamma}_j}^2, \sigma_Y^2, \sigma_{\hat{\gamma}_j}^* \) are known, \( \hat{\gamma}_j \sim N(\gamma_j, \sigma_{\hat{\gamma}_j}^2) \), \( \hat{\Gamma}_j \sim N(\beta_0 \gamma_j, \sigma_Y^2) \), \( \hat{\gamma}_j^* \sim N(\gamma_j, \sigma_{\hat{\gamma}_j}^2) \), and \( \{\hat{\gamma}_j, \hat{\Gamma}_j, \hat{\gamma}_j^*, j = 1, \ldots, p\} \) are mutually independent.

**Assumption 2’.** There are positive constants \( c_\sigma, c_\sigma’ \) such that \( c_\sigma \leq \sigma_{\hat{\gamma}_j}/\sigma_Y \leq c_\sigma’ \) and \( c_\sigma \leq \sigma_{\hat{\gamma}_j}/\sigma_{\hat{\gamma}_j}^* \leq c_\sigma’ \) for all \( j = 1, \ldots, p \).

Let \( \Phi(s) \) be the cumulative distribution function of the standard normal distribution and let \( q_{\lambda,j}^* = P(|\hat{\gamma}_j^*| > \lambda \sigma_{\hat{\gamma}_j}^*) = \Phi(\mu_j^* - \lambda) + \Phi(-\mu_j^* - \lambda) \) be the probability that instrument \( j \) exceeds the threshold \( \lambda \) and is included in the IVW estimator. Let \( \mu_j^* = \gamma_j/\sigma_{\hat{\gamma}_j} \) be the z-score of instrument \( j \) from the selection dataset, and let \( \hat{\mu}_j^* = \hat{\gamma}_j^*/\sigma_{\hat{\gamma}_j} \) be the estimated z-score. Similar to the IVW estimator, we use the expected average strength of IVs under selection

\[
\kappa^*_\lambda = \frac{\sum_{j=1}^{p} \mu_j^2 q_{\lambda,j}^*}{\sum_{j=1}^{p} q_{\lambda,j}^*},
\]

to characterize the behavior of the pre-screened IVW estimator. We also use \( p_{\lambda}^* = \sum_{j=1}^{p} q_{\lambda,j}^* \), the expected number of SNPs that are selected during the pre-screening process, to represent the effective number of IVs used in the pre-screened IVW estimator (3.1). In particular, if \( \lambda = 0 \) so that all the IVs are included in the IVW estimator, \( q_{\lambda,j}^* \), \( \kappa^*_\lambda \), and \( p_{\lambda}^* \) become 1, \( \kappa^* \), and \( p^* \), respectively.

We can obtain an unbiased estimator of \( p_{\lambda}^* \) by counting the number of IVs selected, \( \hat{p}_{\lambda}^* = \sum_{j=1}^{p} I(|\hat{\gamma}_j^*| > \lambda \sigma_{\hat{\gamma}_j}^*) \). We can also obtain a first-order
approximation of \( \kappa_\lambda^* \) by taking the average of F-statistics among IVs that are selected from the selection dataset minus 1, i.e.,

\[
\hat{\kappa}_\lambda^* = \frac{1}{p} \sum_{j=1}^{p} \mu_j^2 I(|\hat{\gamma}_j^*| > \lambda \sigma_{X,j}^*) - 1.
\]

The following theorem states the conditions under which the pre-screened IVW estimator in (3.1) is consistent and asymptotically normal.

**Theorem 3.2.** Suppose Assumptions 1′ and 2′ hold. For a given threshold \( \lambda \), the following conclusions can be made about the pre-screened IVW estimator.

(a) If \( \kappa_\lambda^*/p_\lambda^* \to \infty, \kappa_\lambda^*/\sqrt{p_\lambda^*}/\lambda^2 \to \infty, p \to \infty \) and \( \mu_j^2 q_{\lambda,j}^*/\kappa_\lambda^* p_\lambda^* \to 0 \) for every \( j \), then the pre-screened IVW estimator \( \hat{\beta}_{\lambda,IVW} \) is consistent and asymptotically normal, i.e., \( \hat{\beta}_{\lambda,IVW} \xrightarrow{P} \beta_0 \) and

\[
V_{\lambda,IVW}^{-1/2} \left( \hat{\beta}_{\lambda,IVW} - \beta_0 \right) \xrightarrow{D} N(0,1),
\]

where

\[
V_{\lambda,IVW} = \sum_{j=1}^{p} \left[ v_j^2 (\mu_j^2 + 1) q_{\lambda,j}^* + \beta_0^2 v_j^4 (\mu_j^2 + 3) q_{\lambda,j}^* - \beta_0^2 v_j^4 q_{\lambda,j}^* \right] \left[ \sum_{j:|\hat{\gamma}_j^*|>\lambda} v_j^2 (\mu_j^2 + 1) q_{\lambda,j}^* \right]^2.
\]

The result still holds if we replace \( V_{\lambda,IVW} \) by a plug-in consistent estimator

\[
\hat{V}_{\lambda,IVW} = \sum_{j:|\hat{\gamma}_j^*|>\lambda} v_j^2 \hat{\mu}_j^2 \left[ v_j^2 \hat{\mu}_j^2 + \beta_0^2 v_j^4 (\hat{\mu}_j^2 + 1) \right] \left[ \sum_{j:|\hat{\gamma}_j^*|>\lambda} v_j^2 \hat{\mu}_j^2 \right]^2.
\]

(b) If \( \kappa_\lambda^* \to \infty \) and \( \kappa_\lambda^*/\sqrt{p_\lambda^*/\lambda^2} \to \infty \), then \( \hat{\beta}_{\lambda,IVW} \xrightarrow{P} \beta_0 \).

(c) If \( \kappa_\lambda^* \to \infty \) and \( \sqrt{p_\lambda^*/\max(1,\lambda^2)} \to \infty \), then

\[
\hat{\beta}_{\lambda,IVW} - \beta_0 \frac{\sum_{j=1}^{p} \mu_j^2 v_j^2 q_{\lambda,j}^*}{\sum_{j=1}^{p} (\mu_j^2 + 1) v_j^2 q_{\lambda,j}^*} \xrightarrow{P} 0.
\]

(d) If \( \kappa_\lambda^* \to 0 \) and \( \sqrt{p_\lambda^*/\max(1,\lambda^2)} \to \infty \), then \( \hat{\beta}_{\lambda,IVW} \xrightarrow{P} 0 \).

Each part of Theorem 3.2 mirrors each part of Theorem 3.1. For example, part (a) of Theorem 3.2 states the conditions under which the pre-screened IVW estimator with threshold \( \lambda \) is consistent and asymptotically normal.
The condition $\kappa^*_{\lambda}/p^*_{\lambda} \to \infty$, the counterpart of $\kappa/p \to \infty$ in part (a) of Theorem 3.1, is reasonable when all selected SNPs are strong such that their expected average IV strength $\kappa^*_\lambda$ is going to infinity faster than $p^*_\lambda$. The condition $\mu_j^2 q^*_\lambda,j/(\kappa^*_\lambda p^*_\lambda) \to 0$ for every $j$, like condition $\mu_j^2/(\kappa p) \to 0$ in part (a) of Theorem 3.1, is reasonable if a selected IV's strength is small compared to the total expected IV strength among selected instruments. Part (b) of Theorem 3.2 states the conditions under which the pre-screened IVW estimator is consistent, i.e., $\kappa^*_\lambda \to \infty$, and mirrors the condition $\kappa \to \infty$ in Theorem 3.1(b). Parts (c) and (d) of Theorem 3.2 refer to settings where the pre-screened IVW estimator is biased towards zero. Broadly speaking, these scenarios occur if the pre-screening threshold $\lambda$ is not large enough so that many weak and null IVs are selected, or every IV in the selection dataset is weak. Theorem 4.2 also has an additional assumption involving the selection threshold $\lambda$, which essentially controls the denominator of the IVW estimator after pre-screening.

