Research Synthesis Methodology for Normative Data and Genetic Data

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

WENHAO CAO

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISED BY DR. HAITAO CHU,
DR. LIANNE K. SIEGEL,
DR. SAONLI BASU

November, 2023

Acknowledgements

I would like to express my deepest gratitude to the following individuals and organizations who have played a significant role in the completion of this PhD thesis. First and foremost, I am immensely thankful to my advisors Dr. Haitao Chu, Dr. Lianne Siegel, and Dr. Saonli Basu, for their unwavering guidance, invaluable insights, and endless patience throughout my PhD degree journey. Thank you for encouraging me to become an independent researcher, and patiently advising me on preparing manuscripts and presentations. Your mentorship has been invaluable, and I am truly grateful for the knowledge and wisdom you have shared with me. I extend my sincere appreciation to my advisors and Dr. Sue Duval for serving on my doctoral committee and for their constructive feedback and scholarly contributions. Your expertise has greatly enriched the quality of this thesis. I am indebted to my colleagues and friends who provided support, encouragement, and countless hours of stimulating discussions. I would like to acknowledge the NIH National Center for Advancing Translational Sciences grant (UL1TR002494), National Library of Medicine grant (R01LM012982), National Heart, Lung, and Blood Institute grant (T32HL129956), National Institute on Drug Abuse grant (5R21DA046188), and National Cancer Institute grant (1R01CA266253) for their financial support, which made this research possible. I am grateful to the staff and resources at the University of Minnesota Division of Biostatistics and Health Data Science for providing a conducive environment for research and learning. The facilities and assistance provided were invaluable in the successful execution of this work. I would like to thank Dr. Jincheng Zhou, Dr. Motao Zhu, Dr. Tiejun Tong, Dr. Yong Chen, and Dr. Timothy Hanson for their contributions to the work on meta-analysis for estimating reference intervals in Chapter 3 and 4, as well as Dr. Zhaotong Lin and Dr. Seon-Kyeong Jang for their help to the Mendelian randomization work on Chapter 5.

Finally, I would like to thank Dr. Xianghua Luo for her support and kindness throughout my master's degree journey, thank you for inspiring and encouraging me to pursue this PhD degree.

Dedication

I dedicate this thesis to my parents, Shubin Cao and Airong Zuo, whose unwavering support and belief in my potential have been my driving force. Your love and support have made it possible for me to embark on this academic journey and reach this significant milestone.

Abstract

A reference interval represents the normative range for measurements from a healthy population. It can be interpreted as a prediction interval for a new individual from the overall population. The reference interval based on one study might not be applicable to a broader population. Meta-analysis can provide a more generalizable reference interval based on the combined population by synthesizing results from multiple studies. However, existing random effects methods may give imprecise estimates of the between-study variation with only a few studies. In addition, the normal distribution of underlying study-specific means, and equal within-study variance assumption in these methods may be inappropriate in some settings. In the first paper, we develop a mixture distribution method using the fixed effects model. It combines studies by assuming the overall population is a mixture of subpopulations comprised of individual studies. This mixture distribution method does not explicitly estimate the between-study heterogeneity, which is difficult for a random effects model with few studies. In the second paper, We propose a Bayesian nonparametric (NP) model with more flexible assumptions to extend random effects meta-analysis for estimating reference intervals. The simulation studies show the performance of the mixture distribution and NP approaches when the assumptions of normally distributed study mean and equal within-study variances do not hold. Both methods are applied to real datasets and provide more reasonable estimates for reference intervals compared with existing methods.

The third paper focuses on developing a new Mendelian randomization (MR) approach, which leverages genetic data to estimate the causal effect of an exposure factor on an outcome from observational studies. We utilize genetic correlations to summarize information on a large set of genetic variants associated with the exposure factor. Our proposed two-stage random effects approach (TS-RE) can accommodate many weak and pleiotropic effects. Our approach quantifies the variation explained by all included instrumental variables instead of estimating the individual effects and thus could accommodate weak IVs. This is useful for performing MR estimation in small studies where the selection of valid IVs is unreliable and thus has a large influence on the MR estimation. Through simulation and real data analysis, we demonstrate that our approach provides a robust alternative to

the existing methods.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Meta-analysis has been widely used in medical and genetic studies over the past several decades. A PubMed search for "meta-analysis" returns 264,112 results as of January 2023, with 130,244 hits since 2018. A meta-analysis combines results from multiple studies to increase the sample size, providing more precise estimates of the parameters of interest and thus can be regarded as an integral part of deriving evidence-based results. Reference interval refers to a normative range for measurements from a healthy population. It provides a reference for physicians to determine whether the measurement falls into the normal range or not. However, a reference interval based on a single study population may not be representative of the broader population. [1] Meta-analysis allows for combining results from multiple independent studies and can provide a general reference interval based on the overall population. [2] Since the reference interval can be defined as a prediction interval for the measurement of a new individual, [1, 3] the confidence interval for the pooled mean and the prediction interval for the mean of a new study can not capture the appropriate range for a new individual. Siegel et al. [2] proposed three methods for estimating the reference interval based on a random effects meta-analysis. However, there are some hidden assumptions in a typical random effects model that can be violated: 1) the normal distribution for individuals within each study, 2) the normality assumption of the underlying study-specific means, and 3) the equal within-study variance across studies. The first and second paper of this thesis focuses on relaxing those assumptions by using the fixed effects meta-analysis and nonparametric Bayesian model.

The second chapter of this thesis demonstrates how a fixed-effects model can be applied to meta-analysis with a few number of studies (less than five). The fixed-effects model assumes the included studies are independent the study means are unrelated, and that they are not random samples from one common distribution like the random effects model. [4] More flexible distribution assumptions can be used under a fixed effects model. Chapter 2 proposes a mixture distribution method only assuming parametric distributions (e.g. normal) for individuals within each study and integrating them to form the overall population distribution. The empirical method proposed by Siegel et al. [2] is also extended from normal distribution to any parametric distribution for the overall population. The simulations in Chapter 2 demonstrate the performance of the proposed method, and two case studies for estimating the reference intervals for urination frequency during day time and human postural vertical compare the results of the fixed-effects-based methods and random effects-based methods.

In the random effects model, the underlying normal distribution for study means might not be appropriate (e.g. the data are skewed), then other distributions such as Student's t might be considered. [5] However, the degree of freedom of the t distribution is difficult to determine and the restrictive unimodal and symmetric shape is improper in most applications. Chapter 3 proposes a Bayesian nonparametric method to avoid assuming the parametric forms of underlying distribution by using a Dirichlet process (DP). This Bayesian nonparametric method incorporates infinitely many parameters in order to more flexibly represent uncertainty in the underlying distribution. The DP can be regarded as an infinite mixture distribution. To differentiate this NP method from the mixture distribution method in Chapter 2, it is essential to know that the number of mixture components for DP is unknown, which is equal to the number of included studies for the mixture distribution method in Chapter 2. In addition, the DP process can be also used to assume unequal within-study variance. We apply two independent DP processes to relax both the between-study normal distribution and equal within-study variance assumption, which we call NP-2 in Chapter 3. We apply the proposed NP and NP-2 methods to a meta-analysis focused on the measure of wake time after sleep onset (WASO). The 23 study means are heavily skewed from a normal distribution and the proposed methods provide better estimates for the reference intervals while other existing random effects model based

methods give extremely wide reference intervals that would cover all healthy and unhealthy individuals.

The third paper of this thesis focuses on using genetic data to estimate the causal effect of a risk factor. Recent advances in genotyping technology have delivered a wealth of genetic data, which is rapidly advancing our understanding of the underlying genetic architecture of complex diseases. Mendelian randomization is a method of using measured variation in genes to examine the causal effect of a modifiable exposure on disease in observational studies when there are unmeasured confounders. In Chapter 4, we propose an approach to utilize genetic correlations to summarize information on a large set of genetic variants associated with the risk factor. This approach provides an alternative way of estimating causal effects using Mendelian Randomization in the presence of many weak and invalid instruments. We use a method-of-moment estimator to estimate the causal effect and demonstrate through extensive simulation studies that our approach provides a robust alternative to the existing MR methods. In particular, through theoretical derivations, we show that our approach is conceptually similar to a weighted average of the widely used inverse-variance weighting (IVW) and Egger regression approaches. Our approach focuses on modeling the second-order moments such as the genetic variance and covariance components of multiple valid instruments for both exposure and outcome variables. Through extensive simulations, we compare its performance with other methods and estimate the effect of body mass index on systolic blood pressure in the black British samples of the UK Biobank. The findings demonstrate the superiority of our method in terms of bias reduction, efficiency, and robustness. The proposed method addresses the challenges associated with utilizing genetic variants as instruments and provides more reliable causal effect estimates.

Chapter 5 summarizes the major findings and discusses future work.

Chapter 2

Estimating the Reference Interval from a Fixed Effects Meta-Analysis

2.1 Introduction

In medical sciences, a reference interval is the range of values that is considered normal for continuous measurement in healthy individuals (for example, the range of blood pressure, or the range of hemoglobin level). We generally expect the measurements of a specified proportion (typically 95%) of a healthy population to fall within this interval. This can also be interpreted as a prediction interval for a measurement of a new healthy individual from the population. Reference intervals are provided for many laboratory measurements and are widely used to decide whether an individual is healthy or not. There are two limitations when scientists use the reference interval estimated from a single (particularly small) study for the general population. First, the samples from a single study may not be representative. Second, the reference interval estimated by a small sample size will likely have high uncertainty. [1] In some cases, only 20 to 40 individuals in a particular group are available in a study to estimate the reference range, potentially leading to large variations in the resulting upper or lower limits. [3] Meta-analysis offers a competitive solution by using samples from multiple studies to establish a reference interval. [6–16] The reference interval estimated from a meta-analysis should account for both the within and between-study variation to reflect the distribution of the general population.

The pooled mean has been reported by some meta-analysis studies as a "reference value", which can only provide information on whether an individual might be above or below the average. [13, 14] Although many meta-analysis studies reported the 95% confidence interval (CI) of the pooled mean as the reference interval, [7–9, 12] this interval only explains the uncertainty of the pooled mean, not the predicted range for a new individual. Another interval called the "prediction interval" is also commonly reported in some meta-analyses, but it is for predicting the mean of a new study, and does not capture the appropriate range for a new individual. [17, 18] This article aims to estimate the reference interval for the overall population to predict the range of measurement of a new individual by synthesizing evidence from multiple studies. We are not interested in predicting the mean of a new study, nor the confidence interval of the pooled mean.

Siegel et al. [2] recently proposed a frequentist and a Bayesian method, and an empirical approach for estimating the reference interval from a meta-analysis. However, the frequentist and Bayesian methods, which are based on the random effects model, may lead to inaccurate inference when the number of studies is small (≤ 5). [5, 19] Three assumptions are required in the frequentist and Bayesian methods: 1) the normal distribution for individuals within each study, 2) the normality assumption of the underlying study-specific means, and 3) the equal within-study variance across studies. Those assumptions can also be violated. Following the independent parameters assumption in the fixed effects model, [4] we propose a mixture distribution method to estimate the reference interval, which may be more suitable when the number studies is small and/or when some assumptions required by the random effects model in Siegel et al. [2] are not valid.

