genes

MDPI

Article

Support Interval for Two-Sample Summary Data-Based Mendelian Randomization

Kai Wang

Department of Biostatistics, University of Iowa, 145 N Riverside Dr., Iowa City, IA 52242, USA; kai-wang@uiowa.edu

Abstract: The summary-data-based Mendelian randomization (SMR) method is gaining popularity in estimating the causal effect of an exposure on an outcome. In practice, the instrument SNP is often selected from the genome-wide association study (GWAS) on the exposure but no correction is made for such selection in downstream analysis, leading to a biased estimate of the effect size and invalid inference. We address this issue by using the likelihood derived from the sampling distribution of the estimated SNP effects in the exposure GWAS and the outcome GWAS. This likelihood takes into account how the instrument SNPs are selected. Since the effective sample size is 1, the asymptotic theory does not apply. We use a support for a profile likelihood as an interval estimate of the causal effect. Simulation studies indicate that this support has robust coverage while the confidence interval implied by the SMR method has lower-than-nominal coverage. Furthermore, the variance of the two-stage least squares estimate of the causal effect is shown to be the same as the variance used for SMR for one-sample data when there is no selection.

Keywords: mendelian randomization; summary statistics; SMR; profile likelihood; support

check for updates

Citation: Wang, K. Support Interval for Two-Sample Summary Data-Based Mendelian Randomization. *Genes* **2023**, *14*, 211. https://doi.org/10.3390/genes14010211

Academic Editor: Aristotelis Chatziioannou

Received: 27 November 2022
Revised: 31 December 2022
Accepted: 4 January 2023
Published: 13 January 2023

1. Introduction

A main interest in scientific research is to study the causal effect of an exposure x on an outcome y. When the outcome is continuous, the causal effect is the coefficient b in the following regression model:

$$y = bx + u + \epsilon_y, \tag{1}$$

where u represents the unobserved factors and ϵ_y is normally distributed with mean 0 and variance σ_y^2. Throughout this report, all variables are centered so that the intercept is equal to 0.

When u confounds the effect of x, the least squares estimate of b is biased. Mendelian randomization (MR) is a modern technique for correcting this bias [1–5], thanks to the availability of the large number of genome-wide association studies (GWASs). One appealing feature of the summary-data-based MR methods is that they don't rely on individual-level data.

MR is an application of instrumental variable (IV) analysis to estimate b. IV analysis is able to control for unobserved confounders. MR uses single nucleotide polymorphisms (SNPs) as IVs. Let g denote the genotypic score of an SNP. For an IV to be valid, it must satisfy the following assumptions [6]:

Relevance: It is associated with the exposure x (i.e., $Cov(g, x) \neq 0$);

Exclusion Restriction: It affects the outcome y only through its association with the exposure; and

Exchangeability: It is not associated with any confounders of the exposure–outcome association, which implies $Cov(g, y) = bCov(g, x)$.

Under these assumptions, Equation (1) implies:

$$b_{gy} = bb_{gx},$$

where $b_{gy} = Cov(g,y)/Var(g)$ and $b_{gx} = Cov(g,x)/Var(g)$. Since b_{gx} and b_{gy} can be estimated from the exposure GWAS and the outcome GWAS, respectively, a popular summary-data MR (SMR) estimate of b is [2]:

$$\hat{b}_{\mathrm{SMR}} = \frac{\hat{b}_{gy}}{\hat{b}_{gx}},$$

where $\hat{b}_{gy}$ and $\hat{b}_{gx}$ are GWAS estimates of b_{gy} and b_{gx}, respectively.

In practice, in order to satisfy the relevance assumption on an IV, an SNP is typically selected from the exposure GWAS, often at the genome-wide significance level $p < 5 \times 10^{-8}$. Hence the selected SNPs are subject to a winner's curse that leads to a biased effect estimate. This is an issue that has been recognized in the MR literature for some time [7–13]. A typical solution is to use another GWAS on the exposure to screen for IV SNPs [10,14,15]. However, such a GWAS may not always be available. A simple correction is to transform the false discovery rate to the z-scale [11]. An empirical study on the effect of the winner's curse on a Mendelian randomization study is presented in [12]. A review of methods to overcome the winner's curse in the context of genetic association studies is provided in [13].

As a matter of fact, many applications [2,16,17], including those contained in the original paper that proposed SMR [2], do not use another GWAS for screening. The IV SNPs are simply selected from the exposure GWAS and the selection bias is not corrected for in the downstream analysis [2,18,19]. Furthermore, this approach has been generalized to other settings [20,21].

This research is based on the sampling distribution of the estimated SNP effects on the exposure and on the outcome. Despite the large number of subjects used in the exposure GWAS and the outcome GWAS, there are only 4 summary statistics (i.e., two coefficient estimates and their respective standard errors) needed for MR at an IV SNP. The standard asymptotic theory, which requires a large sample size, does not apply since there is only one "observation" at an SNP. To sidestep this issue, we use a support derived from a profile likelihood as an interval estimate for b and assess its coverage probability through simulation. Support is the set of parameter values at which the log profile likelihood is a certain unit below the maximum log profile likelihood. It can be considered an extension of the confidence interval. Simulation studies demonstrate that the 2-unit support has robust coverage while the confidence interval implied by the SMR method has lower-than-nominal coverage.

In addition, we point out that the standard error of $\hat{b}_{\mathrm{SMR}}$ derived from the delta method is the same as the standard error derived from the standard theory on two-stage least-squares (TSLS) regression for one-sample individual-level data in the absence of SNP selection.

2. Materials and Methods

2.1. One-Sample Individual-Level Data

In this subsection, we consider one-sample individual-level data where the IV SNP is not selected for its significant p-value. In this case, the delta method estimate of $SE(\hat{b}_{\mathrm{SMR}})$ is the same as the estimate derived from the theory on the TSLS method. The delta method is the method used by SMR [2]. This result indicates another connection between SMR and TSLS, in addition to the connection that $\hat{b}_{\mathrm{SMR}}$ is the same as the TSLS estimate of b. James E. Pustejovsky proved this relationship in a blog post [22]. Below we provide a similar proof in our current context.