Comparing the IVW estimator (1.1) to the pre-screened IVW estimator (1.2), the IVW estimator requires far more stringent conditions on IV strength (i.e., all $p$ IVs must be strong) to guarantee consistency or asymptotic normality. The pre-screened IVW estimator also requires that all selected IVs must be strong, but this condition may be satisfied if we carefully choose a threshold $\lambda$. For example, for consistency, we have to choose $\lambda$ so that the average strength of selected IVs $\hat{\kappa}^*_\lambda$ is large and the product of $\hat{\kappa}^*_\lambda$ and the square root of the number of selected instruments $\sqrt{\hat{p}^*_\lambda}$ is much larger than $\lambda^2$.

Unfortunately, the dependence of the pre-screened IVW estimator on the tuning parameter $\lambda$ makes it difficult to verify the pre-screened IVW estimator’s properties in practice. For example, consider the common practice of selecting IVs that pass the p-value threshold of $5 \times 10^{-8}$, which as mentioned earlier is equivalent to setting $\lambda \approx 5.45$. This $\lambda$ may or may not satisfy the conditions for consistency or asymptotic normality of $\hat{\beta}_{\lambda,IVW}$ and we highlight some examples below; Table 1 in Section 6 provides numerical illustrations.

1. If all the non-null IVs are strong and common in the sense that all $\mu_j^2$’s are of order $n_X$, $p$ is in the millions, and the number of samples in the exposure dataset $n_X$ is much larger than the log number of available SNPs, i.e., log $p$, conditions in Theorem 3.2(a) may hold and the pre-screened IVW estimator may be asymptotically normal.

2. If all IVs are very weak, for example $\mu_j^2 \leq c$ for all $j$ and some positive constant $c$, then $\kappa^*_\lambda$ is bounded regardless the choice of $\lambda$. In this case, the GWAS threshold may select weak IVs and bias the pre-screened
dIVW estimator.

3. If every IV strength equals $\lambda$, then $q^*_j \approx 1/2$ and $\kappa^*_\lambda \approx \lambda^2$. But, $\kappa^*_\lambda/p^*_\lambda \approx 2\lambda^2/p$ may be small, implying that Theorem 3.2(a) may not hold.

While the pre-screened dIVW requires less stringent assumptions than the IVW estimator, we believe that finding a selection threshold $\lambda$ and checking that the selected $\lambda$ satisfies these conditions will be cumbersome. The next section provides a way to remedy this and estimate $\beta_0$ without requiring a selection threshold $\lambda$.

4. Debiased IVW Estimator. Motivated by the unrealistic assumptions underlying the consistency and asymptotic normality of the IVW estimator as well as the need to carefully choose a threshold $\lambda$ in the pre-screened IVW estimator, we propose a simple estimator that does not rely on neither. We name the new estimator as the debiased IVW (dIVW) estimator. It is the original IVW estimator multiplied by a bias correction factor, i.e.,

$$
\hat{\beta}_{\text{dIVW}} = \hat{\beta}_{\text{IVW}} \cdot \frac{\sum_{j=1}^{p} w_j}{\sum_{j=1}^{p} (w_j - v_j^2)}.
$$

The bias correction factor $\sum_{j=1}^{p} w_j / \sum_{j=1}^{p} (w_j - v_j^2)$ has an explicit form, where $w_j = \hat{\gamma}_j^2 \sigma_{Y_j}^2$ and $v_j = \sigma_{X_j} / \sigma_{Y_j}$. Simple algebra reveals that this bias correction factor essentially replaces and amplifies the denominator of the IVW estimator in (1.1),

$$
\hat{\beta}_{\text{dIVW}} = \hat{\beta}_{\text{IVW}} \cdot \frac{\sum_{j=1}^{p} w_j}{\sum_{j=1}^{p} (w_j - v_j^2)} = \frac{\sum_{j=1}^{p} w_j \hat{\beta}_j}{\sum_{j=1}^{p} (w_j - v_j^2)} = \frac{\sum_{j=1}^{p} \hat{\gamma}_j \hat{\gamma}_j \sigma_{Y_j}^2}{\sum_{j=1}^{p} (\hat{\gamma}_j^2 - \sigma_{X_j}^2) \sigma_{Y_j}^2}.
$$

Surprisingly, this simple correction to the IVW estimator makes it dramatically more robust to many weak IVs. To see why, consider the difference between the dIVW estimator and the true value $\beta_0$,

$$
\hat{\beta}_{\text{dIVW}} - \beta_0 = \frac{\sum_{j=1}^{p} (\hat{\gamma}_j - \beta_0 \hat{\gamma}_j^2 + \beta_0 \sigma_{X_j}^2) \sigma_{Y_j}^2}{\sum_{j=1}^{p} (\hat{\gamma}_j^2 - \sigma_{X_j}^2) \sigma_{Y_j}^2}.
$$
The numerator of the above expression has exactly mean zero, i.e.,

\[ E\{(\hat{\Gamma}_j \hat{\gamma}_j - \beta_0 \hat{\gamma}_j^2 + \beta_0 \sigma_{Xj}^2)\sigma_{Yj}^{-2}\} = (\beta_0 \gamma_j^2 - \beta_0 \gamma_j^2 - \beta_0 \sigma_{Xj}^2 + \beta_0 \sigma_{Xj}^2)\sigma_{Yj}^{-2} = 0, \]

where the first line follows from \( E(\hat{\Gamma}_j) = \beta_0 \gamma_j \), \( E(\hat{\gamma}_j) = \gamma_j \), \( E(\hat{\gamma}_j^2) = \gamma_j^2 + \sigma_{Xj}^2 \) and the independence between \( \hat{\Gamma}_j \) and \( \hat{\gamma}_j \). The denominator converges to a positive probability limit under mild conditions; see Theorem 4.1 below. Thus, we have \( \hat{\beta}_{dIVW} \overset{P}{\to} \beta_0 \). In contrast, the difference between the IVW estimator \( \hat{\beta}_{IVW} \) and the true value \( \beta_0 \) has the form

\[ \hat{\beta}_{IVW} - \beta_0 = \frac{\sum_{j=1}^p (\hat{\Gamma}_j \hat{\gamma}_j - \beta_0 \hat{\gamma}_j^2) \sigma_{Yj}^{-2}}{\sum_{j=1}^\gamma_j^2 \sigma_{Yj}^{-2}}. \]

The expectation of the numerator is

\[ E \left[ \sum_{j=1}^p (\hat{\Gamma}_j \hat{\gamma}_j - \beta_0 \hat{\gamma}_j^2) \right] = \sum_{j=1}^p (\beta_0 \gamma_j^2 - \beta_0 \gamma_j^2 - \beta_0 \sigma_{Xj}^2)\sigma_{Yj}^{-2} = -\beta_0 \sum_{j=1}^p \sigma_{Xj}^2 \sigma_{Yj}^{-2}, \]

which is not 0 unless \( \beta_0 = 0 \). Thus, in order for the IVW estimator to be consistent, the leftover term in the expectation has to be much smaller than the denominator, or formally, \( \sum_{j=1}^p \hat{\gamma}_j^2 \sigma_{Yj}^{-2} \) has to dominate \( \sum_{j=1}^p \sigma_{Xj}^2 \sigma_{Yj}^{-2} \); this is also the condition \( \kappa \to \infty \) in Theorem 3.1(b).

The following theorem formalizes the above observations and states the properties of the dIVW estimator (4.1).