In Section 2, we first review the fixed effects meta-analysis. Then, we review the empirical method proposed by Siegel et al. [2] which only makes a normal assumption (or more generally a two-parameter exponential family distribution) for the pooled population of all studies. We further extend the fixed effects meta-analysis and proposes the mixture distribution method. The mixture distribution method only makes a distribution assumption for individuals within each study. The simulation in Section 3 shows the performance of the two methods under different data generation processes. We used two real data examples to demonstrate the application of the two methods in Section 4, and a discussion follows in Section 5.

2.2 Methods

Suppose only the sample size, mean and standard deviation are available for each study. There are three meta-analysis models that can be used, which differ by their between-study heterogeneity assumptions: the common effect model, the fixed effects model and the random effects model. [4] The common effect model, also called the fixed effect model, assumes no between-study heterogeneity, and that all studies have the same underlying effect. This model has been criticized for attributing the between-study differences only to the sampling variability. [20] The fixed effects model of Laird and Mosteller, [20] which is sometimes confused with the common effect model, assumes that the means were separate and fixed with different within-study variances. This model considers the between-study heterogeneity but asserts that the study effects are unrelated. The random effects model assumes the underlying effects in different studies are independent and identically drawn from a single distribution. [21] This implies that the study effects are somewhat similar and the similarity is governed by the single distribution. [4] The random effects model is frequently chosen if between-study heterogeneity is expected to be present and there is a sufficient number of studies (larger than 5). However, when there are very few studies, the estimate of between-study variance in a random effects model can be highly variable. [5] As a typical approach to the random effects model uses the estimated between-study variance to calculate the inverse variance weights to estimate the pooled mean, [22] this imprecision may lead to a less desirable estimate of the pooled mean and its confidence interval (CI). [23, 24] The imprecise estimate of between-study variance can also lead to inaccurate assessment of the degree of heterogeneity or the degree of similarity across studies. When between-study heterogeneity is expected and only few studies are available, it may be preferable to consider the study-specific effects unrelated and use the fixed effects model. The independent parameters assumption of the fixed effects model implies that the effects of different studies are unrelated, and that they are not random samples from one common distribution like the random effects model. [4] Thus, the fixed effects model cannot directly estimate the overall population distribution and make predictions for a new individual from the study-level summary statistics without making some additional assumptions. To address this limitation, we assume the pooled population of the included studies is representative of the overall population. With this assumption, one can estimate

the reference range for a new individual from the overall population.

We focus on two methods to combine the results from multiple studies and estimate the pooled population distribution, without making the normality assumption of study-specific underlying means and the equal within-study variance assumption. First, we review and extend the empirical method proposed by Siegel et al. [2] which only requires an distributional assumption for an individual in the overall population. We also propose a second method which treats the overall population distribution as a mixture of study-specific distributions, which we call the mixture distribution method. The main difference between the two methods is whether we make a distributional assumption — which can be any distribution completely decided by the mean and variance — for each study or for the overall population. After estimating the overall population distribution, one can use the estimated quantiles to establish the reference intervals. All analyses were performed using R version 4.0.3 (R Core Team), and the R code for real data analysis is provided in the Supplementary Materials.

2.2.1 The fixed effects model

Let y_{ij} denote the jth observation, θ_i be the underlying true mean, and σ_i^2 be the variance for study $i = 1, \ldots, k$. Typically to estimate a reference range, a parametric (e.g. normal) distribution is assumed within each study. Suppose $\bar{y}_i$ is the observed mean, n_i is the sample size for study i, and ϵ_i is a random variable describing the sampling error of study i. The fixed effects model is given by

$$\bar{y}_i = \theta_i + \epsilon_i, Var(\bar{y}_i) = \frac{\sigma_i^2}{n_i}, \tag{2.1}$$

where σ_i^2 can be different across studies and the independent parameter assumption is that θ_i are unrelated. Let μ_{FE} be the pooled mean of the overall population in the fixed effects model, and μ_{FE} is traditionally estimated as a weighted average of study-specific means:

$$\hat{\mu}_{FE} = \sum_{i=1}^{k} \frac{w_i \bar{y}_i}{\sum_{j=1}^{k} w_j}, Var(\hat{\mu}_{FE}) = \sum_{i=1}^{k} \frac{w_i^2 \sigma_i^2}{n_i (\sum_{j=1}^{k} w_j)^2}, \tag{2.2}$$

where σ_i^2 can be estimated by the sample variance $\hat{\sigma}_i^2$. The two most commonly used weights are the inverse variance weights $w_i = \frac{n_i}{\hat{\sigma}_i^2}$ proposed by Hedges and Vevea [25] and the study sample size weights $w_i = n_i$ proposed by Hunter and Schmidt [26]. Marin et al [27] found that the sample size weighted average was a practically unbiased estimator while the inverse variance weighted estimated was slightly biased but had the lowest mean squared error. The pooled mean $\hat{\mu}_{FE}$ and variance $Var(\hat{\mu}_{FE})$ can be used to construct the confidence interval for the pooled mean, but it cannot be used to construct the reference interval predicting the range of the measurement for a new individual from the overall population. We considered the following two methods to estimate the reference interval in a fixed effects meta-analysis.

2.2.2 The empirical method

The empirical method proposed by Siegel et al [2] does not require that the studies have related means or equal within-study variances and therefore can also be used in a fixed effects meta-analysis. This method does not specify the distribution of y_{ij} within each study. However, it assumes that the overall population follows a normal distribution, or more generally any distribution completely determined by its mean and variance. The overall mean across all study populations can be estimated by the average of the study means weighted by their study sample sizes:

$$\hat{\mu}_{emp} = \frac{\sum_{i=1}^{k} n_i \bar{y}_i}{\sum_{i=1}^{k} n_i}. \tag{2.3}$$

This $\hat{\mu}_{emp}$ is equivalent to the $\hat{\mu}_{FE}$ since they use the same weights. Then the marginal variance across studies can be estimated using the conditional variance formula $Var(y) = E[Var(y_{ij}|S=i)] + Var[E(y_{ij}|S=i)]$:

$$\hat{\sigma}_{emp}^2 = \frac{\sum_{i=1}^{k}(n_i-1)\hat{\sigma}_i^2}{\sum_{i=1}^{k}(n_i-1)} + \frac{\sum_{i=1}^{k}(n_i-1)(\bar{y}_i - \hat{\mu}_{emp})^2}{\sum_{i=1}^{k}(n_i-1)}, \tag{2.4}$$

where the weights $n_i - 1$ give an unbiased estimate of the variance. [2] The limits of the α-level reference interval are then given by the $100 \times \alpha/2$ and $100 \times (1-\alpha/2)$ percentiles of a $N(\hat{\mu}_{emp}, \hat{\sigma}_{emp}^2)$ distribution: $\hat{\mu}_{emp} \pm z_{1-\alpha/2}\hat{\sigma}_{emp}$, where $z_{1-\alpha/2}$ is the standard normal

critical value for the chosen significance level α.

2.2.3 The mixture distribution method

The mixture distribution method estimates the reference interval by integrating the distribution function constructed by each study mean and variance. The study-specific distribution $F_i(y)$ needs to be specified parametrically but there is no need to assume the same parametric distribution for each study, e.g. a normal distribution for all studies. The observations in study i can be assumed to follow any continuous distribution completely determined by the mean θ_i and variance σ_i^2, such as those from the two parameter-exponential families. The variances σ_i^2 can differ across studies. In the fixed effects model, the population mean μ_{FE} is estimated by the weighted average of the study-specific means. Similarly, we assume the overall population has a mixture distribution of individual study populations with weight w_i:

$$F(y) = \sum_{i=1}^{k} \frac{w_i F_i(y)}{\sum_{j=1}^{k} w_j} \tag{2.5}$$

where $F(\cdot)$ is the cumulative distribution function. For each study, the distribution $F_i(y)$ can be determined approximately by the observed sample mean $\bar{y}_i$ and sample variance $\hat{\sigma}_i^2$. Then, a $100 \times (1-\alpha)\%$ reference interval, $[L, U]$, based on the pooled population can be estimated by solving the following equations:

$$\begin{cases} \sum_{i=1}^{k} \frac{w_i \hat{F}_i(L)}{\sum_{j=1}^{k} w_j} = & \alpha/2 \\ \sum_{i=1}^{k} \frac{w_i \hat{F}_i(U)}{\sum_{j=1}^{k} w_j} = & 1-\alpha/2, \end{cases} \tag{2.6}$$

where $\hat{F}_i(\cdot)$ is the estimate of the cumulative distribution function of $F_i(y)$.

When y_{ij} can be assumed to be approximately normally distributed, the study-specific cumulative distribution function can be approximately by $\hat{F}_i = \phi(\bar{y}_i, \hat{\sigma}_i^2)$. When the normality assumption of y_{ij} does not hold, another parametric distribution should be used. For example, if the observed measurements have a skewed distribution or when the values cannot be negative, assuming a log-normal distribution where $ln(y_{ij}) \sim N(\theta_i, \delta_i^2)$ may be more appropriate. In this case, one will need to estimate the mean θ_i and variance δ_i^2 in the log scale from the observed sample mean $\bar{y}_i$ and sample variance $\hat{\sigma}_i^2$ in the original scale

as $\hat{\theta}_i = ln\left(\frac{\bar{y}_i}{\sqrt{1+\hat{\sigma}_i^2/\bar{y}_i^2}}\right)$ and $\hat{\delta}_i^2 = ln\left(1+\hat{\sigma}_i^2/\bar{y}_i^2\right)$. [28]

This mixture distribution method does not require assuming a normal distribution for the overall population but does require a distributional assumption for each study. Moreover, the parametric distributions within each study can be different; we merely use the same distribution in this paper for convenience. We choose the sample sizes as the weights in the mixture distribution method, though other weights such as the inverse variance weights can also be used.

2.3 Simulation

To assess the performance of the mixture distribution method and compare it with the empirical method, we generated the measurements within each study from a normal, a log-normal or a gamma distribution. Following the simulation conducted by Siegel et al, [2] the true overall mean μ_{FE} was set to be 8 and the total variance was 1.25 for all three distributions. A between-study variance, $\tau^2 = 1.25 - E(\sigma_i^2)$, was introduced to generate different study-specific means. The true within-study standard deviations were generated from a doubly-truncated normal distribution $\phi(\mu = X, \sigma^2 = 1, a = X, b = X + 1)$, with both the left truncation point and mean equal to X and the right truncation point equal to $X + 1$, for X ranging from 0 to 0.64, with increments of 0.02. For each X, we estimated $E(\sigma_i^2)$ by simulating from the doubly-truncated normal distribution. We then set τ^2 to be equal to $1.25 - E(\sigma_i^2)$ to keep the total variance constant across conditions. Each individual measurement (y_{ij}) was simulated from the following full conditional distributions:

$$\begin{aligned} \theta_i|\mu_{FE}, \tau^2 &\sim F_i(\mu_{FE}, \tau^2), \tau^2 = 1.25 - E(\sigma_i^2); \\ y_{ij}|\theta_i, \sigma_i^2 &\sim F_i(\theta_i, \sigma_i^2), \end{aligned} \tag{2.7}$$

where the two parameters in $F_i(\cdot)$ were the means and variances for the normal, log-normal and gamma distributions we assumed. The total number of studies was set to be 2, 5, 10 or 20, with 2 and 5 representing cases with few studies. Each study contained 50 participants. We conducted 1000 simulations for each configuration.