Consider the following GWAS model on exposure x:

$$x = b_{gx}g + \epsilon_x,$$

where ϵ_x correlates with the unobserved confounder u. The reduced-form equation for y is:

$$y = b_{gy} g + (b\epsilon_x + u + \epsilon_y),$$

where $b_{gy} = bb_{gx}$. Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{g}$ denote the centered vectors of n observations on x, y, and g, respectively. Estimates of b_{gx} and b_{gy} can be obtained as follows: $\hat{b}_{gx} = \mathbf{g}'\mathbf{x}/\mathbf{g}'\mathbf{g}$ and $\hat{b}_{gy} = \mathbf{g}'\mathbf{y}/\mathbf{g}'\mathbf{g}$. Let $\mathbf{P} = \mathbf{g}\mathbf{g}'/\mathbf{g}'\mathbf{g}$. Their second-order moments are estimated by:

$$\begin{aligned}
\widehat{Var}(\hat{b}_{gx}) &= \frac{n^{-1}\mathbf{x}'(\mathbf{I}-\mathbf{P})\mathbf{x}}{\mathbf{g}'\mathbf{g}}, \\
\widehat{Var}(\hat{b}_{gy}) &= \frac{n^{-1}\mathbf{y}'(\mathbf{I}-\mathbf{P})\mathbf{y}}{\mathbf{g}'\mathbf{g}}, \\
\widehat{Cov}(\hat{b}_{gx}, \hat{b}_{gy}) &= \frac{n^{-1}\mathbf{x}'(\mathbf{I}-\mathbf{P})\mathbf{y}}{\mathbf{g}'\mathbf{g}}.
\end{aligned}$$

We note that $\widehat{Cov}(\hat{b}_{gx}, \hat{b}_{gy}) \neq 0$.

The TSLS estimate of b is the least squares estimate of the coefficient in the regression where the response is $\mathbf{y}$ and the predictor is $\mathbf{Px}$:

$$\hat{b}_{\text{TSLS}} = \frac{\mathbf{x}'\mathbf{P}\mathbf{y}}{\mathbf{x}'\mathbf{P}\mathbf{x}} = \frac{\mathbf{x}'\mathbf{g}(\mathbf{g}'\mathbf{g})^{-1}\mathbf{g}'\mathbf{y}}{\mathbf{x}'\mathbf{g}(\mathbf{g}'\mathbf{g})^{-1}\mathbf{g}'\mathbf{x}} = \frac{(\mathbf{g}'\mathbf{g})^{-1}\mathbf{g}'\mathbf{y}}{(\mathbf{g}'\mathbf{g})^{-1}\mathbf{g}'\mathbf{x}} = \frac{\hat{b}_{gy}}{\hat{b}_{gx}} = \hat{b}_{\text{SMR}}.$$

The delta method estimate of the variance of $\hat{b}_{\text{SMR}}$ is [2]:

$$V_{\text{delta}} = \frac{1}{\hat{b}_{gx}^2}\left[Var(\hat{b}_{gy}) + \hat{b}_{\text{TSLS}}^2 Var(\hat{b}_{gx}) - 2\hat{b}_{\text{TSLS}} Cov(\hat{b}_{gx}, \hat{b}_{gy})\right].$$

Since $\hat{b}_{gx}^2\mathbf{g}'\mathbf{g} = \mathbf{x}'\mathbf{P}\mathbf{x}$, we have:

$$\begin{aligned}
V_{\text{delta}} &= \frac{1}{\hat{b}_{gx}^2} \cdot \frac{1}{n\mathbf{g}'\mathbf{g}}\left[\mathbf{y}(\mathbf{I}-\mathbf{P})\mathbf{y} + \hat{b}_{\text{TSLS}}^2\mathbf{x}'(\mathbf{I}-\mathbf{P})\mathbf{x} - 2\hat{b}_{\text{TSLS}}\mathbf{x}'(\mathbf{I}-\mathbf{P})\mathbf{y}\right] \\
&= \frac{n^{-1}(\mathbf{y}-\hat{b}_{\text{TSLS}}\mathbf{x})'(\mathbf{y}-\hat{b}_{\text{TSLS}}\mathbf{x})}{\mathbf{x}'\mathbf{P}\mathbf{x}} - \frac{n^{-1}(\mathbf{y}-\hat{b}_{\text{TSLS}}\mathbf{x})'\mathbf{P}(\mathbf{y}-\hat{b}_{\text{TSLS}}\mathbf{x})}{\mathbf{x}'\mathbf{P}\mathbf{x}} \\
&= \frac{n^{-1}(\mathbf{y}-\hat{b}_{\text{TSLS}}\mathbf{x})'(\mathbf{y}-\hat{b}_{\text{TSLS}}\mathbf{x})}{\mathbf{x}'\mathbf{P}\mathbf{x}}.
\end{aligned} \tag{2}$$

The last equal sign holds because:

$$\mathbf{P}(\mathbf{y}-\hat{b}_{\text{TSLS}}\mathbf{x}) = \mathbf{g}(\mathbf{g}'\mathbf{g})^{-1}\mathbf{g}'\mathbf{y} - \frac{\mathbf{g}'\mathbf{y}}{\mathbf{g}'\mathbf{x}} \cdot \mathbf{g}(\mathbf{g}'\mathbf{g})^{-1}\mathbf{g}'\mathbf{x} = \mathbf{0},$$

where $\mathbf{0}$ is a vector of 0's. The right-hand side of Equation (2) is exactly the estimated variance of $\hat{b}_{\text{TSLS}}$ defined in the standard theory on TSLS method [23]. Combining all these results, we have $V_{\text{delta}} = V_{\text{TSLS}}$.

V_{delta} can not be computed from GWAS summary data as there is no information on $Cov(\hat{b}_{gx}, \hat{b}_{gy})$. For the same reason, TSLS can also not be computed from GWAS summary data.