**Theorem 4.1.** Suppose Assumptions 1 and 2 hold. The following conclusions can be made about the dIVW estimator.

(a) If \( \kappa \sqrt{p} \to \infty \), \( p \to \infty \) and \( \mu_j^2 / (\kappa r + p) \to 0 \) for every \( j \), then the dIVW estimator is consistent and asymptotically normal, i.e., \( \hat{\beta}_{dIVW} \overset{P}{\to} \beta_0 \) and

\[ V_{dIVW}^{-1/2} \left( \hat{\beta}_{dIVW} - \beta_0 \right) \overset{D}{\to} N(0, 1), \]

where

\[ V_{dIVW} = \sum_{j=1}^p \left[ v_j^2 (\mu_j^2 + 1) + \beta_0^2 v_j^2 (\mu_j^2 + 2) \right] / \left( \sum_{j=1}^p v_j^2 \mu_j^2 \right)^2. \]
The result still holds if $V_{dIVW}$ is replaced with a plug-in consistent estimator
\[
\hat{V}_{dIVW} = \frac{\sum_{j=1}^{p} \left[ v_{j}^{2} \tilde{\mu}_{j}^{2} + \hat{\beta}_{dIVW}^{2} v_{j}^{4} (\tilde{\mu}_{j}^{2} + 1) \right]}{\left[ \sum_{j=1}^{p} v_{j}^{2} (\tilde{\mu}_{j}^{2} - 1) \right]^{2}}.
\]

(b) If we only assume $\kappa \sqrt{p} \to \infty$, then the dIVW estimator is consistent, i.e., $\hat{\beta}_{dIVW} \overset{p}{\to} \beta_{0}$.

Examining the conditions in Theorem 4.1, the dIVW estimator is more robust to weak IVs compared to the original IVW estimator. For example, the main conditions in Theorem 4.1(a), $\kappa \sqrt{p} \to \infty$ and $p \to \infty$, hold even when $\kappa$ is bounded; it even holds if $\kappa \to 0$ but at a slower rate than $1/\sqrt{p}$. In contrast, the original IVW estimator is consistent and asymptotically normal when $\kappa/p \to \infty$ and $p \to \infty$. Furthermore, if IVs are common variants, but are weak in the sense of Staiger and Stock [34] (i.e., $\gamma_{j}$ and $\sigma_{Xj}$ are both of the order $n^{-1/2}$), the dIVW estimator still remains consistent and asymptotically normal as $p \to \infty$. We remark that the consistency condition in part (b) does not necessarily require $p \to \infty$ and as such, can also approximate non-growing $p$ regimes.

The rate condition $\kappa \sqrt{p} \to \infty$ in Theorem 4.1 can be interpreted as an effective sample size for the dIVW estimator and can be estimated by $\hat{\kappa} \sqrt{p}$ with $\hat{\kappa}$ defined in (2.4). In our simulation studies (i.e., Figure 1), we provide some guidelines on what would be considered a large value of $\hat{\kappa} \sqrt{p}$ for the asymptotics promised in Theorem 4.1 to kick in. This is akin to qualitative guidelines on what would be a large enough $n$ for a normal approximation of an estimator to hold. The rate condition is also related to conditions imposed by the limited information maximum likelihood (LIML) estimator in the one-sample individual-level data setting [15] and the robust adjusted profile score (MR-raps) estimator [40]. We remark that the latter requires all SNPs to be common variants whereas the dIVW estimator can handle both common and rare variants.

4.1. Improving Efficiency of dIVW with Pre-Screening. While the dIVW estimator is consistent and asymptotically normal even if many IVs are weak, a closer inspection of its variance $V_{dIVW}$ in Theorem 4.1 suggests that including null IVs increases variance. The goal of this section is to explore how to make the dIVW estimator more efficient by pre-screening for non-null IVs. We remark that in this section, we use pre-screening to improve efficiency of the dIVW estimator; in contrast, the IVW estimator uses pre-screening to reduce bias.
Formally, consider the dIVW estimator using only IVs selected from the selection dataset.

\[
\hat{\beta}_{\lambda, \text{dIVW}} = \frac{\sum_{p} \hat{\gamma}_{j} \hat{\sigma}_{Y_j}^2 I(|\hat{\gamma}_j^*| > \lambda \sigma_{X_j}^*)}{\sum_{j=1}^{p} (\hat{\gamma}_j^2 - \sigma_{X_j}^2) \hat{\sigma}_{Y_j}^2 I(|\hat{\gamma}_j^*| > \lambda \sigma_{X_j}^*)}.
\]

Theorem 4.2 states the properties of \( \hat{\beta}_{\lambda, \text{dIVW}} \).

**Theorem 4.2.** Suppose Assumptions 1’ and 2’ hold. For a given \( \lambda \), the following conclusions can be made about the pre-screened dIVW estimator.

(a) If \( \kappa \sqrt{p} \max(1, \lambda^2) \rightarrow \infty \), \( p \rightarrow \infty \) and \( \mu_j^2 q_{\lambda,j}/(\kappa^* p_{\lambda}^* + p_{\lambda}^*) \rightarrow 0 \) for every \( j \), then the pre-screened dIVW estimator \( \hat{\beta}_{\lambda, \text{dIVW}} \) is consistent and asymptotically normal, i.e., \( \hat{\beta}_{\lambda, \text{dIVW}} \overset{P}{\to} \beta_0 \) and

\[
V_{\lambda, \text{dIVW}}^{-1/2} \left( \hat{\beta}_{\lambda, \text{dIVW}} - \beta_0 \right) \overset{D}{\to} N(0,1),
\]

where

\[
V_{\lambda, \text{dIVW}} = \frac{\sum_{j=1}^{p} \hat{\mu}_j^2 (\mu_j^2 + 1) + \hat{\beta}_{\lambda, \text{dIVW}}^2 (\mu_j^2 + 2) \hat{q}_{\lambda,j}}{\left[ \sum_{j=1}^{p} \hat{v}_j^2 \hat{\mu}_j^2 \hat{q}_{\lambda,j} \right]^2}.
\]

The same result holds if we replace \( V_{\lambda, \text{dIVW}} \) by a plug-in consistent estimator

\[
\hat{V}_{\lambda, \text{dIVW}} = \frac{\sum_{j: |\hat{\mu}_j^*| > \lambda} \hat{v}_j^2 \hat{\mu}_j^2 + \hat{\beta}_{\lambda, \text{dIVW}}^2 (\hat{\mu}_j^2 + 1)}{\left[ \sum_{j: |\hat{\mu}_j^*| > \lambda} \hat{v}_j^2 (\hat{\mu}_j^2 - 1) \right]^2}.
\]

(b) If we only assume \( \kappa \sqrt{p} \max(1, \lambda^2) \rightarrow \infty \), then the pre-screened dIVW estimator is consistent, i.e., \( \hat{\beta}_{\lambda, \text{dIVW}} \overset{P}{\to} \beta_0 \).

We remark that the conditions for the pre-screened dIVW estimator \( \hat{\beta}_{\lambda, \text{dIVW}} \) are weaker than those for the pre-screened IVW estimator \( \hat{\beta}_{\lambda, \text{IVW}} \) stated in Theorem 3.2. For example, the condition for consistency of the pre-screened IVW estimator implies the condition \( \kappa \sqrt{p} \max(1, \lambda^2) \rightarrow \infty \) which is required for the consistency of pre-screened dIVW estimator. Also, the rate condition \( \kappa \sqrt{p} \max(1, \lambda^2) \rightarrow \infty \) acts like an effective sample size for the pre-screened dIVW estimator and can be estimated by \( \hat{\delta}_1 \sqrt{p_{\lambda}^*}/ \max(1, \lambda^2) \).

In our simulation studies, specifically Figure 1, we provide some guidelines on what would be considered a large effective sample size for the asymptotics promised in Theorem 4.2 to kick in.
The variance of the pre-screened dIVW estimator in Theorem 4.2 provides a basis to choose the threshold $\lambda$ to improve efficiency of the dIVW estimator. In general, the threshold $\lambda$ has to satisfy the conditions in Theorem 4.2 as well as reduce the variance $V_{\lambda,dIVW}$.