Under each scenario we calculated the fraction of the true population distribution captured by each of the two reference interval methods, which we call the "coverage". The

ratio of between-study variance (τ^2) to the total variance ($\tau^2 + E(\sigma_i^2)$) and the number of studies k included in the meta-analysis influenced the median coverage and the variation (Figure 1). For normally distributed data in Figure 1a, both the mixture distribution and empirical methods generally had coverages near 95% when the between-study variance was small. The median coverage decreased as the between-study variance increased as a fraction of the total variance; this decrease was most pronounced when k was very small ($k = 2$). The extreme heterogeneity would be a problem for the case with very few studies. Compared with the empirical method, the median coverage of the mixture distribution method decreased slightly more quickly with the between-study heterogeneity. The variation in coverage increased as the between study heterogeneity τ^2 increased and decreased as k increased, while the mixture distribution method had a larger variation than the empirical method. The results for the log-normal distribution are shown in Figure 1b with very similar pattern to the normal distribution. Figure 1c showed that two methods provided almost the same results under a gamma distribution assumption.

2.4 Two Case Studies

2.4.1 A meta-analysis of urination frequency during day time

Accurate reference intervals for measurements of bladder function (storage, emptying and bioregulatory) are useful to promote bladder health. They can be used to identify lower urinary tract symptoms and determine whether further evaluation and treatments are needed. Wyman et al [10] conducted a meta-analysis with 24 studies to establish normative reference values for bladder function parameters of noninvasive tests in women, including urination frequencies, voided and postvoid residual volumes and uroflowmetry parameters.

Here, we only focused on the daytime urination frequency data which was available in 5 studies to demonstrate our methods with few studies. The high degree of observed heterogeneity across studies, the large I^2 value (0.859), and the small number of studies suggest that a fixed effects model is more appropriate than a random effects model. We used the log-normal assumption since the urination frequency data could not be negative and the distribution is skewed. Figure 2 used the urination frequency data to illustrate the mixture distribution method. We first estimated the densities for 5 studies and weighted

them by their sample sizes (the blue dashed curves). Then, the 95% reference interval was obtained by letting $\alpha = 0.05$, which is the region of x-axis between two vertical dashed lines. The solid black curve is the density of the pooled population. Figure 3a shows the means (95% CI) and the prediction interval for a new individual for each study, the 95% CI of the overall mean estimated by the fixed effects model, and the reference intervals based on the methods introduced in Section 2. The overall 95% CI based on the pooled mean gave the narrowest interval ($[6.50, 6.76]$), which represents only the precision in the point estimate. The reference intervals for the empirical method ($[3.56, 11.32]$) and mixture distribution method ($[3.53, 11.31]$) were much wider and overlapped with all studies' 95% prediction interval. Wyman et al [10] used the same mixture distribution method and reported a 90% reference interval $[4, 10]$ for the day time urination frequency, and our mixture distribution method had the same result after changing the quantiles to 90%. We also considered a gamma distribution for the measurement in Figure 3b to see the performance of our methods under different distribution assumption. The reference intervals for the empirical method ($[3.31, 11.09]$) and mixture distribution method ($[3.29, 11.09]$) were shifted to the left by 0.2 compared with the results under log-normal assumption. We provided the 95% prediction intervals for a new individual of each study, which is the estimated reference interval if only a single study was available. The prediction intervals for a new observation of each study showed obvious differences representing variation in the study populations. This suggests that a reference interval calculated from a meta-analysis of these studies is more generalizable to the overall population.

2.4.2 A meta-analysis of human postural vertical

The second case study is a meta-analysis of human subjective postural vertical (SPV) measurements, [15] which reflect an individual's ability to perceive whether they are vertical or not. Maintaining vertical posture is an important ability when engaging in daily activities. [29] Vertical perception is also associated with postural control and functionality and can be altered in stroke patients. [30] To measure the SPV, the participants usually sit on a tilting chair with their eyes closed, and verbally instruct an examiner to set the chair to their perceived upright body orientation.

We used the data for frontal SPV from 15 studies that measured the deviation (in

degree) of the specified position from true verticality in the frontal planes. This case study was used to demonstrate the application of our methods when the number of studies is relatively large. The meta-analysis included 15 studies measuring frontal SPV and the heterogeneity I^2 is 0.909. Conceição et al [15] used the empirical method to estimate the pooled mean and standard deviation, then estimated the normal reference interval as $\hat{\mu} \pm 2\hat{\sigma}$ ([-2.87, 3.31]). Siegel et al [2] proposed the frequentist method and Bayesian method, and estimated the reference intervals as [-2.92, 3.15] and [-3.07, 3.20], respectively. We analyzed the data using the fixed effects model to estimate the pooled mean. As we expected, the 95% CI for the pooled mean was narrow ([-0.04,0.27]) and did not reflect the variation between individuals. The reference interval calculated using the empirical method was [-2.89, 3.13], the same as Siegel et al's [2] result. Conceição et al's [15] interval was slightly wider since they used 2 times standard deviation instead of 1.96 and weighted by n when estimating the overall variance. [15] The mixture distribution method gave a relatively narrower interval [-2.97, 3.10], which was still very close to the results of Conceição et al [15] and Siegel et al [2]. Figure 4 shows that the reference intervals estimated using the mixture distribution and empirical methods overlapped with all individual studies' 95% CIs for the mean, while the 95% CI for the pooled mean only included 4 study means and did not account for the variation. The 95% prediction intervals for a new observation of each study demonstrated a high degree of heterogeneity across studies like the first example. These results reflect how our methods incorporate the full variation in the overall population into the estimated reference intervals.

2.5 Discussion

Meta-analysis is a useful method for synthesizing the results of multiple independent studies to address a particular question. In this paper, we described two methods based on the fixed effects assumption to estimate the normal reference intervals for an individual. One method was a mixture distribution method assuming the overall population distribution is a mixture of individual study distribution. The other method was an empirical method assuming a normal distribution of the overall population. [2] The simulation results showed that when using the fixed effects model with a very small number of studies (2 or 5), both methods performed well if the between-study variation was relatively small. However, it

is important to consider whether separate results from individual studies would be more informative than a meta-analysis with few studies. [5] We recommend choosing the meta-analysis when 1) establishing the reference interval based on the pooled population is necessary, and 2) the estimated between-study variation is no more than $30 - 50\%$ of the total estimated variation. It may be preferable to calculate separate reference intervals for each study population rather than using a meta-analysis when the number of studies is very small and the heterogeneity between studies is extremely large. The example of frontal SPV demonstrated that the two methods can give very similar reference intervals as the random effects model when the number of studies is relatively large. It is difficult to predict the study-level mean of a new study or the range of a new individual in a fixed effects model since the included study effects are assumed unrelated and there is no distribution assumption for the underlying study means. Thus, if the random effects assumption that the underlying study-specific means are from the same distribution is valid, and the number of studies is large, then the between-study variance can be precisely estimated. In this case, the random effects model may be preferred to draw inferences about a hypothetical future study and/or individual not included in the meta-analysis.

The within-study normality assumption in meta-analysis "might not always be appropriate", especially for small sample size studies or skewed data. [19, 31] Log or other transformations can be used for skewed data. If the underlying distribution of the data is not normal and the transformation to a normal population is impossible, the mixture distribution method is still feasible as long as a parametric distribution for each study can be assumed. The empirical method does not make any within-study normality assumption but assumes the overall population follows a known distribution belonging to the two-parameter exponential family. Based on the information obtained from the included studies, the choice of the two methods depends on which assumption is more appropriate. In addition to the flexible assumption for the distribution, both methods do not require equal within-study variances.

This paper focuses on the situation that only the study-level data is available, making it impossible to avoid making an assumption for within-study distributions or the overall distribution. If the individual patient data (IPD) are available, other methods for estimating the reference intervals for a single study could likely be extended to the meta-analysis

setting. For example, nonparametric methods could be used without making assumptions about the specific form of the underlying distribution of the data within each study. [3]

Finally, it is important to determine whether the studies included in a meta-analysis have reported measurements from the target population whose reference range is being sought. One suggestion is evaluating the inclusion and exclusion criteria of the meta-analysis based on the population of interest. For example, 16 studies were excluded in Conceição et al [15] because their SPV protocol was not in a seated position or used control groups with non-healthy participants. Furthermore, considering different instruments can be used to get measurements, the reference interval for measurements obtained by one instrument might not be applicable for measurements from other instruments.

Although in this paper we assume that each study reports the sample size, mean, and standard deviation (SD) of the outcome, some articles report the sample median, the minimum and maximum values, and/or the first and third quartiles, especially when the data are skewed. Multiple methods have been proposed to estimate the sample mean and SD by using those summary statistics. [32–35] With the estimated sample mean and SD, those studies can be included in the meta-analysis. Furthermore, for studies that reported those summary statistics in addition to the mean and SD, the quantile-matching estimation (QME) may be used to better estimate the parameters of the within-study distribution. [36,37] While the methods presented in this paper only use aggregated study-level data, future studies may consider estimating the reference interval by combining studies with individual participant data and studies with aggregated study-level data. Future work could also investigate the effect of subject characteristics, such as age, on the normal reference range by incorporating covariates in the meta-regression model. In addition, it may also be fruitful to investigate the impact of small study effects, publication, and other biases on the estimation of reference range. [38–40]

Figure 2.1: **Simulation Results:** The median (line), 2.5%, and 97.5% (shaded area) of the proportion of the true population distribution captured by the estimated 95% reference interval, for different numbers N of studies. The horizontal axis, proportion of between-study variance to the total variance, represent the degree of heterogeneity across studies. Three distributions are assumed: (a) normal distribution; (b) log-normal distribution; (c) gamma distribution.