However, when $\hat{b}_{gx}$ and $\hat{b}_{gy}$ are derived from two independent samples, $Cov(\hat{b}_{gx}, \hat{b}_{gy}) = 0$ and V_{delta} can be computed from GWAS summary data, as is shown in the SMR method [2]. SMR tests whether the exposure has a causal effect on the outcome using a statistic T_{SMR} defined by:

$$T_{\text{SMR}} = \frac{\hat{b}_{\text{SMR}}^2}{V_{\text{delta}}}, \tag{3}$$

where $\hat{b}_{\mathrm{SMR}} = \hat{b}^2_{gy}/\hat{b}^2_{gx}$ and

$$V_{\mathrm{delta}} = \frac{1}{\hat{b}^2_{gx}}\Big[Var(\hat{b}_{gy}) + \hat{b}^2_{\mathrm{SMR}}Var(\hat{b}_{gx})\Big].$$

On the other hand, the TSLS method is not defined for two-samples MR although there are some extensions [24].

2.2. Two Independent Samples with a Selected SNP

In this subsection, we consider two-sample MR where the IV SNP is selected from the exposure GWAS. The purpose of this selection is to ensure that the IV SNP is associated with the exposure. This practice is commonly used in empirical MR studies [2,16]. The selection criterion for an SNP is typically $|\hat{b}_{gx}|/SE(\hat{b}_{gx}) \geq \tau$ for a prespecified τ. For the genome-wide significance level 5×10^{-8}, $\tau = 5.45131$.

The summary statistics used in an MR analysis are $\hat{b}_{gx}, SE(\hat{b}_{gx}), \hat{b}_{gy}$, and $SE(\hat{b}_{gy})$. To simplify notations, they will be denoted by x, σ_x, y, and σ_y, respectively, and b_{gx} and b_{gy} will be denoted by μ_x and μ_y, respectively. We ignore the sampling variation in σ_x and σ_y as they typically are derived from GWASs of very large sample sizes. A similar assumption is made elsewhere, for instance, [25]. In these notations, x and y have the following sampling distributions, respectively:

$$x \sim CN(\mu_x, \sigma^2_x), \ \ y \sim N(\mu_y, \sigma^2_y),$$

where $CN(\cdot,\cdot)$ stands for a conditional normal given $|x/\sigma_x| \geq \tau$ and $N(\cdot,\cdot)$ a normal distribution. The distribution function $CN(\cdot,\cdot)$ was used to construct an approximate conditional likelihood for estimating μ_x [26]. The term "approximate" comes from the fact that the distributions of x (prior to selection) and y are approximately normal.

Let $\alpha_1 = -\tau - \mu_x/\sigma_x$, $\alpha_2 = \tau - \mu_x/\sigma_x$, and

$$\begin{aligned} A &= \Pr(x/\sigma_x \geq \tau) + \Pr(x/\sigma_x \leq -\tau) \\ &= 1 - \Phi(\alpha_2) + \Phi(\alpha_1), \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal. The density function of x is:

$$\frac{1}{A\sigma_x}\phi\left(\frac{x-\mu_x}{\sigma_x}\right),$$

where $\phi(\cdot)$ is the density function of standard normal. The expected value of x is:

$$\mu_x + \frac{\sigma_x}{A}[\phi(\alpha_2) - \phi(\alpha_1)] \tag{4}$$

and its variance is:

$$\tilde{\sigma}^2_x = \sigma^2_x\left\{1 + \frac{\alpha_2\phi(\alpha_2) - \alpha_1\phi(\alpha_1)}{A} - \frac{[\phi(\alpha_2) - \phi(\alpha_1)]^2}{A^2}\right\}.$$

Note that $\tilde{\sigma}^2_x$ is no longer constant; its value depends on μ_x. When there is no selection, $\tau = 0$ and $\alpha_1 = \alpha_2$. The mean and variance reduce to μ_x and σ^2_x, respectively.

Since x is independent of y, the likelihood function $L(\mu_x, \mu_y)$ is:

$$L(\mu_x, \mu_y) = L_x(\mu_x)L_y(\mu_y),$$

where $L_x(\mu_x)$ and $L_y(\mu_y)$ are the likelihood functions based on x and y, respectively.

The likelihood function $L_y(\mu_y)$ is:

$$L_y(\mu_y) = \frac{1}{\sigma_y}\phi\left(\frac{y-\mu_y}{\sigma_y}\right).$$

The MLE of μ_y is apparently $\hat{\mu}_y = y$.

The likelihood function $L_x(\mu_x)$ is:

$$L_x(\mu_x) = \frac{1}{A\sigma_x}\phi\left(\frac{x-\mu_x}{\sigma_x}\right).$$

Its score equation is:

$$x = \mu_x + \frac{\sigma_x}{A}[\phi(\alpha_2) - \phi(\alpha_1)]. \quad (5)$$

This equation determines the MLE $\hat{\mu}_x$ for μ_x. However, there is no explicit form for $\hat{\mu}_x$. The MLE $\hat{\mu}_x$ can be obtained by maximizing $L_x(\mu_x)$ numerically.

From Equations (4) and (5), x is an unbiased estimate of the mean of the conditional normal distribution for x. However it is biased for μ_x as the mean shown in Equation (4) is a nonlinear function of μ_x. Similar comments are made elsewhere [26].

Since $\sigma_x > 0$ and $A > 0$, Equation (5) indicates that when $x > 0$ μ_x must be positive. Otherwise $x - \mu_x$ would be positive and $\phi(\alpha_2) - \phi(\alpha_1)$ is negative (because α_1 is closer to 0 than α_2 is). There would be a contradiction. Because $\mu_x > 0$ implies that the second term of Equation (5) is positive, there is $x > \hat{\mu}_x > 0$. Following the same logic, when $x < 0$, there is $x < \hat{\mu}_x < 0$. In either case, the naïve Wald ratio y/x underestimates $b = \mu_y/\mu_x$. The MLE of b, denoted by $\hat{b}$, is $\hat{b} = y/\hat{\mu}_x$. $\hat{b}$ is biased. The expectation of $\hat{b}$ is $E(y)E(1/\hat{\mu}_x) \neq b$ because $E(1/\hat{\mu}_x) \neq 1/\mu_x$.

Figure 1 shows the $\hat{\mu}_x/\sigma_x$ as a function of x/σ_x. The larger the value of x/σ_x, the smaller the absolute difference $|x/\sigma_x - \hat{\mu}_x/\sigma_x|$. When $|x/\sigma_x| \geq 7.5$ (corresponding to a p-value less than or equal to 6.38×10^{-14}), the absolute difference $|x/\sigma_x - \hat{\mu}_x/\sigma_x|$ is < 0.05 and seems to be negligible.