A natural choice is the threshold $\lambda = \sqrt{2 \log p}$ which guarantees that the probability of selecting any null instrument is very small.

\[
P(\text{at least one null IV is selected}) = 2(p - s)\Phi(-\lambda)
\leq 2(p - s)\frac{1}{\lambda\sqrt{2\pi}}e^{-\lambda^2/2}
= \frac{2}{\sqrt{2\pi}} \frac{1}{\sqrt{2\log p}} \frac{p - s}{p}
\to 0 \quad \text{as } p \to \infty
\]

For example, suppose we have an MR study where we have a few strong and common variant SNPs with $z$-scores on the order of $n_X$ and the other relevant SNPs satisfy $\mu_j^2 \geq \sqrt{2 \log p - c}$ for some constant $c$. Then, the condition in Theorem 4.2(a), $\kappa^*_{\lambda} \sqrt{p_{\lambda}} / \max(1, \lambda^2) \to \infty$ is satisfied by $n_X / (\sqrt{5} \log p) \to \infty$. We remark that $\lambda = \sqrt{2 \log p}$ is an improvement over the common cut-off based on the $p$-value threshold of $5 \times 10^{-8}$ (i.e., $\lambda \approx 5.45$) because $\lambda = \sqrt{2 \log p}$ adapts to the total number of IVs. More concretely, if the initial number of candidate SNPs in the selection dataset is smaller than a million, i.e., $p < 10^6$, then $\sqrt{2 \log p} < 5.45$ and the proposed pre-screening cutoff $\lambda = \sqrt{2 \log p}$ selects more IVs than the genome-wide $p$-value cutoff $5 \times 10^{-8}$. The net effect is that we can include more non-null IVs to improve the efficiency of dIVW estimator.

However, $\lambda = \sqrt{2 \log p}$ is not necessarily optimal in the sense that it also removes many weak, but non-null IVs. Indeed, the $\sqrt{2 \log p}$ threshold is mainly designed to minimize the probability of not selecting any null IVs; it isn’t designed to minimize the asymptotic variance of $\hat{\beta}_{\lambda,dIVW}$. Therefore, we propose another approach, which we call the MR-EO algorithm, to choose $\lambda$ that directly minimizes the asymptotic variance $V_{\lambda,dIVW}$; MR-EO stands for Mendelian Randomization Estimation-Optimization. In a nutshell, MR-EO considers a range of estimates $\hat{\beta}_{\lambda,dIVW}$ for $\lambda \in [0, \sqrt{2 \log p}]$, where $\lambda = 0$ amounts to including every IV and $\lambda = \sqrt{2 \log p}$ may select very few IVs. It then tries to find the optimal $\lambda$ in this range such that the asymptotic variance is minimized.

Formally, suppose $\kappa^*_{\lambda} \sqrt{p_{\lambda}} / \max(1, \lambda^2) \to \infty$ holds for every $\lambda$ in the range $[0, \sqrt{2 \log p}]$. Then, every $\hat{\beta}_{\lambda,dIVW}$ is consistent. Also, for a given $\lambda$, its vari-
The debiased IVW estimator is
\[
\hat{V}_{\lambda, \text{dIVW}}(\beta_0) = \frac{\sum_{j:|\hat{\mu}^*_j| > \lambda} v^2_j \hat{\mu}^2_j + \beta_0^2 v^4_j (\hat{\mu}^2_j + 1)}{\left[ \sum_{j:|\hat{\mu}^*_j| > \lambda} v^2_j (\hat{\mu}^2_j - 1) \right]^2}.
\]

Then, the most efficient estimator \( \hat{\beta}_{\lambda, \text{dIVW}} \) among \( \lambda \in [0, \sqrt{2 \log p}] \) uses \( \lambda \) that minimizes \( \hat{V}_{\lambda, \text{dIVW}}(\beta_0) \). However, \( \hat{V}_{\lambda, \text{dIVW}}(\beta_0) \) can only be computed with a known \( \beta_0 \). MR-EO resolves this by alternating between estimating the exposure effect \( \beta_0 \) (i.e., the E-Step), and finding the optimal \( \lambda \) given an estimated \( \beta_0 \) (i.e., the O-Step); see Algorithm 1 for details.

Initialize \( t = 0, t_{\text{max}}, \lambda_0 = \sqrt{2 \log p}, V = \infty; \)
while \( t \leq t_{\text{max}} \) do
  E-Step: for a given \( \lambda_t \), estimate \( \hat{\beta}_0 \) with the dIVW estimator \( \hat{\beta}_{\lambda_t, \text{dIVW}} \);
  if \( V \leq \hat{V}_{\lambda_t, \text{dIVW}}(\hat{\beta}_{\lambda_t, \text{dIVW}}) \) then
    exit the while loop;
  else
    \( V = \hat{V}_{\lambda_t, \text{dIVW}}(\hat{\beta}_{\lambda_t, \text{dIVW}}) \);
  end
  O-Step: Plug \( \hat{\beta}_{\lambda_t, \text{dIVW}} \) into the variance estimator and find
  \( \lambda_{t+1} = \arg \min_{\lambda \in [0, \sqrt{2 \log p}]} \hat{V}_{\lambda, \text{dIVW}}(\hat{\beta}_{\lambda, \text{dIVW}}) \)
  Set \( t = t + 1; \)
end
Output \( \lambda_{t-1} \).

**Algorithm 1**: MR-EO algorithm to determine the optimal \( \lambda \)

We make some comments regarding the implementation of MR-EO and its final output. First, we initialize MR-EO to \( \lambda_0 = \sqrt{2 \log p} \) and force the algorithm to stop at \( t = t_{\text{max}} \) with a reasonably large \( t_{\text{max}} \), mainly for computational efficiency. Second, if the estimator of \( \beta_0 \) is inconsistent for some values of \( \lambda \), the algorithm may fail to find the optimal \( \lambda \). To avoid this, we can empirically evaluate \( \hat{\kappa}_\lambda \sqrt{\hat{p}_\lambda}/\max(1, \lambda^2) \) and make sure this quantity is reasonably large for all \( \lambda \) in the specified range. In our simulation studies, we find that setting the range of \( \lambda \) to be \( 0 \) and \( \sqrt{2 \log p} \) works well. Other range of \( \lambda \) besides \( [0, \sqrt{2 \log p}] \) can also be used so long as all the estimators with \( \lambda \) in the range is consistent for \( \beta_0 \). Third, there is no theoretical guarantee that the estimator chosen by MR-EO is asymptotically normal. Nevertheless, in our simulation studies in Section 6, we find that the estimator chosen by MR-EO is consistent and the constructed confidence interval has desirable coverage probability.
5. Balanced Horizontal Pleiotropy. In this section, we study the property of the dIVW estimator under one type of pleiotropy in MR, balanced horizontal pleiotropy [23, 40, 8]. Under balanced horizontal pleiotropy, the third core IV assumption described in Section 2 is violated. Specifically, the pleiotropic effects of p SNPs $\alpha_1, \ldots, \alpha_p$ are random effects that are normally distributed with mean zero. These effects are added into the relationship between $Y$, $X$, and $Z_j$'s.

\begin{equation}
Y = \beta_0 X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y U + E_Y.
\end{equation}

In short, under balanced horizontal pleiotropy, model (5.1) replaces (2.2) and model (2.1) remains the same as before. The random effects $\alpha_j$'s are independent of $X$, $Z_j$'s, $U$, $E_Y$ and $E_X$.

To incorporate balanced pleiotropy, we replace Assumptions 1' and 2' with the following assumptions.

**Assumption 3.** For every $j = 1, \ldots, p$, $\sigma^2_X j$, $\sigma^2_Y j$, $\sigma^2_{X j}$ are known, $\hat{\gamma}_j \sim N(\gamma_j, \sigma^2_{X j})$, $\hat{\Gamma}_j \sim N(\alpha_j + \beta_0 \gamma_j, \sigma^2_{Y j})$, $\hat{\gamma}^*_j \sim N(\gamma_j, \sigma^2_{X j})$, $\alpha_j \sim N(0, \tau_0^2)$, and the variables in the set \{\hat{\gamma}_j, \hat{\Gamma}_j, \hat{\gamma}^*_j, j = 1, ..., p\} are mutually independent.

**Assumption 4.** There are positive constants $c_\sigma, c'_\sigma, c_\tau$ such that $c_\sigma \leq \sigma_{X j}/\sigma_{Y j} \leq c'_\sigma$, $c_\sigma \leq \sigma_{X j}/\sigma^*_{X j} \leq c'_\sigma$ and $\tau_0 \leq c_\tau \sigma_{Y j}$ for all $j$.