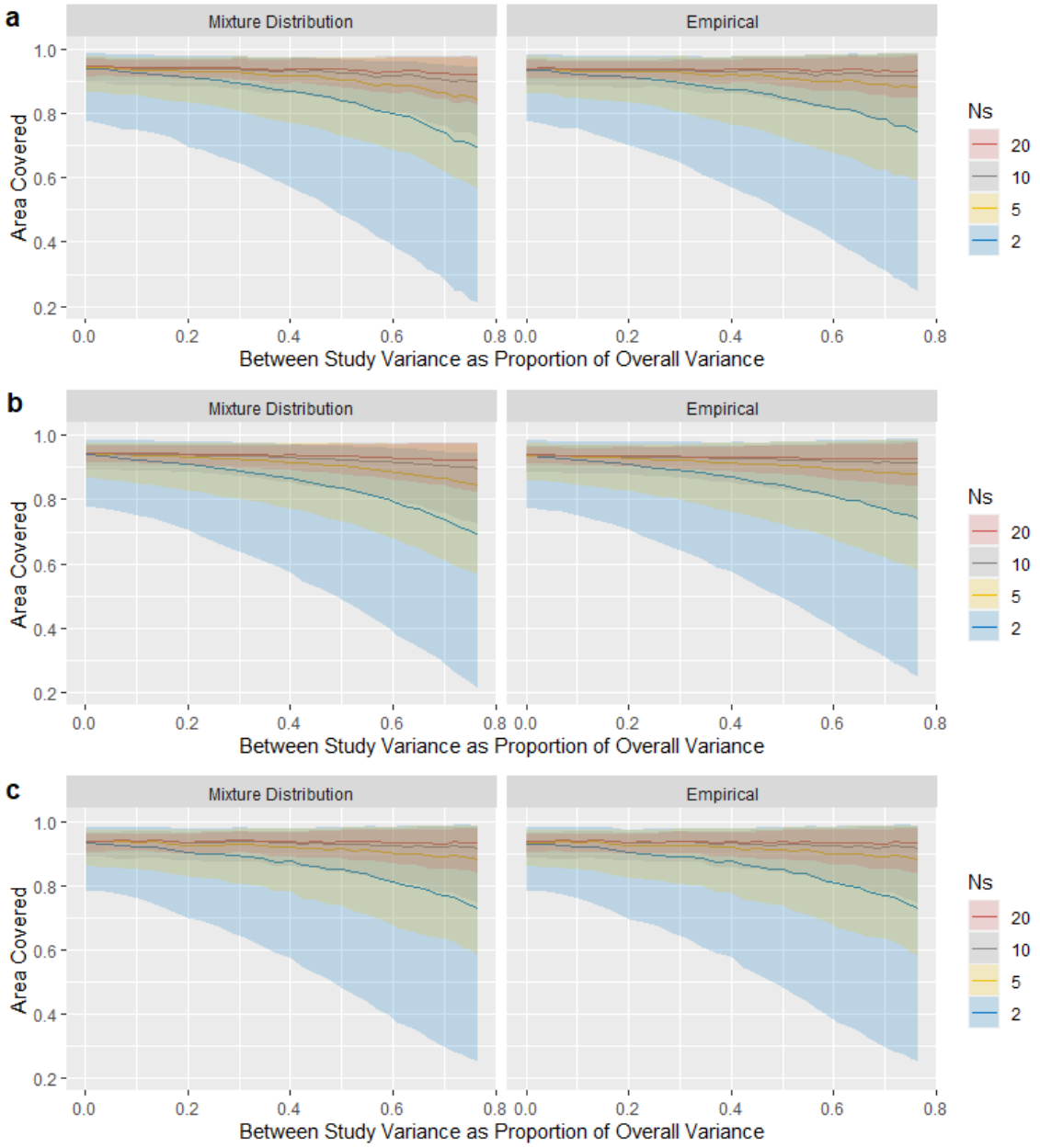


Figure 2.2: **An illustration of the 95% reference interval estimated by the mixture distribution method**: The blue dashed curves are the estimated densities for 5 studies weighted by the sample sizes, and the solid black curve represents the pooled population distribution density. The 95% reference interval is the region of x-axis between two vertical lines, and the sum of area under each blue curve outside the vertical line on each side is equal to 0.025

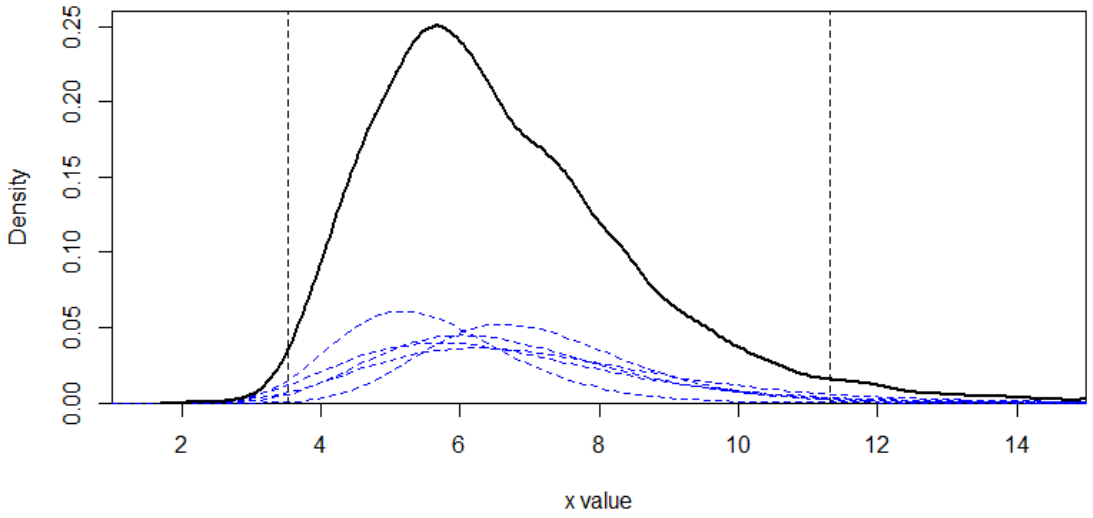


Figure 2.3: **A Meta-analysis of Daytime Frequency:** Mean (95% CI) and 95% prediction interval for a new individual for each study; Overall is the 95% CI for pooled mean estimated by the fixed effects model; 95% reference ranges are estimated from the mixture distribution and the empirical methods under: (a) the log-normal distribution; (b) the gamma distribution.

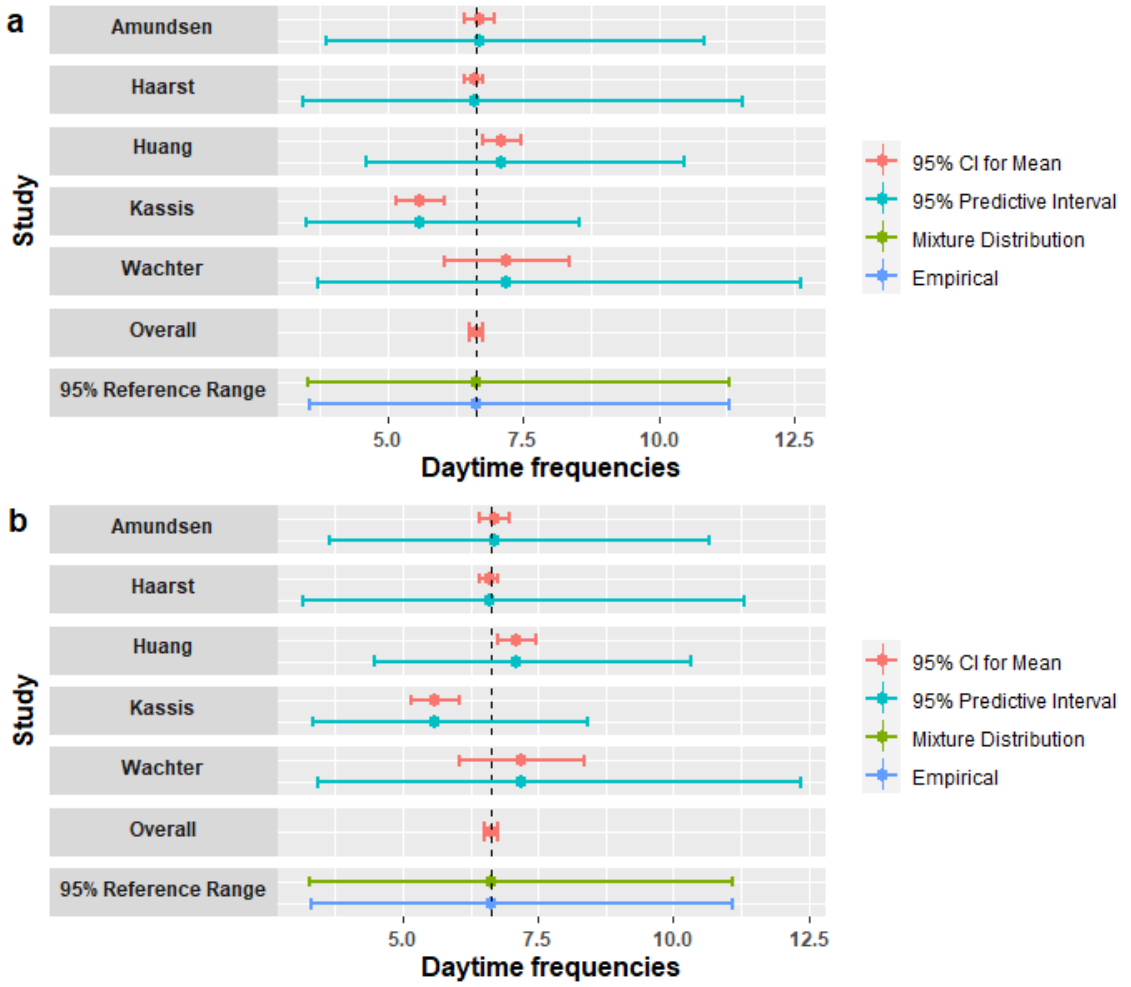
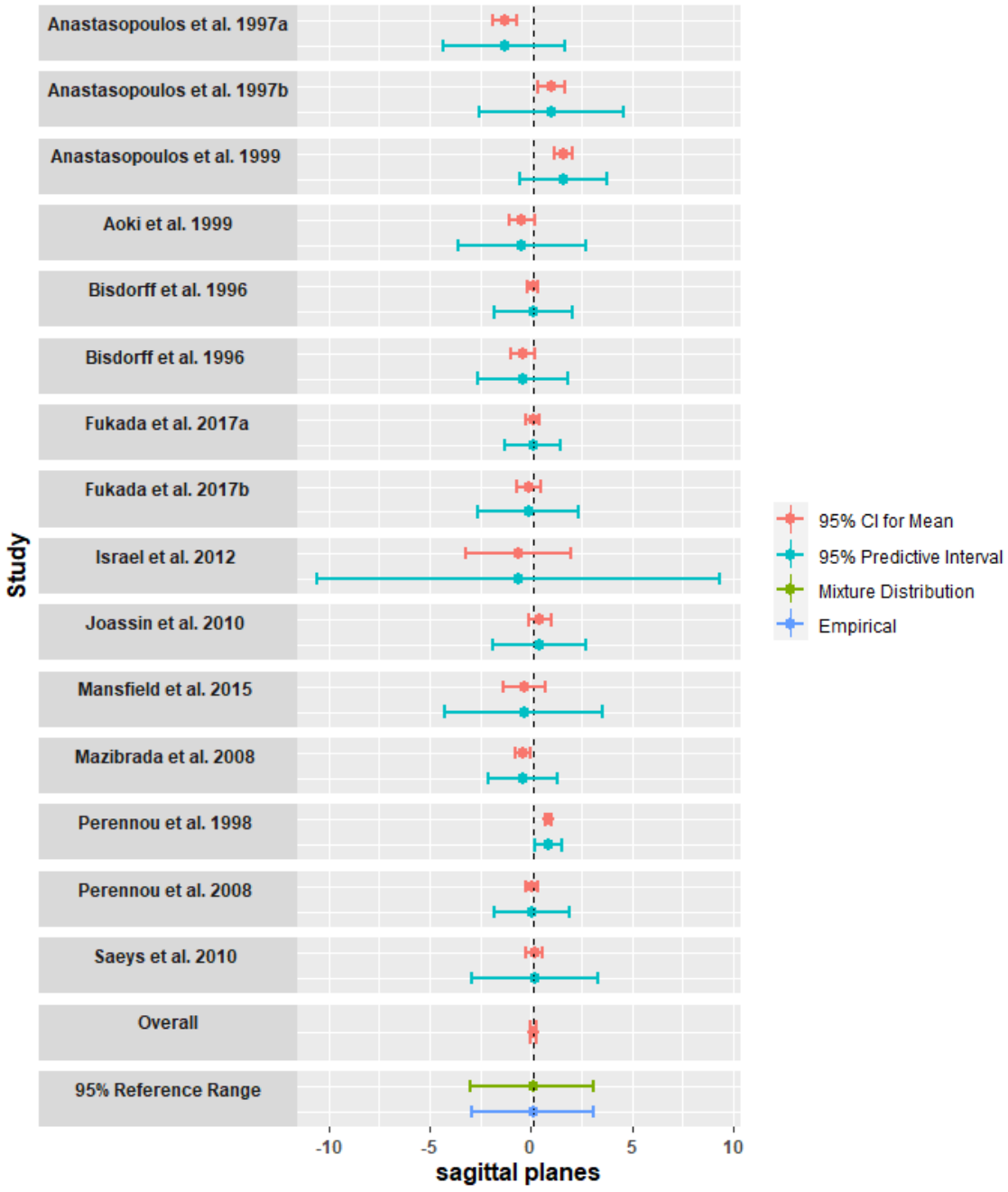