Under the null:

$$H_0 : b = \mu_y/\mu_x = 0, \mu_x \neq 0,$$

there is $L(\mu_x, \mu_y) = L_x(\mu_x)L_y(0)$. The MLE of μ_x is equal to $\hat{\mu}_x$ determined by Equation (5). Under the alternative:

$$H_1 : b = \mu_y/\mu_x \neq 0, \mu_x \neq 0,$$

the MLE of μ_y is y and the MLE of μ_x is still $\hat{\mu}_x$. The likelihood ratio statistic for testing H_0 against H_1 is:

$$T = 2\log\frac{L(\hat{\mu}_x, y)}{L(\hat{\mu}_x, 0)} = 2\log\frac{L_y(y)}{L_y(0)} = \frac{y^2}{\sigma_y^2} \sim \chi_1^2.$$

This is the "conditional test" we proposed previously [9]. It is more powerful than the SMR statistic shown in Equation (3).

The SMR statistic T_{SMR} shown in Equation (3) does not taking into account the effect of the selection of the IV SNP on the inference. The variance of x is no longer σ_x^2 because x is selected. Even if σ_x^2 is replaced by the variance $\tilde{\sigma}_x^2$ of $CN(\mu_x, \sigma_x^2)$, the resulting statistic is less powerful than the T statistic: Replacing $\hat{b}_{\text{SMR}}$ by $y/\hat{\mu}_x$ and σ_x by $\tilde{\sigma}_x$ in Equation (3), a modification of the SMR statistic would be:

$$\tilde{T}_{\text{SMR}} = \frac{y^2/\hat{\mu}_x^2}{(\sigma_y^2 + y^2/\hat{\mu}_x^2 \cdot \tilde{\sigma}_x^2)/\hat{\mu}_x^2} = \frac{y^2}{\sigma_y^2 + \tilde{\sigma}_x^2 \cdot y^2/\hat{\mu}_x^2} < \frac{y^2}{\sigma_y^2} = T.$$

That is, $\tilde{T}_{\text{SMR}}$ is less powerful than T; p-values for both statistics are calculated from the same distribution, which is chi-square with 1 df. The larger test statistic corresponds to the smaller p-value.

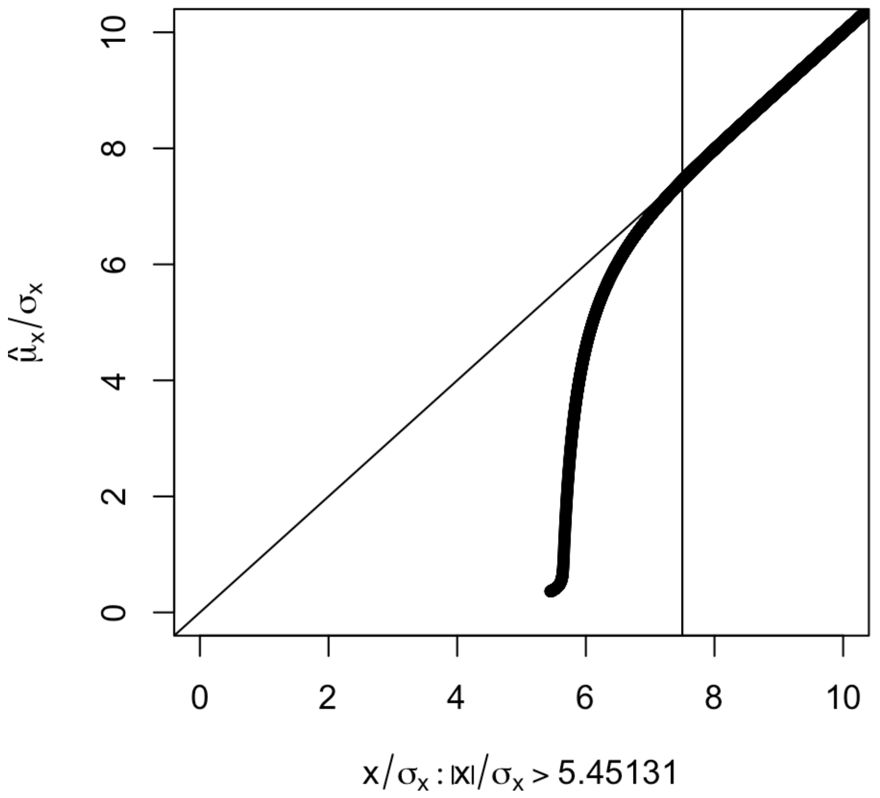

Figure 1. Plot of $\hat{\mu}_x/\sigma_x$ against x/σ_x selected under $|x/\sigma_x| > 5.45131$ (corresponding to $p < 5 \times 10^{-8}$). The vertical line is at $|x/\sigma_x| = 7.5$, which corresponds to a p-value of 6.38×10^{-14}. The part corresponding to $x/\sigma_x < 0$ is not shown since $\hat{\mu}_x/\sigma_x$ is an odd function of x/σ_x.

2.3. Support of Profile Likelihood

We now turn to an interval estimate for $b = \mu_y/\mu_x$. Such an estimate is not trivial since the asymptotic theory is irrelevant as there is effectively only one observation in $L(\mu_x, u_y)$. For this reason, the distribution of $\hat{b} = y/\hat{\mu}_x$, which is the MLE of $b = \mu_y/\mu_x$, is far from normal. To demonstrate this point, the following simulation study is conducted.

We generate 100,000 x's from a normal distribution with $\mu_x = 4$ and $\sigma_x^2 = 1$ (so that there are a reasonable amount of x's to be selected), 7.412 of them satisfy $|x| > 5.45131$ and are selected. The same number (i.e., 7.412) of y's are generated independently from a normal distribution with mean $\mu_y = b\mu_x$ and variance $\sigma_y^2 = 1$. For each (x, y) pair, $\hat{b} = y/\hat{\mu}_x$ is calculated. Histograms of $\hat{b}$ for $b = 0$ and $b = 2$ are shown in Figure 2. For $b = 0$, the mean of $\hat{b}$ is 0.0073 and the median is 0.0022. The distribution has a high probability in the neighborhood of 0. For $b = 2$, the distribution of $\hat{b}$ seems to be bimodal and is skewed to the right with a mean equal to 7.4755 and a median equal to 2.5700. Both values are larger than the true value $b = 2$. These means and medians are also shown in Table 1 together with results from another simulation study described later.