Theorem 5.1 shows that the pre-screened dIVW estimator remains consistent and asymptotically normal under balanced horizontal pleiotropy.

**Theorem 5.1.** Suppose Assumptions 3 and 4 hold. For a given $\lambda$, the following conclusions can be made about the pre-screened dIVW estimator.

(a) If $\kappa^*_\lambda p_X^*/\max(1, \lambda^2) \to \infty$, $p \to \infty$ and $\mu_j^2 q^*_{\lambda, j} / (\kappa^*_\lambda p_X^* + p^*_\lambda) \to 0$ for every $j$, then the pre-screened dIVW estimator is consistent and asymptotically normal, i.e., $\hat{\beta}_{\lambda, \text{dIVW}} \xrightarrow{P} \beta_0$, and

$W_{\lambda, \text{dIVW}}^{-1/2} \left( \hat{\beta}_{\lambda, \text{dIVW}} - \beta_0 \right) \xrightarrow{D} N(0, 1),$

where

\[
W_{\lambda, \text{dIVW}} = \sum_{j=1}^{p} \left[ v_j^2 (1 + \tau_0^2 \sigma^*_Y) (\mu_j^2 + 1) + \beta_0^2 v_j^4 (\mu_j^2 + 2) \right] q^*_{\lambda, j}^2 - \frac{q^*_{\lambda, j}^2}{\sum_{j=1}^{p} v_j^2 \mu_j^2 q^*_{\lambda, j}^2}.
\]
Additionally, if there exists a constant \( c > 0 \) such that

\[
\frac{p\sigma_{Y_j}^2}{\sigma_{Y_k}^2} \leq c \sum_{k=1}^{p} \sigma_{Y_k}^2 \quad \text{for every } j,
\]

and \( \kappa \sqrt{p} \to \infty \), then the same result holds if we replace \( W_{\lambda,dIVW} \) with a plug-in consistent estimator

\[
\hat{W}_{\lambda,dIVW} = \sum_{j:|\hat{\mu}_j^*| > \lambda} \left[ \frac{\nu_j^2 (1 + \hat{\tau}_j^2 \sigma_{Y_j}^2) \hat{\mu}_j^2 + \hat{\beta}_{\lambda,dIVW}^2 \nu_j^4 (\hat{\mu}_j^2 + 1)}{\sum_{j:|\hat{\mu}_j^*| > \lambda} \nu_j^2 (\hat{\mu}_j^2 - 1)^2} \right],
\]

\[
\hat{\tau}^2 = \frac{\sum_{j=1}^{p} \left[ (\hat{\Gamma}_j - \hat{\beta}_{dIVW} \hat{\gamma}_j)^2 - \sigma_{Y_j}^2 - \beta_{dIVW}^2 \sigma_{X_j}^2 \right] \sigma_{Y_j}^{-2}}{\sum_{j=1}^{p} \sigma_{Y_j}^{-2}}.
\]

(b) If we only assume \( \kappa^*\sqrt{\lambda^*}/\max(1,\lambda^2) \to \infty \), then the pre-screened dIVW estimator is consistent, i.e., \( \hat{\beta}_{\lambda,dIVW} \xrightarrow{P} \beta_0 \).

We remark that Theorem 5.1 recovers the dIVW estimator without pre-screening by setting \( \lambda = 0 \). The major difference between the performance of the dIVW estimator without balanced pleiotropy and with balanced pleiotropy is in the variance \( W_{\lambda,dIVW} \). In particular, Condition (5.2) is assumed to guarantee the consistency of the plug-in variance estimator \( \hat{W}_{\lambda,dIVW} \) and is reasonable if \( \sigma_{Y_j}^2 \)'s are not too different from each other. Note that we use the dIVW without pre-screening to estimate \( \hat{\tau}^2 \) for simplicity. Also, we can still use the aforementioned methods (e.g., MR-EO) to choose \( \lambda \) and improve efficiency.

6. Simulation Studies.

6.1. Comparison Between Estimators and Pre-Screening Thresholds. We conduct simulation studies to evaluate the finite sample behavior of the proposed estimators. To closely mirror what is done in practice, we take a real two-sample summary-level MR data from the BMI-CAD dataset in the \texttt{mr.raps} R package (version 0.3.1) of Zhao et al. [41] and use the values from the dataset as our simulation parameters. The dataset is used to estimate the effect of body mass index (BMI), i.e., \( X \), on the risk of coronary artery disease (CAD), i.e., \( Y \). There are three non-overlapping GWASs in the dataset:

1. Selection dataset: A GWAS for BMI in the Japanese population (sample size: 173,430) [2];
2. Exposure dataset: A GWAS for BMI in round 2 of the UK BioBank (sample size: 336,107) [1];
3. Outcome dataset: A GWAS for CAD from the CARDIoGRAMplusC4D consortium (sample size: ≈185,000), with genotype imputation using the 1000 Genome Project [36].

The three datasets have been cleaned so that (i) SNPs appear in all three datasets and (ii) SNPs are far apart in genetic distance; see Zhao et al. [41] for details. The data cleaning leads to \( p = 1119 \) SNPs available for analysis. Each GWAS contains publicly available summary statistics that include the coefficients and standard errors from marginal linear or logistic regression of the trait in question to each SNP. From the exposure dataset, the estimated average IV strength is \( \hat{\kappa} = 6.8 \). From the selection dataset, the estimated average IV strength is \( \hat{\kappa} = 7.7 \).

Let \( s \) denote the number of non-null instruments. We consider three simulation settings for the exposure and selection datasets that are plausible in an MR study. Broadly speaking, each of these settings change the number of null IVs from the set of 1119 SNPs.

**Case 1** (Some strong IVs, many null IVs): \( p = 1119, \ s = 20, \ \kappa = 2.90 \). For the \( s = 20 \) non-null instruments, we set the true marginal IV-exposure association parameters \( \gamma_j \)'s to be those from the top 20 genetic variants with the smallest \( p \)-values in the exposure dataset.

**Case 2** (Many weak IVs, many null IVs): \( p = 1119, \ s = 100, \ \kappa = 1.05 \). This setting is similar to Case 1, except we use the first 100 SNPs in the exposure dataset.

**Case 3** (Many weak IVs, no null IVs): \( p = s = 1119, \ \kappa = 7.78 \). This setting is similar to Case 1, except we use all 1119 SNPs.

For each setting, we generate the simulated exposure and selection datasets based on Assumptions 1’ and 2’. Specifically, for non-null IVs, we set \( \gamma_j \)'s according to each simulation setting and set the true variances \( \sigma^2_{Xj} \) and \( \gamma_j^2 \) for non-null IVs to be the corresponding estimated variances in the BMI datasets. For null IVs, we set their \( \gamma_j \)'s to be zero and set their variances \( \sigma^2_{Xj} \) and \( \gamma_j^2 \) based on taking a random sample of the estimated variances in the BMI dataset. Finally, for the outcome dataset, we set the true marginal IV-outcome association parameter \( \Gamma_j \), to be \( \Gamma_j = \beta_0 \gamma_j \) with \( \beta_0 = 0.4 \) and \( \gamma_j \) from the exposure dataset. Like the exposure and selection datasets, the true variances \( \sigma^2_{Yj} \) of the marginal IV-outcome model correspond to the estimated variances from the outcome dataset.

We compare seven MR methods: the IVW estimator introduced in Section 3, the dIVW estimator proposed in Section 4, and five other methods
in the literature, MR-Egger regression [6], weighted median estimator (MR-median) [7], weighted mode estimator (MR-mode) [22], profile score estimator (MR-raps) [40], and profile score with empirical partially Bayes shrinkage weights (MR-raps-shrink) [41]. MR-Egger, MR-median and MR-mode are implemented using the MendelianRandomization R package (version 0.4.1) [39]. To make the comparisons fair, we use the $l_2$ loss while implementing MR-raps via the mr.raps package. For every estimator except MR-raps and MR-rap-shrink, we also use different pre-screening procedures, including $\lambda = 0$ (no thresholding, all SNPs are included), $\lambda = 5.45131$ (p-value cutoff based on the p-value threshold of $5 \times 10^{-8}$), and $\lambda = \sqrt{2 \log p}$ ($\approx 3.75$ when $p = 1119$). We also include the dIVW estimator with $\lambda$ determined by the MR-EO algorithm with the maximum number of iterations set to $t_{\text{max}} = 5$ and using the optimize function from R in the O-step. The standard errors of the IVW and dIVW estimators are estimated from the variance estimators in Theorem 4.2 or 5.1.