Figure 2.4: **A Meta-analysis of Sagittal Plane SPV:** Mean (95% CI) and 95% prediction interval for a new individual for each study; Overall is the 95% CI for pooled mean estimated by the fixed effects model; 95% reference ranges are estimated from the mixture distribution and the empirical methods under the normal distribution.



Chapter 3

A Bayesian Nonparametric Meta-Analysis Model for Estimating the Reference Interval

3.1 Introduction

In laboratory medicine and clinical studies, reference intervals are used to identify extreme or abnormal measurements. Further clinical investigation or medical diagnosis is indicated if an individual's value is outside the interval. [41] The reference interval, also called the "normal range" or "reference range", is usually defined by the range of values between the 2.5th and 97.5th percentiles of measurements from a healthy population. From a statistical perspective, this can also be regarded as a prediction interval for the measurement of a new healthy individual from the population. Traditional methods for estimating the reference interval from a single study may not fully capture the variability of the entire healthy population and thus have limited applicability due to incomplete information. [1] Thus, a meta-analysis of multiple studies aiming to represent the whole population, incorporating both within and between-study heterogeneity by synthesizing summary statistics reported by multiple studies, may be preferable.

Methods based on a random effects model [2, 42] and a fixed effects model [43] for estimating the reference interval from a meta-analysis have been recently proposed. Based

on the random effects model, the frequentist and Bayesian parametric methods [2] assume both the study-specific means, also called the random effects, and the within-study individual measurements are normally distributed. In this paper, we refer to those two assumptions as between-study normality and within-study normality. A previously proposed empirical method [2] does not require specifying a (normal) distribution for the population within each study but does assume a normal distribution of measurements for the overall population. However, these normality assumptions might not be appropriate for some meta-analyses, in which case a misspecified between-study distribution could induce a systematic bias for the estimated reference intervals. [44–46] The other limitation of the random effects model is that the between-study variance is often inaccurately estimated when only a small number of studies are included, [5] which can lead to imprecise inference about the reference interval. To estimate the reference interval with very few studies and relax these normality assumptions, a mixture distribution method based on the fixed effects model has been proposed, [43] which integrates the distribution function constructed for each study to form an overall population mixture distribution. Any two-parameter exponential family distribution determined by the mean and variance can be used for the mixture distribution method. However, this mixture distribution method does not account for the uncertainty in the estimated means and standard deviations for each study. In addition, the fixed effects model cannot make predictions for a new individual unless assuming the pooled studies completely represent the true overall population. This additional assumption might be inappropriate in the absence of strong clinical evidence or the presence of significant between-study heterogeneity.

This paper aims to relax the between-study normality assumption in the random effects model, with the belief that the study means are related while accounting for the uncertainty in the between-study and within-study variances. The related study means assumption can allow us to make predictions for new individuals or study means without assuming the included studies comprise the entire target population, as in the fixed effects model. Specifically, we assume the random effects follow a nonparametric model while facilitating clustering among studies; this will allow us to approximate a non-normal between-study distribution. This does require that we include a larger number of studies to estimate the clustering structure. There is a rich literature on using semi- and nonparametric Bayesian

hierarchical models to allow for uncertainty in the distribution of the underlying study means, [44,47–51] in which case discrete Dirichlet Processes (DPs) are used to approximate this unknown distribution density. A DP method, which assumes that there are infinitely many clusters in the random effects distribution, is an attractive alternative to mixture models. [52] The DP method can update the components naturally while allocating the samples to those components. Three advantages of the DP method for random effects meta-analysis are (a) allowing the random effects distribution to be unknown without making a parametric specification; (b) avoiding allocating all studies to a fixed number of components; (c) the sub-clustering structure of the DP can capture extra variation in the random effects that do not follow the between-study normality assumption. [50] The other class contains finite mixture models using a reversible jump procedure with an unknown number of mixture components, [53, 54] which requires a dimension-jumping technique to create and delete components. However, the reversible jump procedure is difficult to implement since the jump of dimensions in the model requires more data, and the summary statistics might not be adequate when conducting a meta-analysis to estimate a reference interval.

A random effects meta-analysis assumes the data have a hierarchical structure with two sources of variation: within-study variation and between-study variation. [55] Most meta-analysis methods involve directly plugging in the sample variances from each study as the within-study variances, which is convenient for making inferences for the study-specific means and overall mean. However, it is important to estimate the underlying population within-study variances to make an accurate prediction for a new individual observation. The within-study variances are assumed equal across different studies in previous methods, [2] but this assumption may be violated in some scenarios. Thus, we specify a DP prior to the within-study variances which allows them to vary across studies and follow an unknown distribution. Table 1 summarizes the methods for estimating the reference interval that has been discussed thus far. Our new nonparametric method combines the benefits of all other methods by 1) relaxing the between-study normality assumption and the assumption of equal within-study variances; 2) accounting for the uncertainty in both between-study and within-study variance parameters; 3) predicting a reference interval for a new individual or a new study.

In Section 2 we provide a brief review of random effects models and theoretical properties of the DP and describe our model's likelihood and its hierarchical prior. We first relax the normality assumption on the study-specific means and keep the equal within-study variance assumption in Section 2.1; we then relax the equal within-study variance assumption in Section 2.2. Simulation studies are given in Section 3. In Section 4, we give two real data examples to illustrate the application of our method, and we conclude with a discussion in Section 5. The data and code for real data analysis is provided in the Supplementary Materials.

3.2 Methods

Let y_{ij} denote the jth observation for study $i = 1, \ldots, k$, θ_i be study $i's$ true underlying mean, and σ_i^2 be study $i's$ within-study variance. We consider a random effects model and let τ^2 be the between-study variance. Suppose y_i is the observed mean and n_i is the sample size of study i, then

$$y_i \sim N\left(\theta_i, \frac{\sigma_i^2}{n_i}\right), \quad \theta_i \sim G. \tag{3.1}$$

Although the assumption of a normally distributed $G = N(\mu, \tau^2)$ for random effect θ_i is often used in practice, this assumption is made purely for convenience and can bias inference if not appropriate. [45] In particular, the normal distribution has light tails, encouraging study means to stay close to μ, and discouraging clusters of means commonly seen in heterogeneous collections of studies. Hence, the 'outlier' studies will have excessive influence over the estimated mean treatment effect. In addition, the between-study normality assumption may be too strong when there is considerable heterogeneity among studies. One can relax the between-study normality assumption on the distribution of θ_i by choosing a distribution such as Student's t, Cauchy, or Laplace that allows for the possibility of skewness. [56] However, a heavy-tailed distribution, such as the t, has a restrictive unimodal and symmetric shape, and in many applications assuming such a shape can lead to bias. [57] It is also difficult to derive a reliable analytical estimate of τ^2 for non-normal distributions under a frequentist framework. Flexible models for random-effects distributions can be assumed [58] but these methods still require parametric forms for the underlying distribution. The nonparametric approach to meta-analysis developed here is Bayesian and

uses Markov chain Monte Carlo (MCMC) sampling to generate an accurate approximation to the posterior distribution, which can be used to assess evidence of statistical heterogeneity or variation in the underlying means across studies while relaxing distributional assumptions.

In terms of assessing the assumption of normally distributed study means in a meta-analysis, [59] one can develop a formal test or use a weighted quantile-quantile plot. It has also been suggested to use a Bayes factor for testing a Gaussian null versus a nonparametric Polya-tree alternative. [60] Another alternative is fitting six alternative three- or four-parameter parametric models allowing for varying degrees of heavy-tailedness and/or skew and choosing among them via the deviance information criterion (DIC). [56] However, all models considered are unimodal. Note that the normal-normal assumption implies that the marginal distribution of $y_1, \ldots, y_k$ i.e. $f(y) = \iint \phi(\theta|\mu,\tau)\phi(y|\theta,\sigma^2)g(s)\mathrm{d}\theta\mathrm{d}s$ is symmetric and unimodal. [61] Thus, before using a normal-normal model, a histogram or Q-Q plot of the observed effects $y_1, \ldots, y_k$ can be checked. Standard tests for skewness, e.g. D'Agostino test [62] can be used to provide a check of model assumptions, although this is likely underpowered for meta-analyses with a small number of studies, e.g. $k < 5$. [63]

3.2.1 Equal within-study variances

The Bayesian nonparametric model flexibly incorporates uncertainty in G. Specifically, we assume that the random effects are generated from a DP. Under an equal within-study variance assumption,

$$y_i \sim f\left(\theta_i, \frac{\sigma^2}{n_i}\right), \quad \theta_i \sim G, \quad G \sim DP(\alpha, G_0), \quad i = 1, \ldots, k, \tag{3.2}$$

where $f(\theta_i, \frac{\sigma^2}{n_i})$ is called the kernel distribution with mean θ_i and variance $\frac{\sigma^2}{n_i}$, and G follows a Dirichlet process with baseline distribution G_0 and concentration parameter α. G_0 is a "distributional location" parameter on which the NP distribution is centered, typically chosen to achieve conditional conjugacy. The parameter α is a measure of the strength in the belief that G is G_0; note this happens with probability one when $\alpha \to \infty$. The constructive stick-breaking representation [64] is helpful to illustrate what choosing a $DP(\alpha, G_0)$ prior for G implies about prior beliefs regarding G. $G \sim DP(\alpha, G_0)$ is

equivalent to letting:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\gamma_i}, \gamma_i \sim G_0, \tag{3.3}$$

where $\pi_i = V_i \Pi_{l<i}(1 - V_l)$ is a probability weight that is formulated from a stick-breaking process, with $V_i \sim Beta(1, \alpha)$ for $i = 1, \ldots, \infty$, and δ_γ is a point mass at γ. For our method, we used a truncated DP where the maximum number of distinct components k is the number of included studies, i.e. $i = 1, \ldots, k$. It has been showed the joint distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is the product of successive conditional distributions [65]

$$\theta_i | \alpha, \theta_1, \ldots, \theta_{i-1} \propto \frac{\alpha}{i - 1 + \alpha} \phi_0(\theta_i) + \frac{1}{i - 1 + \alpha} \sum_{j=1}^{i-1} \delta(\theta_j), \tag{3.4}$$

where $\phi_0(\cdot)$ is the density function of $G_0(\cdot)$ and $\delta(s)$ is a degenerate Dirac distribution with point mass at s. The (conditionally) conjugate prior for Gaussian kernel $G_0 = N(\mu, \tau^2)$ is a normal distribution for μ and inverse gamma distribution for τ^2. However, such an informative prior distribution could be problematic when making Bayesian inference for τ^2 when not much is known beyond the data included in the analysis at hand. A uniform prior density on τ with an upper limit larger than the standard deviation of the observed effects, e.g. $U(0, 10)$, is recommended. [66] Under the DP normal model, $Var(\theta_i) = \tau^2 \frac{\alpha}{\alpha+1}$ and it converges to τ^2 when α tends to infinity. Therefore, for reasonably large α , τ^2 can be still interpreted as a heterogeneity parameter in a typical random effects model.