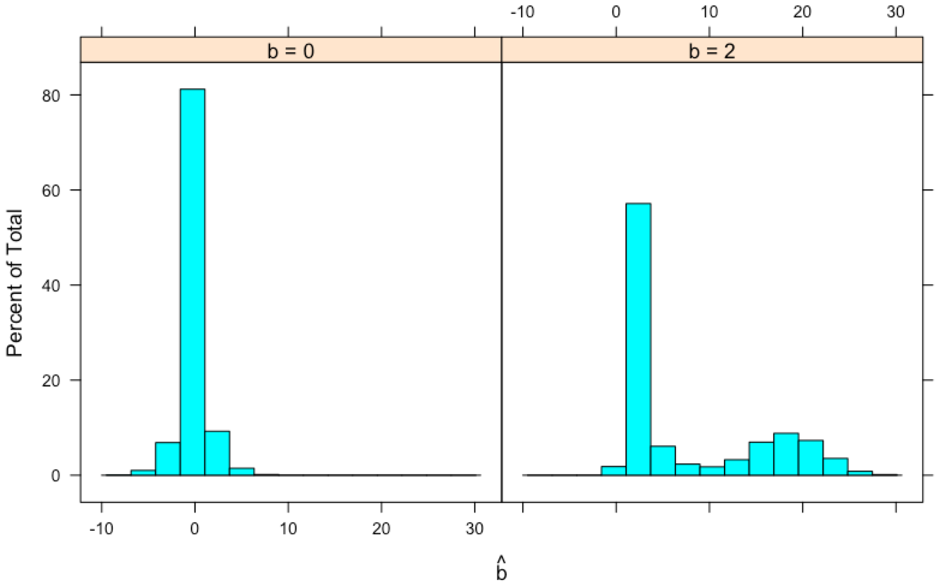

Figure 2. Histogram of simulated $\hat{b} = y/\hat{\mu}_x$, the MLE of b. Data simulation procedure is described in the text.

We consider the profile log-likelihood function $pl(b)$ defined by:

$$pl(b) = \max_{\mu_x}[\log L_x(\mu_x) + \log L_y(b\mu_x)].$$

This function is maximized at $\hat{b} = y/\hat{\mu}_x$ and the maximum is equal to $pl(\hat{b}) = \log L(\hat{\mu}_x, y) = \log L_x(\hat{\mu}_x) + \log L_y(y)$, which is also the maximum of $\log L(\mu_x, \mu_y)$.

A natural interval estimate would be a $1 - \alpha$ profile confidence interval defined as the set of b_0 such that $H_0 : b = b_0$ is not rejected at significance level α. However, the distribution of the log profile likelihood ratio

$$2\left[pl(\hat{b}) - pl(b_0)\right] = 2[\log L(\hat{\mu}_x, y) - pl(b_0)]$$

is unknown for an arbitrary b_0. The only exception is $b_0 = 0$ at which

$$\begin{aligned} 2[\log L(\hat{\mu}_x, y) - pl(0)] &= 2\left\{\log\left[L_x(\hat{\mu}_x)L_y(y)\right] - \log\left[L_x(\hat{\mu}_x)L_y(0)\right]\right\} \\ &= 2\left[\log L_y(y) - \log L_y(0)\right] \\ &= \frac{y^2}{\sigma_y^2} \sim \chi_1^2. \end{aligned}$$

Intuitively, the log partial likelihood function $pl(b)$ can not be approximated by a quadratic function in the vicinity of $\hat{b}$ when $b_0 \neq 0$. As a result, the profile confidence interval for b can not be constructed. An example log partial likelihood function $pl(b)$ is shown in Figure 3 for $x/\sigma_x = 5.4599$ and $y/\sigma_y = 12.3155$.