Table 1 shows the mean and standard deviation of each estimator. It also shows the average length and coverage probability of 95% confidence intervals (CI). Under all scenarios, the IVW estimator without pre-screening (i.e., $\lambda = 0$) is biased towards zero, which agrees with our theoretical result since the average IV strength $\kappa$’s are relatively small. The IVW estimator’s coverage probabilities are far from 95% due to the downward bias and the inaccurate normal approximation. The pre-screened IVW estimator under the threshold $\lambda = 5.45$ or $\sqrt{2 \log p}$ improves the performance of the IVW estimator substantially, which again agrees with our theory that the pre-screened IVW estimator requires less stringent assumptions for consistency and asymptotic normality.

The proposed dIVW estimator is robust against weak and null IVs and shows great performance across all simulation scenarios, with and without pre-screening. This observation agrees with our theoretical assessment that the dIVW estimator requires far less stringent conditions for consistency and asymptotic normality than the IVW estimator. The pre-screened dIVW estimator does improve upon the dIVW estimator by having a smaller variance and a shorter confidence interval, while having no impact on bias and coverage.

Without pre-screening, MR-Egger, MR-median and MR-mode are biased when the average IV strength is small. In particular, the MR-mode without pre-screening can be severely biased with unrealistically wide confidence interval. MR-Egger, MR-median and MR-mode thresholded at 5.45 or $\sqrt{2 \log p}$ generally have wider confidence intervals compared to the dIVW estimators thresholded at the same level. Also, even with thresholding, these three
Table 1
Simulation results for Case 1- Case 3 based on 10,000 repetitions with \( p = 1119 \) candidate instruments. The true effect is \( \beta_0 = 0.4 \). SD represents standard deviation, length represents the average length of the 95% CI, and CP represents the empirical coverage probability of the 95% CI.

| Case | Method      | \( \lambda \) | mean SD | length CP |
|------|-------------|----------------|---------|-----------|
| 1    | IVW         | 0              | 0.260   | 0.069     | 0.269     | 46.92   |
| \( s = 20 \) | IVW         | 5.45           | 0.398   | 0.094     | 0.366     | 94.82   |
| \( \kappa = 2.90 \) | IVW         | \( \sqrt{2\log \hat{p}} \) | 0.398   | 0.087     | 0.341     | 95.10   |
| dIVW | 0           | 0.402          | 0.107   | 0.419     | 95.20     |
| dIVW | 5.45        | 0.401          | 0.095   | 0.368     | 94.86     |
| dIVW | \( \sqrt{2\log \hat{p}} \) | 0.401  | 0.087   | 0.343     | 95.12     |
| dIVW | MR-EO       | 0.400          | 0.086   | 0.337     | 95.06     |
| MR-Egger | 0         | 0.335          | 0.082   | 0.322     | 87.52     |
| MR-Egger | 5.45   | 0.390          | 0.240   | 1.005     | 96.00     |
| MR-Egger | \( \sqrt{2\log \hat{p}} \) | 0.389  | 0.205   | 0.840     | 95.60     |
| MR-median  | 0         | 0.371          | 0.110   | 0.477     | 96.27     |
| MR-median | 5.45   | 0.398          | 0.118   | 0.501     | 96.51     |
| MR-median | \( \sqrt{2\log \hat{p}} \) | 0.397  | 0.113   | 0.486     | 96.70     |
| MR-mode    | 0         | 0.033          | 0.106   | 296293 100 |
| MR-mode | 5.45      | 0.395          | 0.139   | 0.591     | 97.14     |
| MR-mode | \( \sqrt{2\log \hat{p}} \) | 0.395  | 0.142   | 0.616     | 97.15     |
| MR-raps   | 0         | 0.401          | 0.105   | 0.413     | 95.16     |
| MR-raps-shrink | 0  | 0.400          | 0.086   | 0.336     | 95.12     |
| 2    | IVW         | 0              | 0.159   | 0.091     | 0.354     | 23.94   |
| \( s = 100 \) | IVW         | 5.45           | 0.397   | 0.206     | 0.806     | 95.05   |
| \( \kappa = 1.05 \) | IVW         | \( \sqrt{2\log \hat{p}} \) | 0.394   | 0.183     | 0.717     | 94.88   |
| dIVW | 0           | 0.404          | 0.233   | 0.912     | 95.39     |
| dIVW | 5.45        | 0.400          | 0.207   | 0.811     | 95.10     |
| dIVW | \( \sqrt{2\log \hat{p}} \) | 0.400  | 0.186   | 0.728     | 94.87     |
| dIVW | MR-EO       | 0.396          | 0.167   | 0.655     | 95.00     |
| MR-Egger | 0         | 0.231          | 0.122   | 0.483     | 72.28     |
| MR-Egger | 5.45   | 0.388          | 0.948   | 3.787     | 96.22     |
| MR-Egger | \( \sqrt{2\log \hat{p}} \) | 0.385  | 0.359   | 1.508     | 95.77     |
| MR-median  | 0         | 0.276          | 0.152   | 0.668     | 91.31     |
| MR-median | 5.45   | 0.396          | 0.228   | 0.959     | 96.60     |
| MR-median | \( \sqrt{2\log \hat{p}} \) | 0.394  | 0.216   | 0.925     | 96.93     |
| MR-mode    | 0         | -1.062         | 130     | 306245 100 |
| MR-mode | 5.45      | 0.387          | 0.267   | 1.141     | 97.00     |
| MR-mode | \( \sqrt{2\log \hat{p}} \) | 0.39   | 0.237   | 1.123     | 97.10     |
| MR-raps   | 0         | 0.398          | 0.224   | 0.884     | 94.91     |
| MR-raps-shrink | 0  | 0.399          | 0.160   | 0.625     | 95.24     |
methods (MR-Egger, MR-median and MR-mode) have larger biases than the dIVW estimator because they inherently rely on using the ratio estimator $\hat{\beta}_j$.

The MR-raps without thresholding performs well with respect to bias and coverage and is comparable to the dIVW estimator with respect to bias, coverage length, and coverage probability. The dIVW estimator with a threshold chosen by MR-EO performs as well as the MR-raps-shrink, the most efficient among the MR-raps estimators. MR-raps-shrink can have slightly smaller standard errors than the most efficient dIVW estimator. But, the method is computationally more complicated and may not have unique solution as mentioned in [41]. In contrast, the dIVW estimator is much simpler and has an explicit, unique solution.