Define the vector of distinct means $\boldsymbol{y} = (y_1, \ldots, y_k)$ and vector of sample variances as $\boldsymbol{s^2} = (s_1^2, \ldots, s_k^2)$. First, we assume the within-study variances are equal and use the normal-theory sampling distribution of the sample variance to capture uncertainty about the within-study variances $p(s_i^2 | \sigma^2)$:

$$(n_i - 1) s_i^2 \sim gamma\left(\frac{n_i - 1}{2}, \frac{1}{2\sigma^2}\right). \tag{3.5}$$

Then we can assume a wide uniform prior density (e.g. $U(0, 10)$) on σ. The conditional

posterior distribution is

$$p(\theta_i|\boldsymbol{y}, \boldsymbol{s^2}, \boldsymbol{\theta_{-i}}, \sigma^2, \tau^2) \propto \sum_{j \neq i} \phi(y_i|\theta_j, \frac{\sigma^2}{n_i}) \cdot \delta_{\theta_j} + \left\{ \alpha \int \phi(y_i|\theta_i, \frac{\sigma^2}{n_i})) \phi(\theta_i|\mu, \tau^2) d\theta_i \right\} \times \phi(\theta_i|\mu, \tau^2) \phi(y_i|\theta_i, \frac{\sigma^2}{n_i})),$$

where $\phi(\cdot)$ is the density function of the corresponding distribution. Then, we can get the conditional posterior distribution of other parameters $p(\sigma^2|\boldsymbol{y}, \boldsymbol{s^2}, \boldsymbol{\theta}, \tau^2)$, $p(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}, \boldsymbol{s^2})$ and use a Gibbs sampler for the full posterior. [48] The posterior predictive distribution of y_{new} given $\{y_i, \ldots, y_k\}$ is:

$$f(y_{new}|\{y_i, \ldots, y_k\}) = \iint f(y_{new}|\boldsymbol{\theta}, \sigma_i^2) f(\boldsymbol{\theta}, \sigma^2|\{y_i, \ldots, y_k\}) d\boldsymbol{\theta} d\sigma^2.$$

The limits of the α-level reference interval can then be estimated by the $\frac{\alpha}{2} \times 100$ and $(1 - \frac{\alpha}{2}) \times 100$ percentiles of y_{new}'s predictive distribution; note here y_{new} is the predication for a new individual not for a new study.

The DP avoids assuming that all individuals can be clustered into a fixed number of groups, N, but assumes that there are infinitely many clusters represented in the overall population, with an unknown number observed in a finite sample of k studies. One common concern with the blocked Gibbs sampler used for the DP is that this approach relies on the truncation of the stick-breaking representation to a finite number of terms, thus passing the infinite-dimensional representations and fitting a model with finite clusters. For example, if we set $N = 30$, which is the number of studies included in the meta-analysis, as the truncation level, a natural question is how this is better or intrinsically different than fitting a model with 30 components. However, the truncation level for DP is not the number of components occupied by the subjects in the sample, but merely an upper bound on the number of components. In most cases, taking a conservative upper bound as the number of included studies, which is used for our nonparametric method, should be sufficient since clustering models are most useful when there are relatively few components.

3.2.2 Unequal within-study variances

In a meta-analysis, the source of variation is usually of great scientific interest, and it is important to account for within-study variances, which we call study-specific sampling errors and between-study variances. [55] Within-study variances can be conceptualized as quantifying the variation that would arise if a specific study were replicated multiple times, each time with a different sample of replicates but with the same study effect. In many random-effects meta-analyses, sample variances from each study are utilized as the within-study variances, often assuming a uniform underlying σ^2 for simplicity. First, assuming equal within-study variances overlooks the uncertainty associated with these variances within the model. Moreover, it may lead to a mischaracterization of the differences across studies in the "true" within-study variances, thus misspecifying the total heterogeneity across studies and affecting prediction accuracy. Similar to the approach employed in previously proposed frequentist and Bayesian parametric methods for reference interval estimation [2], our nonparametric method can assume unequal within-study variances to account for the uncertainty and heterogeneity in within-study variances. Considering that the within-study variances σ_i^2 may vary across studies and it is difficult to determine the true underlying parametric distribution, it is intuitive to model them nonparametrically, similar to the study means which are already assumed to follow a DP. We can use a DP prior on the unequal σ_i instead of the uniform prior under the equal within-study variances assumption in Section 2.1:

$$\begin{aligned} y_i \sim f\left(\theta_i, \frac{\sigma_i^2}{n_i}\right), \theta_i \sim G, G \sim DP(\alpha, G_0) \\ (n_i - 1)s_i^2 \sim gamma(\frac{n_i - 1}{2}, \frac{1}{2\sigma_i^2}),\ \sigma_i \sim G_\sigma,\ G_\sigma \sim DP(\alpha_\sigma, G_{0\sigma}). \end{aligned} \tag{3.6}$$

The baseline $G_{0\sigma}$ is an inverse gamma distribution $IG(a_0, b_0)$ with shape parameter a_0 and scale parameter b_0, which can achieve conditional conjugacy for the normal distribution $N(\theta_i, \sigma_i^2/n_i)$ for each study. [57] The two DPs for random effects θ_i and within-study variances σ_i^2 are assumed to be independent. The posterior predictive distribution of y_{new} given $\{y_i, \ldots, y_n\}$ can be obtained by using a similar Gibbs sampler to that described in Section 2.1 after adding this new DP for σ_i.

3.3 Simulation

The aim of the simulation study is to investigate the performance of the nonparametric method for estimating reference intervals under non-normal conditions. Since the common assumption of normally distributed study means will often be violated in real data sets, we simulate data under alternative non-normal distributions for study means θ_i: a mixture of normal distributions, log-normal, and gamma distributions. The total number of studies k was set to be 5, 10 or 30 and the sample size for each study is 100. We conducted 2000 simulations for each condition. Each simulation compared the nonparametric method with the frequentist method, Bayesian parametric method, and empirical methods proposed by Siegel et al. [2] and the mixture distribution method proposed by Cao et al. [43] For our Bayesian nonparametric method, we used the R package Nimble; [67] the code can be found in the Supplementary Materials. For the Bayesian parametric method, [2] we used JAGS version 4.3.0 with the packages rjags and coda in R version 4.2.1. [68–70].

For the scenarios where the study means followed a mixture of normal distributions, the density of this mixture was formulated as:

$$\begin{aligned} f(\theta; \mu_1, \ldots, \mu_k, \tau_1, \ldots, \tau_k) &= \sum_{i=1}^{k} p_i f_i(\theta; \mu_i, \tau_i), \\ f_i(\theta; \mu_i, \tau_i) &= \frac{1}{\sqrt{2\pi\tau_i^2}} exp\left(\frac{(\theta - \mu_i)^2}{2\pi\tau_i^2}\right), \quad \sum_{i=1}^{k} p_k = 1, \end{aligned} \tag{3.7}$$

We also generated θ_i from a log-normal distribution and gamma distribution. The measurements within each study are generated from a normal $y_{ij} \sim N(\theta_i, \sigma_i^2)$. For the mixture of normal distributions, the means and variances are $\mu = (8, 10, 11)$, $\tau^2 = (1.5^2, 0.8^2, 0.5^2)$ and we set the mixing proportion to be $p = (0.4, 0.4, 0.2)$. We choose this mixture p to generate a skewed distribution as opposed to the symmetric normal distribution assumed by existing methods. The logarithm of mean and standard deviation (SD) of the log-normal distribution is 8 and 3.5, the mean of the gamma distribution is 5 and the SD is 3.5. We considered both equal and unequal within-study variances for all three data-generation scenarios. The summary statistics i.e. means and SDs, for each study were used to fit each of the six models, frequentist, Bayesian parametric, empirical, mixture (fixed effects), the

nonparametric model using one DP for study means (NP), and the nonparametric model using two DPs for study means and within-study variances (NP-2).

We used the following prior distributions in the Bayesian framework for the NP method:

$$\sigma \sim U(0,10), \quad \alpha \sim gamma(1,1), \quad \tau \sim U(0,10), \quad \mu \sim N(0,100). \tag{3.8}$$

For the NP-2 method, the prior of σ was set to be:

$$\sigma_i \sim IG(\alpha_0, \beta_0), \alpha_0 \sim IG(1,2), \beta_0 \sim IG(1,2). \tag{3.9}$$

For the [2] Bayesian parametric method the prior distributions for μ, σ, τ were the same as Equation (8).

3.3.1 Equal Within-Study Variances

For the equal variance scenarios, after generating the true study means θ_i according to the specified non-normal distribution, we then generated the individual-level data according to a normal distribution $N(\theta_i, \sigma^2)$, where $\sigma^2 = 1.25$ was constant across studies. We also compared the results when using a half-Cauchy prior distribution $p(\tau) \propto (1+\tau^2)^{-1}$ for the variance parameter τ under the mixture of normal distributions condition in the case where $k = 5$.

3.3.2 Unequal Within-Study Variances

In addition, we conducted a set of simulations where the within-study variances were unequal across different studies. The σ_i were generated from a truncated normal distribution with a mean of 1, the left truncation point equal to 0.5, and the right truncation point equal to 2.5. θ_i were generated from the mixture of normal distributions, $\mu = (8, 10, 11)$, $\tau^2 = (1.5^2, 0.8^2, 0.5^2)$, and $p = (0.4, 0.4, 0.2)$. Here μ_i and σ_i are generated independently.

3.3.3 Outliers

To show the robustness of our proposed methods to the presence of outliers, we conducted additional scenarios where we introduced outliers to each simulated study. After generating

the individual level data for each study, using the 25% (Q1) and 75% (Q3) quantiles to get the interquartile range (IQR), which is defined as the difference between Q1 and Q3. Outliers were defined as values smaller than $Q1 - 1.5 \times IQR$ or larger than $Q3 + 1.5 \times IQR$ [71]. Based on this definition, original data generated from the mixture normal distribution with equal within-study variances in Section 3.1 contained approximately 0.8% "outliers". We took random sampling with replacement from the outliers and added them into the original data, which made the overall proportion of outliers close to 2.5%.