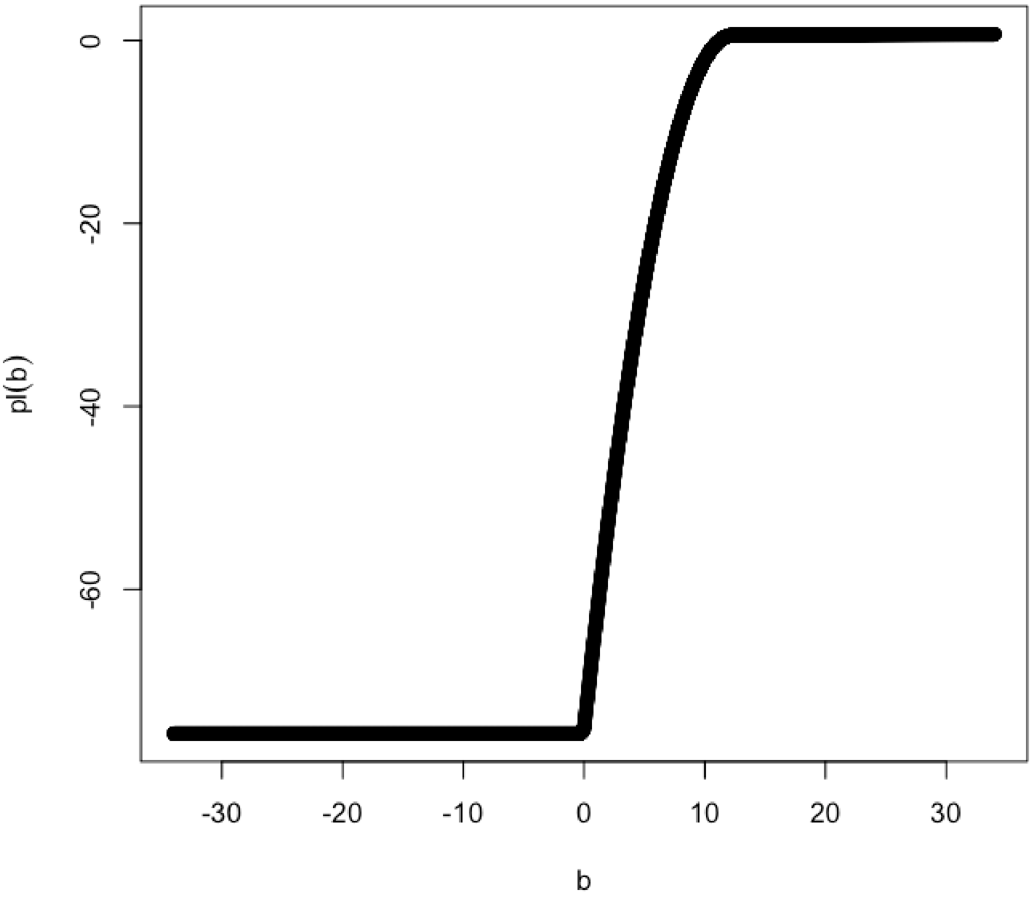

Figure 3. Profile likelihood for $x/\sigma_x = 5.4599$ and $y/\sigma_y = 12.3155$. The MLE of b is $\hat{b} = 33.416$. The lower limit of the 2-unit support is 2.146 and the upper limit is greater than 43.406. The exact value of the upper limit is unknown due to numerical issues. It may be unbounded.

For an interval estimate of b, we use the k-unit support defined by [27]:

$$\left\{b_0 : pl(\hat{b}) - pl(b_0) < k\right\} = \{b_0 : pl(b_0) > \log L(\hat{\mu}_x, y) - k\},$$

where k is a prespecified number. This interval consists of b_0 for which $pl(b_0)$ is greater than $\log L(\hat{\mu}_x, y) - k$. It can be regarded as a generalization of the usual confidence interval. For instance, when $b_0 = 0$ and $k = 2$,

$$\begin{aligned} 0.95 &= \Pr(2[\log L(\hat{\mu}_x, y) - pl(0)] < 3.84) \\ &= \Pr(\log L(\hat{\mu}_x, y) - pl(0) < 1.92) \\ &\approx \Pr(\log L(\hat{\mu}_x, y) - pl(0) < 2). \end{aligned}$$

This approximation worsens as $|b_0|$ moves further away from 0. For the example shown in Figure 3 (i.e., $x/\sigma_x = 5.4599$ and $y/\sigma_y = 12.3155$), the lower limit of the 2-unit support is 2.146 and the upper limit is greater than 43.406. The exact value of the upper limit is unknown due to numerical issues. It may be unbounded.

By the way the support is constructed, the null $H_0 : b_0 = 0$ is rejected by the statistic T at significance level α if and only if the k-unit support, where $k = [\Phi^{-1}(1 - \alpha/2)]^2/2$, contains 0.

We use a simulation study to investigate the coverage of a 2-unit support. For this purpose, data are generated as before but more values for b, i.e., $b = 0, 0.5, 1, 1.5$, and 2, are considered. For each value of b, we compare the winner's-0curse-corrected method and the SMR method in terms of a point estimate of b, an interval estimate of b, and a test of $H_0 : b = 0$. Results are reported in Table 1. Both the winner's-curse-corrected method and SMR method are biased in terms of the mean and median. The SMR 95% confidence interval, computed as $\hat{b}_{\mathrm{SMR}} \pm 1.96 \times V_{\mathrm{delta}}^{1/2}$, has worse coverage as b increases while the 2-unit support has rather stable coverage. In addition, the test statistic T is more powerful than the SMR method.

Table 1. Results of simulation studies with $\mu_x = 4$ and $\sigma_x = \sigma_y = 1$. The statistic T is $T = y^2/\sigma_y^2$.

	b				
Method	**0**	**0.5**	**1**	**1.5**	**2**
Winner's-curse-corrected					
Mean of $\hat{b}$	0.0073	1.8743	3.7414	5.6084	7.4755
Median of $\hat{b}$	0.0022	0.6843	1.3091	1.9327	2.5700
Coverage of 2-unit support	0.9587	0.9725	0.9803	0.9816	0.9811
Power of T for testing $H_0 : b = 0$	0.0471	0.5217	0.9807	1.0000	1.0000
SMR					
Mean of $\hat{b}_{\mathrm{SMR}}$	0.0019	0.3424	0.6829	1.0234	1.3639
Median of $\hat{b}_{\mathrm{SMR}}$	−0.3310	0.3405	0.6795	1.0199	1.3615
Coverage of 95% CI	0.9648	0.8524	0.6511	0.4966	0.3958
Power for testing $H_0 : b = 0$	0.0353	0.4721	0.9726	1.0000	1.0000

3. An Empirical Data Analysis

We conducted a Mendelian randomization analysis of the effect of age of menarche on total pubertal height growth and late pubertal height growth using the winner's-curse-corrected method and the SMR method. Previously, we used the inverse-variance weighted (IVW) method [5] and the MR-Egger regression method [6] on these exposures and outcomes [15]. In that study, to avoid the winner's curse caused by the selection of IV SNPs, two other GWAS studies on age at menarche from an MR-Base database were used for validation. IV SNPs were significant in the main GWAS for age at menarche but not in the other two other GWASs which were removed. Such a procedure helps to avoid IV SNPs that are close to the selection threshold. In this study, we use all significant IV SNPs without further validation.

GWAS summary data were retrieved from the MR-Base database (http://www.mrbase.org/ accessed on 27 November 2022). At the genome-wide significance level 5×10^{-8}, 117 instrument SNPs were selected from a previous study on age at menarche with 182,413 females of European ancestry [28]. After pruning for linkage disequilibrium, there are 84 SNPs left. The GWAS summary statistics on adult height were obtained from a study with 4946 females of European ancestry [29]. Thus, the population of this study matches that of the study on age at menarche.

For each SNP, the winner's-curse-corrected estimate of b and a support are computed in addition to the SMR estimate and the associated confidence interval. To correct for the 84 IV SNPs, the support is 5.9-unit since $\Pr(\chi_1^2 > 2 \times 5.9) = 0.05/84$ and the nominal coverage of the confidence interval is 0.9994 (=1 − 0.05/84). As discussed previously, this support excludes 0 if and only if the T statistic is significant at the level 0.05/84. The p-value for the winner's-curse-corrected method is based on the T statistic. SNPs whose supports or confidence intervals do not contain 0 are shown in Table 2.