Table 2 presents the results on the number of total IVs and non-null IVs selected based on different pre-screening thresholds. We see from Table 2 that the number of non-null IVs selected based on genome-wide significance ($p$-value $\leq 5 \times 10^{-8}$ or $\lambda \approx 5.45$) is very small compared with $s$, the true number of non-null IVs; this is especially true under Case 2 and Case 3 when IVs are weak. Taking the theoretically motivated $\lambda = \sqrt{2\log p}$ improves upon the genomic threshold, but it still eliminates too many IVs when IVs are weak (Cases 2 and 3). The proposed MR-EO algorithm does

|               | IVW     | 3     | 0.352 | 0.047 | 0.185 | 82.60 |
|---------------|---------|-------|-------|-------|-------|-------|
| $s = 1119$    | IVW     | 5.45  | 0.395 | 0.086 | 0.340 | 95.41 |
| $\kappa = 7.78$ | IVW     | $\sqrt{2\log p}$ | 0.392 | 0.068 | 0.270 | 94.99 |
| dIVW          | 0       | 0.400 | 0.054 | 0.210 | 94.70 |
| dIVW          | 5.45    | 0.399 | 0.087 | 0.343 | 95.43 |
| dIVW          | $\sqrt{2\log p}$ | 0.399 | 0.070 | 0.275 | 95.42 |
| dIVW          | MR-EO   | 0.400 | 0.054 | 0.210 | 94.79 |
| MR-Egger      | 0       | 0.372 | 0.066 | 0.262 | 93.11 |
| MR-Egger      | 5.45    | 0.383 | 0.189 | 0.776 | 95.37 |
| MR-Egger      | $\sqrt{2\log p}$ | 0.372 | 0.132 | 0.534 | 95.01 |
| MR-median     | 0       | 0.375 | 0.079 | 0.354 | 96.64 |
| MR-median     | 5.45    | 0.394 | 0.114 | 0.49  | 96.81 |
| MR-median     | $\sqrt{2\log p}$ | 0.391 | 0.100 | 0.436 | 96.86 |
| MR-mode       | 0       | 0.750 | 84    | 91149 | 100  |
| MR-mode       | 5.45    | 0.391 | 0.125 | 0.551 | 96.77 |
| MR-mode       | $\sqrt{2\log p}$ | 0.385 | 0.260 | 1.974 | 97.62 |
| MR-raps       | 0       | 0.400 | 0.054 | 0.210 | 94.90 |
| MR-raps-shrink| 0       | 0.400 | 0.053 | 0.209 | 94.71 |
the best, although it doesn’t always select the correct number of non-null IVs especially under Case 2. Nevertheless, as seen above, MR-EO as applied to the dIVW estimator, increases efficiency and shortens CI length without increasing bias or losing coverage.

Overall, there are three takeaways from the simulation study. First, the dIVW estimator with or without pre-screening always outperforms the IVW estimator. Second, for an MR estimator that is robust to weak IVs (e.g., dIVW), if a selection dataset is available and pre-screening is applied, we suggest the threshold used in the pre-screening process should be based on \( \lambda = \sqrt{2 \log p} \) instead of the usual cutoff \( \lambda \approx 5.45 \). Third, to improve efficiency of an estimator that is robust to weak IVs, (e.g., dIVW), we recommend using the MR-EO algorithm to adaptively choose \( \lambda \).

In the supplementary material, we conduct a simulation study under balanced horizontal pleiotropy. In summary, the results are nearly identical to Case 3 without pleiotropy in that our dIVW estimator shows robustness against weak and balanced pleiotropic IVs.

6.2. *Empirical Guidelines for Asymptotics.* In practice with real data, it is important to have some sense of what is “a large enough” sample size for the asymptotic results in the paper to be a good approximation. Many researchers in MR have conducted such analysis for the IVW estimator, most notably [10]. Here, we conduct a similar analysis for the dIVW estimator through a small simulation study. Specifically, we examine what would be a “large” effective sample size, as measured by \( \kappa_{\lambda, dIVW}^* \sqrt{p_{\lambda}^*} / \max(1, \lambda^2) \), for the asymptotics promised by Theorem 4.2 to be plausible.

The setting is identical to Case 3. We choose a grid of 100 equally spaced \( \lambda \)'s between 0 and 10. For each \( \lambda \), we generate 1,000 simulation datasets with the given simulation parameters and calculate the corresponding \( \hat{\beta}_{\lambda, dIVW} \) for each dataset. Figure 1 plots these \( \hat{\beta}_{\lambda, dIVW} \) values against \( \kappa_{\lambda}^* \sqrt{p_{\lambda}} / \max(1, \lambda^2) \) (dots) together with the two standard error bands centered at \( \hat{\beta}_0 \) (shaded area).

We find that for any \( \lambda \), the coverage probability for the pre-screened
dIVW estimator is around 95% and ranges from 93.5% to 96.0%. However, as $\hat{\kappa}_j^* \sqrt{p_j^*} / \max(1, \lambda^2)$ grows larger, we saw fewer estimates far from $\beta_0$, an indication that asymptotics have “kicked in.” This appears to occur when $\hat{\kappa}_j^* \sqrt{p_j^*} / \max(1, \lambda^2)$ is greater than 20. Based on this, we recommend that users of dIVW check to make sure that $\hat{\kappa}_j^* \sqrt{p_j^*} / \max(1, \lambda^2)$ is at least greater than 20 as part of diagnostic checks for the dIVW estimator.

7. Real Data Example. We apply our methods to the BMI-CAD example described in Section 6. Table 3 summarizes the results, where dIVW$_{\alpha}$ denotes the dIVW estimator developed under balanced horizontal pleiotropy, MR-raps$_{\alpha}$ and MR-raps-shrink$_{\alpha}$ are MR-raps estimators that account for balanced pleiotropy by setting the over.dispersion parameter in the mr.raps R package to be TRUE.

Overall, except for the MR-mode estimator without any pre-screening
Table 3

Point estimates and the estimated standard errors (in parentheses) from different MR methods in the BMI-CAD example. Each column represents different $\lambda$ thresholds for pre-screening. MR-EO selects slightly different $\lambda$ depending on whether the dIVW or dIVW$\alpha$ are used; the left-hand side of the semicolon refers to the dIVW estimator and the right-hand size of the semicolon refers to the dIVW$\alpha$ estimator.

| $\lambda$ | 0  | 5.45 | $\sqrt{2\log p} = 3.75$ | MR-EO |
|-----------|----|------|----------------|-------|
| number of IVs selected | 1119 | 44 | 165 | 1029; 1023 |
| $\hat{\kappa}\sqrt{\hat{p}/\max(1,\lambda^2)}$ | 6.8 | 72.9 | 28.1 | 7.2; 7.3 |
| $\hat{\kappa}\sqrt{\hat{p}/\max(1,\lambda^2)}$ | 226.8 | 16.3 | 25.7 | 232.4; 233.1 |
| IVW | 0.315 (0.050) | 0.282 (0.084) | 0.319 (0.068) | |
| dIVW | 0.365 (0.058) | 0.287 (0.085) | 0.331 (0.071) | 0.345 (0.058) |
| dIVW$\alpha$ | 0.365 (0.067) | 0.287 (0.100) | 0.331 (0.082) | 0.345 (0.067) |
| MR-raps | 0.382 (0.061) | 0.291 (0.086) | 0.339 (0.072) | |
| MR-raps$\alpha$ | 0.367 (0.067) | 0.297 (0.120) | 0.337 (0.090) | |
| MR-raps-shrink | 0.388 (0.060) | 0.292 (0.086) | 0.341 (0.072) | |
| MR-raps-shrink$\alpha$ | 0.374 (0.067) | 0.298 (0.120) | 0.339 (0.090) | |
| MR-Egger | 0.356 (0.077) | 0.513 (0.184) | 0.390 (0.129) | |
| MR-median | 0.322 (0.097) | 0.278 (0.124) | 0.304 (0.116) | |
| MR-mode | 0.739 (402.9) | 0.499 (0.402) | 0.488 (4.241) | |

and the MR-Egger method using the threshold $\lambda = 5.45$, all the estimated exposure effects are within two standard errors of each other. We also suspect the IVW estimator is slightly biased towards zero when $\lambda = 0$ since its point estimator is smaller than that from the dIVW estimator, which is generally unbiased with weak IVs. We notice that selecting IVs based on genome-wide significance (i.e., $\lambda = 5.45$) generally leads to larger estimated standard errors across all methods. Except for the IVW estimator, the dIVW estimator has the smallest standard errors among all MR estimators across the three choices of $\lambda$. Also, the dIVW estimator using $\lambda$ selected by the MR-EO algorithm has the smallest standard error while having a point estimator that is similar to the other methods. Comparing thresholds, we generally recommend using $\lambda = \sqrt{2\log p}$ for all methods that are not sensitive to weak IVs because this threshold adapts to the dimension $p$ and includes more relevant IVs than a static threshold of $\lambda = 5.45$ based on the genome-wide significance level.