3.3.4 Simulation Results

For each scenario, we calculated the means of the estimated 95% reference interval, the proportion of the true population distribution captured by each of the reference interval methods, which we call the "coverage", and the median, 2.5th percentile and 97.5th percentiles of the coverage (Tables 2 and 3). We use this coverage to evaluate the capability of each method of including the pre-specified proportion (e.g., 95%) of measurements from the healthy population; this is a metric of interest when evaluating a prediction interval. This is the same as the coverage that has been used in previous papers for estimating the reference range [2, 43],. We also drew the mean of the lower and upper bounds of the 2000 estimated reference intervals (Figures 1-5 of the Supplementary Materials) for each scenario. The solid lines are the true 2.5th and 97.5th percentiles of the marginal distribution, and the horizontal lines are the mean of the estimated interval limits.

Under the equal within-study variances setting, two nonparametric methods gave similar results. The two nonparametric methods and the mixture distribution method had better performance in capturing 95% of the marginal distribution than the other three random effects model-based methods when the true underlying distribution was a mixture of normal distributions. Specifically, the mean of the estimated 95% reference interval limits from the two nonparametric models and mixture distribution methods were closer to the true 2.5th and 97.5th percentiles (Table 2). Table 2 and Supplemental Figures S1-S3 also show that when the study means were generated according to a log-normal or a gamma distribution and the number of studies is large (30), the performance of two nonparametric models and mixture distribution methods have obvious advantages in comparison to the other methods. Firstly, our nonparametric methods exhibited less bias in estimating

the reference interval limits compared to other methods. Additionally, the other methods produced lower bounds for the gamma distribution that were negative, an improbable outcome for a non-negative distribution. Reducing the number of included studies had a detrimental effect on both coverage and the accuracy of reference interval estimates for all methods except the Bayesian parametric method. Surprisingly, the Bayesian parametric method estimated an exceptionally wide and biased interval, resulting in high coverage but potentially misclassifying observations as coming from healthy individuals. For the case where $k = 5$, we employed the half-Cauchy prior for the between-study standard deviation parameter τ. The estimated means of the reference interval limits were $[5.96, 12.54]$, with a median coverage of $0.89(0.71, 0.98)$. These results closely resembled those obtained using the uniform prior, which produced an estimate of $[5.97, 12.63]$ and a coverage of $0.89(0.71, 0.98)$

Under conditions of unequal within-study variances, as depicted in Table 3, our two nonparametric models and mixture distribution methods continued to outperform other approaches in accurately capturing 95% of the marginal distribution, especially when the study number (k) was set to 30. In this scenario, which provided more precise estimates for the 95% reference interval, our methods excelled. However, as the study number (k) decreased, the performance of the nonparametric methods declined, both in terms of coverage and the accuracy of the 95% reference interval. This decrease can be attributed to the scarcity of information available for estimating the mixture components when utilizing two DPs. It's noteworthy that the results for the NP and NP-2 methods exhibited similar trends under both equal and unequal within-study variances settings.

In Table 4 and Supplementary Figure S4, we investigate the impact of introducing outliers into the dataset. Our nonparametric methods NP and NP-2 both displayed robustness even in the presence of additional outliers. In contrast, other methods tended to exaggerate the reference intervals. Moreover, the incorporation of more studies resulted in enhanced estimates for the Bayesian parametric method. However, despite this augmentation, the Bayesian parametric method still yielded a wide reference interval.

3.4 Real Data Analysis

3.4.1 A meta-analysis of human postural vertical measurements

The first case study is a meta-analysis of human subjective postural vertical (SPV) measurements [15] which reflect an individual's ability to perceive whether they are oriented vertically or not. Vertical perception is an important ability associated with postural control. [30] To measure the SPV, participants sit on a tilting chair with their eyes closed and verbally instruct an examiner to set the chair to their perceived upright body orientation.

We used data from 15 studies that measured frontal SPV, or the deviation (in degrees) of the specified position from true verticality in the frontal plane. We evaluated the normality of the study means using the D'Agostino test (p-value=0.41) and the normal Q-Q plot; there was no apparent departure from normality. Thus, this case study was used to compare the different reference interval methods under a scenario where the between-study normality assumption was likely valid. [2] used these data as a case study when proposing the frequentist method, Bayesian parametric method, and empirical method, and estimated the reference intervals as $[-2.92°, 3.15°]$, $[-3.07°, 3.20°]$ and $[-2.89°, 3.13°]$, respectively. The fixed effects method estimated the reference interval as [-2.97, 3.10]. The DP method assuming equal within-study variances resulted in a smaller upper limit $[-2.97°, 3.24°]$ than the unequal within-study variances method $[-2.96°, 3.40°]$. As we expected, the 95% CI for the pooled mean was narrow ($[-0.04°, 0.27°]$) and did not reflect the variation between individuals. The reference interval calculated using the empirical method was $[-2.87°, 3.11°]$; Previous study [15] used a similar empirical method but their interval was slightly different since they used 2 times standard deviation instead of 1.96 and weighted by n when estimating the overall variance. Figure 1 shows that the reference intervals estimated using our proposed methods overlapped with all individual studies' 95% CIs for the mean, and that the 95% CI for the pooled mean and the overall 95% prediction CI for a new study do not reflect the full individual-level variation in measurements. The Watanabe–Akaike information criterion (WAIC) and the effective number of parameters (PAIC) [72] for the Bayesian parametric method was 250 (PAIC=19); for our NP equal-variance and NP-2 unequal-variance methods, these were 272 (PAIC=26) and 247 (PAIC=66).

3.4.2 A meta-analysis of Pediatric nighttime sleep

Our second case study uses data from a systematic review and subsequent meta-analysis of pediatric sleep measures. [13] In particular, we focus on the measure of wake time after sleep onset (WASO). Of the 79 studies included in the systematic review, 23 studies reported WASO. There was no apparent departure from the normality of the study means based on the D'Agostino test (p-value=0.93). However, in this case study, the wake time can not be negative and the lower bound of the estimated reference intervals must be truncated at 0. Thus, we considered log-transforming the included data using the method proposed by Siegel et al. [2] to make the within-study normality assumption more reasonable. After this transformation, the study means became significantly skewed based on the D'Agostino test (p-value=0.04) and the weighted test (p-value=0.02) [59], as well as visual inspection. Therefore, this provides an example of using the DP method under a likely non-normal between-study distribution. The estimated 95% reference intervals are (0.07, 4.86), (0.06, 6.02), (0.01, 4.79), (0.06, 2.59), (0.06, 2.84), and (0.06, 2.88) hours for the frequentist, Bayesian, empirical, mixture distribution, NP equal within-study variance, and NP unequal within-study variances methods, respectively. As shown in Figure 2, the frequentist, Bayesian, and empirical methods gave extremely wide reference intervals that would cover all healthy and unhealthy individuals; this is likely due to a violation of their normality assumptions. The mixture distribution and DP methods (equal and unequal within-study variances) gave similar reference intervals that overlapped with all the included study means. The WAIC (PAIC) for the Bayesian parametric method was 428 (PAIC=30), and for our NP equal-variance and NP-2 unequal-variance methods were 442 (PAIC=31) and 182 (PAIC=77). This suggested that the proposed nonparametric methods provide a good alternative to the previously proposed random effects methods when the between-study normality assumption does not hold.

3.5 Discussion

The random effects model is a commonly used meta-analysis model to combine multiple independent studies on a particular question and draw inferences about the general population. In this paper, we used a Bayesian nonparametric method to relax the between-study

normality assumption for the underlying study means and the equal within-study variances assumption imposed by previous random effects methods for estimating the reference interval from a meta-analysis. The reference interval discussed in this paper is conceived as a prediction interval for new, healthy individuals within the population, delineating a region in the sample space defined by quantiles. Some authors advocate for the use of tolerance intervals in medical diagnostics. [73, 74] A tolerance interval aims to ensure that the coverage of a reference interval is at least 95%, effectively controlling the false positive rate for identifying healthy individuals as abnormal to less than 5%. However, it's important to note that tolerance intervals do not safeguard against selecting overly wide intervals. In contrast, diagnostic procedures relying on tolerance intervals may bolster specificity but often at the cost of sacrificing sensitivity. [75] Therefore, in this paper, we have chosen to define reference intervals as prediction intervals rather than tolerance intervals.

This Bayesian nonparametric method has similar properties with the frequentist fixed effects mixture distribution method [43] since they both assume the underlying distribution is a mixture. However, the Bayesian nonparametric method follows a fully specified Bayesian hierarchical random effects model framework, which assumes a flexible stochastic relationship between the study parameters, thus allowing for predictive inference on future study means and individual measurements without assuming that the studies included in the meta-analysis represent the entire target population. The fixed effects mixture distribution method in general cannot be used to predict future study means as the observed study means are assumed to be fixed. The proposed Bayesian nonparametric method can be regarded as a complement to the existing methods recently proposed. [2, 43] The Bayesian nonparametric method provides a flexible alternative when the study effects in the random effects model cannot be assumed to follow a normal distribution. The simulation study in Section 3 shows the advantage of using the Bayesian nonparametric method for log-normal and gamma distributions. An appropriate choice of aforementioned methods should be based on the specific scientific question and the collected data.

Our proposed Bayesian nonparametric method uses a truncated DP assuming the study means and within-study variances follow a DP with an unknown number of clusters. The upper bound on the number of clusters is simply the number of included studies; this is a flexible assumption that lets the data determine the posterior predictive interval and

interpretation. When there are many studies, assuming such a large number of components might increase the model complexity and result in overfitting in the DP implementation.

In this paper, we also allowed the within-study variances to be heterogeneous across studies in the simulations and used another independent DP for the σ_i^2. This flexible assumption does not require further information about the study means and variances and can also estimate the reference intervals with 95% coverage in the simulation. A large number of studies is recommended to provide enough information for the nonparametric methods with two DPs to estimate the mixture components. While the independent DPs assumption is simple to implement, it may lack interpretability. [76] This is because θ_i, θ_j belonging to same component implies that study i and j are similar, thus suggesting that σ_i^2, σ_j^2 may also belong to same component. However, dependent DPs would increase the model complexity and may require more data to accurately estimate the dependence parameter. The simulation study in Section 3 also demonstrates that a model using independent DPs can still estimate the reference interval well.In our simulations, we generated data assuming equal and unequal within-study variances and compared two Bayesian nonparametric models that assumed equal and unequal within-study variances under each scenario. The estimated intervals and the coverage of the true marginal distributions are similar for two Bayesian nonparametric models under each scenario (See Table 2 and Table 3). We hypothesize this is because we are primarily interested in the total variance (sum of between-study variance and within-study variance) of the individual measurements across studies, not the variance within each specific study. In the case study of pediatric nighttime sleep, assuming equal vs. unequal within-study variances only affects the results for the reference interval slightly, but there is a large difference in the WAIC used for model selection. This is because, from the model selection perspective, the full model for the joint distribution is evaluated while the reference interval only considers the marginal distribution. It is possible for the equal variances model to not fit individual study variances well but average out to approximately the correct marginal variance, thus giving similar reference interval results but very different results for the WAIC. It becomes apparent that these two nonparametric models may yield different goodness of fit to the data.