The estimates of b from the winner's-curse-corrected method and the SMR method are pretty close to each other for the SNPs shown in Table 2, as are the support and the confidence interval. This is due to the high significance of the association of these SNPs with the age at menarche (p-values: 4.552×10^{-15} for rs7514705 and rs7642134; $<4.552 \times 10^{-15}$ for rs7759938). For both total and late pubertal height growth, the T statistic is more significant than the T_{SMR} statistic. For late pubertal height growth, SNP rs7514705 is significant for the T statistic but not for the T_{SMR} statistic.

Table 2. Results for the effects of age at menarche on total pubertal height growth and late pubertal height growth. To correct for the 84 IV SNPs, the support is 5.9-unit and the nominal coverage of the CI is $0.9994 (= 1 - 0.05/84)$. This support excludes 0 if and only if the T statistic is significant at the level $0.05/84$. The p-value is for the null $H_0 : b = 0$. It is computed from the T statistic (the winner's-curse-corrected method) or the T_{SMR} statistic (the SMR method).

Winner's-Curse-Corrected Method			
SNP	**Gene Name**	$\hat{b}$ **(5.9-Unit Support)**	***p*-Value**
Total pubertal height growth			
rs7514705	TNNI3K	2.048 (0.889, 3.807)	8.856×10^{-6}
rs7642134	POU1F1	2.474 (1.264, 4.433)	1.117×10^{-7}
Late pubertal height growth			
rs7514705	TNNI3K	1.822 (0.057, 5.091)	5.024×10^{-4}
rs7759938	LIN28B	0.931 (0.335, 1.571)	2.756×10^{-7}
SMR Method			
SNP	**Gene Name**	$\hat{b}_{\text{SMR}}$ **(99.94% CI)**	***p*-Value**
Total pubertal height growth			
rs7514705	TNNI3K	2.042 (0.330, 3.754)	1.108×10^{-4}
rs7642134	POU1F1	2.466 (0.647, 4.284)	1.110×10^{-5}
Late pubertal height growth			
rs7759938	LIN28B	0.931 (0.330, 1.533)	5.142×10^{-7}

Another empirical application on the conditional test T is the study of schizophrenia, which was shown in our previous publication [9]. The T statistic identified some strong candidate genes (e.g., AKT3, RGS6, and KCNN3) for schizophrenia that are missed by the SMR method.

4. Discussion

Previously, we proposed a test statistic T for testing $H_0 : b = 0$ [15]. The current work extends the previous work by focusing on the point and interval estimate of the causal effect b. Because the "sample size" for the MR analysis is 1, the standard likelihood theory does not apply. As a result, it is not straightforward to construct a confidence interval.

We considered two extreme scenarios: one being the one-sample individual-level data and the other being independent-sample summary data. In addition, on of these scenarios is without the winner's curse caused by the selection of IV SNPs and the other suffers from the winner's curse. For one-sample individual-level data that is free of the winner's curse, the SMR method is the same as the TSLS method, not only in terms of the estimates of the causal effect size but also in terms of the variance of the estimates. For two independent-sample summary data with a selected SNP, the SMR test for $H_0 : b = 0$ is less powerful than the conditional test we proposed earlier [9]. Confidence intervals derived from the SMR method have poor coverage compared to their nominal levels. In comparison, the supports we proposed have stable coverage, at least in our simulation studies.

There are reports (also see our empirical data analyses) showing that the winner's curse may not have substantial impact on the MR estimates [12]. This is because in these cases the SNPs are strong. As indicated by Figure 1, the winner's curse affects the relatively weak IV SNPs most. These are the SNPs with $|\hat{b}_{gx}/SE(\hat{b}_{gx})| < 7.5$ (i.e., $p < 6.38 \times 10^{-14}$). For the strong SNPs, there is not much difference between x and its maximum likelihood estimate. An SNP is strong when either the effect size b or the sample size in the exposure GWAS is, or both, are large. The three-sample design [14] also helps in making an SNP strong by increasing the chance that the selected SNPs are highly significant. Theoretically, however, it does not eliminate the winner's curse as the probability that a weak SNP is significant in the discovery GWAS *and* the exposure GWAS is non-zero.

In the previous paragraph, the meaning of the term "weak" may be different than weak instrument in the usual sense although there is no universally-accepted definition of weak instrument. It is relative to the threshold for selecting IV SNPs. SNPs that barely pass the threshold are always weak. In comparison, a weak instrument in the usual sense seems to be characterized in absolute sense, for instance, the F-statistic for testing $H_0 : b_{gx} = 0$ is less than 10 [30].

Although Equation (1) is on continuous traits, the proposed winner's-curse-corrected method works for dichotomous traits because it is based on the approximate normality on $\hat{b}_{gx}$ and $\hat{b}_{gy}$.

The current study focuses on a single SNP analysis. A major advantage of such an analysis over multiple SNPs such as the IVW method and the MR-Egger regression method is that it involves less assumptions. For example, the causal effects at different SNPs are allowed to be different. An interesting topic would be to generalize the current work to the case of using multiple SNPs simultaneously.

Our winner's-curse-corrected method is designed for two independent (i.e., non-overlapping) samples only. This is a limitation although it is not uncommon for methodology development, for instance, [14]. In practice, the study subjects for the exposure GWAS and the outcome GWAS may overlap [12]. The likelihood function $L(\mu_x, \mu_y)$ will be different than what is presented here. The conditional test T needs to be revised and the concept of support is still applicable. Future research on this topic is warranted.

The winner's-curse-corrected method has been implemented in the R package `iGasso`.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: GWAS summary data used in this research were retrieved from the MR-Base database (http://www.mrbase.org/ accessed on 27 November 2022).