We also conduct diagnostics to assess the plausibility of Assumption 1 in the real dataset. Specifically, when Assumption 1 holds, we have $\kappa\sqrt{p} \to \infty$, $\hat{\beta}_{dIVW}$ is close to $\beta_0$, and the standardized residuals,

$$
\frac{\Gamma_j - \hat{\beta}_{dIVW}\hat{\gamma}_j}{\sqrt{\hat{\sigma}_j^2 + \hat{\beta}_{dIVW}^2\hat{\sigma}_j^2}}, \quad j = 1, ..., p,
$$

should follow a standard normal distribution. Following [40], we construct
a Quantile-Quantile (QQ) plot of the standardized residuals to check Assumption 1. Figure 2 shows the result. Since the residuals line up close to the 45-degree line, Assumption 1 is likely to hold for this example. In the supplementary material, we also assess the plausibility of Assumption 3 by using a similar strategy as in Figure 2. The QQ plot for Assumption 3 looks similar to Figure 2. In total, Assumptions 1 and 3 are plausible with the data.

8. Discussion. In this paper, we analyze the behavior of the inverse-variance weighted (IVW) estimator and the pre-screened IVW estimator in two-sample summary-data Mendelian randomization studies. We show that the IVW estimator requires stringent assumptions on the average strength of instruments for consistency or asymptotic normality. The pre-screened IVW estimator requires careful choice of the p-value threshold in the selection dataset for the desired statistical properties. We then propose a simple modification of the IVW estimator, called the debiased IVW (dIVW) estimator. The dIVW estimator is drastically more robust to many weak instruments than the IVW estimator and performs well with or without pre-screening. We also show that the common pre-screening practice of selecting strong instruments based on a p-value cutoff at $5 \times 10^{-8}$ may eliminate majority of relevant IVs. In order to increase estimation efficiency of the dIVW esti-
mator, we then propose two approaches of choosing SNPs via $\lambda$, one based on $\lambda = \sqrt{2\log p}$ that effectively eliminates null IVs and the other that uses the MR-EO algorithm to directly maximize efficiency.

Based on our theoretical and simulation works, we make two recommendations for two-sample summary-data Mendelian randomization studies. First, we argue that the dIVW estimator without pre-screening should be the default, baseline estimator for two-sample summary-data Mendelian randomization studies instead of the IVW estimator. It is as simple as the IVW estimator, requiring one to multiply the IVW estimator with a bias-correction factor, and has provable robustness against many weak instruments and balanced horizontal pleiotropy. If an independent selection dataset is not available, including every SNP in dIVW can still lead to a consistent and asymptotically normal dIVW estimator. Also, the dIVW estimator’s performance should serve as the baseline for investigating more complex pleiotropy as the dIVW estimator effectively eliminates concerns for weak instruments in most MR settings. Second, to improve efficiency of the dIVW estimator, and if an independent selection dataset is available, we generally recommend a thresholding value of $\lambda = \sqrt{2\log p}$ on the selection dataset instead of the commonly used genome-wide significance p-value threshold so that the threshold adapts to the initial pool of IVs. We also recommend using the proposed MR-EO algorithm to adaptively select $\lambda$ and maximize estimation efficiency. Future work will explore how to modify the dIVW estimator to be robust against other types of pleiotropy.

**SUPPLEMENTARY MATERIAL**

**Supplementary Material: Debiased Inverse-Variance Weighted Estimator in Two-Sample Summary-Data Mendelian Randomization**

(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide simulation results under balanced horizontal pleiotropy and theoretical proofs for the theorems in the paper.

**References.**

[1] Abbott, L., Bryant, S., Churchhouse, C. and et al. (2018). Round 2 GWAS Results of Thousands of Phenotypes in the UK Biobank. http://www.nealelab.is/uk-biobank (14 November 2018, date last accessed).

[2] Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikekawa, S., Hirata, M., Matsuda, K., Iwasaki, M., Yamaji, T., Sawada, N., Hachiya, T., Tanno, K., Shimizu, A., Hozawa, A., Minegishi, N., Tsugane, S., Yamamoto, M., Kubo, M. and Kamatani, Y. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nature Genetics 49 1458-1467.
Andrews, D. W. K. and Stock, J. H. (2005). Inference with Weak Instruments. Working Paper No. 313, National Bureau of Economic Research.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 15 69-85.

Baiocchi, M., Cheng, J. and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine* 33 2297–2340.

Bowden, J., Davey Smith, G. and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* 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. *Genetic Epidemiology* 40 304–314.

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

Bowden, J., Del Greco M. F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J. and Davey Smith, G. (2019). Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International journal of epidemiology* 48 728–742.

Burgess, S., Butterworth, A. and Thompson, S. G. (2013). Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology* 37 658–665.

Burgess, S., Small, D. S. and Thompson, S. G. (2015). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* 26 2333–2355.

Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine* 30 1312–1323.

Burgess, S., Thompson, S. G. and Collaboration, C. C. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology* 40 755–764.

Burgess, S. and Thompson, S. G. (2012). Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine* 31 1582–1600.

Chao, J. C. and Swanson, N. R. (2005). Consistent Estimation with a Large Number of Weak Instruments. *Econometrica* 73 1673–1692.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* 11 415–425.

Consortium, I. H. (2005). A haplotype map of the human genome. *Nature* 437 1299–1320.

Corbin, L. J., Richmond, R. C., Wade, K. H., Burgess, S., Bowden, J., Smith, G. D. and Timpson, N. J. (2016). BMI as a Modifiable Risk Factor for Type 2 Diabetes: Refining and Understanding Causal Estimates Using Mendelian Randomization. *Diabetes* 65 3002.

Davey Smith, G. and Ebrahim, S. (2003). ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 32 1–22.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 16
[21] Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13** 135 - 145.

[22] Hartwig, F. P., Davey Smith, G. and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology* **46** 1985–1998.

[23] Hemani, G., Bowden, J. and Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human molecular genetics* **27** R195–R208.

[24] Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Davey Smith, G., Gaunt, T. R., Haycock, P. C. and Loos, R. (2018). The MR-Base platform supports systematic causal inference across the human phenotype. *eLife* **7** e34408.

[25] Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

[26] Kamstrup, P. R., Tyrlérg-Hansen, A., Steffensen, R. and Nordestgaard, B. G. (2009). Genetically Elevated Lipoprotein(a) and Increased Risk of Myocardial Infarction. *JAMA* **301** 2331–2339.

[27] Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27** 1133–1163.

[28] Pe’er, I., Yelensky, R., Altshuler, D. and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology* **32** 381–385.

[29] Pierce, B. L. and Burgess, S. (2013). Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *American Journal of Epidemiology* **178** 1177–1184.

[30] Pingault, J.-B., O’Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F. and Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics* **19** 566–580.

[31] Sawa, T. (1969). The Exact Sampling Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators. *Journal of the American Statistical Association* **64** 923–937.

[32] Smith, G. D. and Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33** 30–42.

[33] Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14** 483 - 495.

[34] Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica* **65** 557–586.

[35] Stock, J. H., Wright, J. H. and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics* **20** 518–529.

[36] the CARDiOGRAMplusC4D Consortium, Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C. and et al. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47** 1121 -1130.

[37] Verbanck, M., Chen, C.-Y., Neale, B. and Do, R. (2018). Detection of
widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* 50 693–698.

[38] Wang, S. and Kang, H. (2019). Weak-Instrument Robust Tests in Two-Sample Summary-Data Mendelian Randomization. *arXiv 1909.06950*.

[39] Yavorska, O. O. and Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International Journal of Epidemiology* 46 1734–1739.

[40] Zhao, Q., Wang, J., Hemani, G., Bowden, J. and Small, D. S. (2019a). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *To appear in Annals of Statistics*.

[41] Zhao, Q., Chen, Y., Wang, J. and Small, D. S. (2019b). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *International Journal of Epidemiology*.

[42] Zheng, J., Baird, D., Borges, M.-C., Bowden, J., Hemani, G., Haycock, P., Evans, D. M. and Smith, G. D. (2017). Recent Developments in Mendelian Randomization Studies. *Current Epidemiology Reports* 4 330–345.