Acknowledgments: The author would like to thank two anonymous reviewers for their helpful comments.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GWAS	genome-wide association study
IV	instrumental variable
MR	Mendelian randomization
SNP	single nucleotide polymorphism
TSLS	two-stage least-squares

References

1. Hemani, G.; Tilling, K.; Smith, G.D. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **2017**, *13*, e1007081.
2. Zhu, Z.; Zhang, F.; Hu, H.; Bakshi, A.; Robinson, M.R.; Powell, J.E.; Montgomery, G.W.; Goddard, M.E.; Wray, N.R.; Visscher, P.M.; others. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **2016**, *48*, 481. [CrossRef]
3. Morrison, J.; Knoblauch, N.; Marcus, J.H.; Stephens, M.; He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.* **2020**, *52*, 740–747. [CrossRef]
4. Davey Smith, G.; Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **2003**, *32*, 1–22. [CrossRef] [PubMed]
5. Burgess, S.; Butterworth, A.; Thompson, S.G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **2013**, *37*, 658–665. [CrossRef] [PubMed]
6. Bowden, J.; Davey Smith, G.; Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **2015**, *44*, 512–525. [CrossRef] [PubMed]

7. Hemani, G.; Zheng, J.; Elsworth, B.; Wade, K.H.; Haberland, V.; Baird, D.; Laurin, C.; Burgess, S.; Bowden, J.; Langdon, R.; others. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **2018**, *7*, e34408. [CrossRef]
8. Zhao, Q.; Wang, J.; Hemani, G.; Bowden, J.; Small, D.S.; others. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Stat.* **2020**, *48*, 1742–1769. [CrossRef]
9. Wang, K.; Han, S. Effect of selection bias on two sample summary data based Mendelian randomization. *Sci. Rep.* **2021**, *11*, 7585. [CrossRef]
10. Ye, T.; Shao, J.; Kang, H. Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *Ann. Stat.* **2021**, *49*, 2079–2100. [CrossRef]
11. Bigdeli, T.B.; Lee, D.; Webb, B.T.; Riley, B.P.; Vladimirov, V.I.; Fanous, A.H.; Kendler, K.S.; Bacanu, S.A. A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans. *Bioinformatics* **2016**, *32*, 2598–2603. [CrossRef]
12. Jiang, T.; Gill, D.; Butterworth, A.S.; Burgess, S. An empirical investigation into the impact of winner's curse on estimates from Mendelian randomization. *medRxiv* **2022**. [CrossRef]
13. Forde, A.; Hemani, G.; Ferguson, J. Review and further developments in statistical corrections for Winner's Curse in genetic association studies. *bioRxiv* **2022**. [CrossRef]
14. Zhao, Q.; Chen, Y.; Wang, J.; Small, D.S. Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *Int. J. Epidemiol.* **2019**, *48*, 1478–1492. [CrossRef]
15. Jo, E.J.; Han, S.; Wang, K. Estimation of Causal Effect of Age at Menarche on Pubertal Height Growth Using Mendelian Randomization. *Genes* 2022, *in press*. [CrossRef]
16. Hannon, E.; Gorrie-Stone, T.J.; Smart, M.C.; Burrage, J.; Hughes, A.; Bao, Y.; Kumari, M.; Schalkwyk, L.C.; Mill, J. Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *Am. J. Hum. Genet.* **2018**, *103*, 654–665. [CrossRef]
17. Lee, B.; Yao, X.; Shen, L. Integrative analysis of summary data from GWAS and eQTL studies implicates genes differentially expressed in Alzheimer's disease. *BMC Genom.* **2022**, *23*, 414. [CrossRef]
18. Porcu, E.; Rüeger, S.; Lepik, K.; Santoni, F.A.; Reymond, A.; Kutalik, Z. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **2019**, *10*, 3300. [CrossRef]
19. Porcu, E.; Sadler, M.C.; Lepik, K.; Auwerx, C.; Wood, A.R.; Weihs, A.; Sleiman, M.S.B.; Ribeiro, D.M.; Bandinelli, S.; Tanaka, T.; others. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nat. Commun.* **2021**, *12*, 5647. [CrossRef]
20. Zhu, Z.; Zheng, Z.; Zhang, F.; Wu, Y.; Trzaskowski, M.; Maier, R.; Robinson, M.R.; McGrath, J.J.; Visscher, P.M.; Wray, N.R.; others. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **2018**, *9*, 224. [CrossRef]
21. Jin, C.; Lee, B.; Shen, L.; Long, Q.; Initiative, A.D.N.; others. Integrating multi-omics summary data using a Mendelian randomization framework. *Briefings Bioinform.* **2022**, *23*, bbac376. [CrossRef] [PubMed]
22. Pustejovsky, J.E. 2SLS Standard Errors and the Delta-Method. 2017. Available online: https://www.jepusto.com/delta-method-and-2sls-ses/ (accessed on 11 November 2022).
23. Greene, W.H. *Econometric Analysis*, 6th ed.; Pearson-Prentice Hall: New York, NY, USA, 2008.
24. Zhao, Q.; Wang, J.; Spiller, W.; Bowden, J.; Small, D.S. Two-sample instrumental variable analyses using heterogeneous samples. *Stat. Sci.* **2019**, *34*, 317–333. [CrossRef]
25. Burgess, S.; Dudbridge, F.; Thompson, S.G. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat. Med.* **2016**, *35*, 1880–1906. [CrossRef] [PubMed]
26. Ghosh, A.; Zou, F.; Wright, F.A. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am. J. Hum. Genet.* **2008**, *82*, 1064–1074. [CrossRef]
27. Edwards, A.W.F. *Likelihood*; CUP Archive: New York, NY, USA, 1984.
28. Perry, J.R.; Day, F.; Elks, C.E.; Sulem, P.; Thompson, D.J.; Ferreira, T.; He, C.; Chasman, D.I.; Esko, T.; Thorleifsson, G.; others. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **2014**, *514*, 92–97. [CrossRef]
29. Cousminer, D.L.; Berry, D.J.; Timpson, N.J.; Ang, W.; Thiering, E.; Byrne, E.M.; Taal, H.R.; Huikari, V.; Bradfield, J.P.; Kerkhof, M.; others. Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal height growth, pubertal timing and childhood adiposity. *Hum. Mol. Genet.* **2013**, *22*, 2735–2747. [CrossRef]
30. Staiger, D.O.; Stock, J.H. *Instrumental Variables Regression with Weak Instruments*; Cowles Foundation Discussion Papers: London, UK, 1994.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.