Here we extend the DP model [51] to obtain reference intervals for non-normal study effects. We opt to not constrain the median [77] as this constraint is primarily for the

estimation of one overall median study effect, and the focus here is on flexible methods for estimating a reference interval for study effects. In practice, assuming a discrete mixture is likely an oversimplification. However, we believe this approach can still approximate the likely continuous distribution of study means. A continuous Polya tree prior can be used to the mean distribution, [60] obviating the clustering in the discrete DP methods; this approach considers a flexible, continuous discrete mixture for the study means, disallowing clustering (i.e. two or more studies with the same exact mean) and offering more flexibility in the mean distribution that a Gaussian-centered Polya tree. However, their approach centers the mean distribution at a Gaussian distribution, and this centering markedly affects inference in meta-analyses with a smaller number of studies. A model that allows the random effects distribution to change flexibly and non-linearly (not a linear regression), such as a generalized linear model for some study-level covariates can be further considered. [78] This feature permits a flexible meta-regression analysis that can account for covariate information. Future work may include implementing the two approaches for estimating the reference interval.

The proposed method uses aggregated data from published papers; however, individual participant data (IPD) can provide more information for estimating the reference interval if they are available. It has been shown how to estimate the reference interval using previously proposed methods in one step by using IPD; [42] the reference interval based on IPD can be considered a "gold standard". Our future work will include further developing nonparametric meta-analytic methods with IPD.

Table 3.1: Methods for Estimating the Reference Interval

Method	Assumptions	Limitations
Frequentist	1) Random effects model.	
	2) Normal distribution for study means.	Few studies.
	3) Normal distribution for within-study measurements.	Skewed data, outliers.
	4) Constant within-study variance.	Unequal variances.
Bayesian	1), 2), 3), and 4) are the same as the frequentist method.	
	5) Accounts for the uncertainty in variance parameters.	Wider estimated reference intervals.
Empirical	1) Applied with the assumption for the overall distribution.	Cannot predict a new study.
	2) Measurements across all studies follow any given. distribution	
Mixture distribution	1) Fixed effects model: study means are unrelated.	Cannot predict a new study.
	2) Measurements in each study. follow any given distribution.	
	3) Overall population is the mixture of each study distribution.	Assumption 3) is too strong.
Nonparametric	1) Random effects model.	
	2) Dirichlet process for study means.	Model complexity.
	3) Dirichlet process for within-study variances.	

Table 3.2: Data generated with equal within-study variances $\sigma^2 = 1.25$ for all three scenarios. We considered using two NP models: NP used DP for the study means, and the second NP-2 used two DPs for both study means and within-study variances

Method	Mixed Normal		LogNormal		Gamma	
	95% RI[1]	Coverage[2]	95% RI	Coverage	95% RI	Coverage
True Range	[5.17, 12.79]		[2.60, 16.90]		[0.07, 14.07]	
$k = 5$						
NP[3]	[5.97, 12.63]	0.89 (0.71, 0.98)	[3.33, 13.64]	0.82 (0.57, 0.97)	[0.40, 10.61]	0.81 (0.55, 0.99)
NP-2	[5.95, 12.55]	0.89 (0.71, 0.97)	[3.36, 13.66]	0.82 (0.55, 0.97)	[0.41, 10.65]	0.82 (0.54, 0.99)
Mix	[5.88, 12.59]	0.90 (0.72, 0.98)	[3.30, 13.70]	0.83 (0.56, 0.97)	[0.36, 10.73]	0.83 (0.56,0.99)
Freq	[5.65, 13.15]	0.93 (0.73, 0.99)	[1.46, 14.57]	0.91 (0.62, 1.00)	[-1.54, 11.54]	0.91 (0.59, 0.99)
Emp	[5.89, 12.91]	0.91 (0.72, 0.99)	[2.05, 13.98]	0.88 (0.59, 1.00)	[-0.95, 10.95]	0.88 (0.57, 0.99)
Bayes	[2.93, 15.81]	0.99 (0.81, 1.00)	[-4.56, 20.41]	0.99 (0.82, 1.00)	[-7.48, 17.41]	0.99 (0.77, 1.00)
$k = 10$						
NP	[5.60, 12.67]	0.92 (0.80, 0.98)	[2.99, 14.92]	0.89 (0.71, 0.98)	[0.17, 12.44]	0.90 (0.70, 0.99)
NP-2	[5.60, 12.69]	0.92 (0.80, 0.98)	[2.99, 14.91]	0.89 (0.71, 0.98)	[0.10, 12.28]	0.90 (0.72, 0.99)
Mix	[5.50, 12.70]	0.93 (0.81, 0.98)	[2.93, 15.06]	0.89 (0.71, 0.98)	[0.07, 12.41]	0.91 (0.72,0.99)
Freq	[5.58, 13.21]	0.94 (0.82, 0.99)	[1.20, 14.70]	0.93 (0.75, 0.99)	[-1.88, 11.91]	0.94 (0.75, 0.99)
Emp	[5.70, 13.10]	0.93 (0.81, 0.99)	[1.51, 14.42]	0.92 (0.73, 0.99)	[-1.57, 11.60]	0.93 (0.73, 0.99)
Bayes	[4.78, 13.98]	0.98 (0.86, 1.00)	[-0.67, 16.48]	0.96 (0.83, 1.00)	[-3.78, 13.72]	0.97 (0.81, 0.99)
$k = 30$						
NP	[5.31, 12.76]	0.94 (0.89, 0.97)	[2.80, 16.38]	0.93 (0.84, 0.98)	[0.004, 13.95]	0.95 (0.85, 0.99)
NP-2	[5.30, 12.77]	0.94 (0.89, 0.97)	[2.75, 16.72]	0.93 (0.85, 0.98)	[0.02, 14.10]	0.95 (0.86, 0.99)
Mix	[5.20, 12.80]	0.95 (0.90, 0.97)	[2.68, 17.21]	0.94 (0.85, 0.98)	[-0.05, 14.45]	0.95 (0.87,0.99)
Freq	[5.52, 13.26]	0.95 (0.89, 0.98)	[0.94, 15.05]	0.95 (0.87, 0.99)	[-2.11, 12.13]	0.95 (0.87, 0.99)
Emp	[5.56, 13.22]	0.95 (0.89, 0.98)	[1.04, 14.95]	0.95 (0.86, 0.99)	[-2.01, 12.03]	0.95 (0.88, 0.99)
Bayes	[5.32, 13.48]	0.96 (0.90, 0.99)	[0.44, 15.61]	0.96 (0.88, 0.99)	[-2.61, 12.69]	0.96 (0.88, 0.99)

[1] Mean of 2000 estimates of the 95% reference interval.

[2] Mean, 2.5th and 97.5th percentile of the proportion of the true population captured by the estimated 95% reference interval.

[3] NP: nonparametric model using one DP for study means; NP-2: the nonparametric model using two DPs for study means and within-study variances; Mix: mixture distribution method; Freq: frequentist method; Emp: empirical method; Bayes: Bayesian parametric method.

Table 3.3: Data generated with unequal within-study variances, where σ_i were generated from a truncated normal distribution with a mean of 1, the left truncation point equal to 0.5, and the right truncation point equal to 2.5. We considered using two NP models, one used DP for the study means, and the second NP-2 used two DPs for both study means and within-study variances

	k=5		k=10		k=30	
	95% RI[1]	Coverage[2]	95% RI	Coverage	95% RI	Coverage
True Range	[4.82, 13.36]		[4.82, 13.36]		[4.82, 13.36]	
NP[3]	[5.47, 13.15]	0.91 (0.74, 0.98)	[5.20, 13.22]	0.93 (0.83, 0.98)	[4.95, 13.29]	0.94 (0.90, 0.97)
NP-2	[5.51, 13.09]	0.91 (0.74, 0.97)	[5.18, 13.20]	0.93 (0.83, 0.97)	[4.93, 13.27]	0.94 (0.90, 0.97)
Mix	[5.35, 13.14]	0.91 (0.75, 0.98)	[5.07, 13.29]	0.93 (0.82, 0.98)	[4.85, 13.32]	0.95 (0.90, 0.97)
Freq	[5.16, 13.59]	0.93 (0.77, 0.99)	[5.15, 13.67]	0.94 (0.82, 0.99)	[5.12, 13.67]	0.95 (0.90, 0.99)
Emp	[5.39,13.36]	0.92 (0.75, 0.99)	[5.26, 13.55]	0.94 (0.82, 0.98)	[5.16, 13.63]	0.95 (0.90, 0.99)
Bayes	[2.51,16.36]	0.99 (0.86, 1.00)	[4.40, 14.40]	0.97 (0.86, 1.00)	[4.95, 13.86]	0.96 (0.91, 0.98)

[1] Mean of 2000 estimates of the 95% reference interval.

[2] Mean, 2.5th and 97.5th percentile of the proportion of the true population captured by the estimated 95% reference interval.

[3] NP: nonparametric model using one DP for study means; NP-2: nonparametric model using two DPs for study means and within-study variances; Mix: mixture distribution method; Freq: frequentist method; Emp: empirical method; Bayes: Bayesian parametric method.

Table 3.4: Simulation Results for Mixed Normal adding outliers: outliers defined as values smaller than $Q1 - 1.5 \times IQR$ or larger than $Q3 + 1.5 \times IQR$. The overall proportion of outliers was approximately 2.5%.

	k=5		k=10		k=30	
	95% RI[1]	Coverage[2]	95% RI	Coverage	95% RI	Coverage
True Range	[5.17, 12.79]		[5.17, 12.79]		[5.17, 12.79]	
NP [3]	[5.77,12.72]	0.91 (0.74, 0.98)	[5.47,12.83]	0.93 (0.83, 0.98)	[5.20,12.90]	0.95 (0.90, 0.98)
NP-2	[5.76,12.73]	0.91 (0.74, 0.98)	[5.48,12.86]	0.94 (0.83, 0.98)	[5.17,12.92]	0.95 (0.90, 0.98)
Mix	[5.70, 12.80]	0.91 (0.75, 0.98)	[5.39,12.87]	0.94 (0.83, 0.98)	[5.13,12.93]	0.95 (0.91, 0.98)
Freq	[5.47, 13.29]	0.92 (0.75, 0.99)	[5.45,13.33]	0.95 (0.83, 0.99)	[5.41,13.35]	0.96 (0.90, 0.98)
Emp	[5.70, 13.10]	0.95 (0.89, 0.98)	[5.57,12.21]	0.94 (0.83, 0.99)	[5.44,13.32]	0.95 (0.90, 0.98)
Bayes	[2.77, 15.91]	0.99 (0.95, 1)	[4.86,14.07]	0.94 (0.83, 0.98)	[5.22,13.56]	0.97 (0.91, 0.99)

[1] Mean of 2000 estimates of the 95% reference interval.

[2] Mean, 2.5th and 97.5th percentile of the proportion of the true population captured by the estimated 95% reference interval.

[3] NP: nonparametric model using one DP for study means; NP-2: nonparametric model using two DPs for study means and within-study variances; Mix: mixture distribution method; Freq: frequentist method; Emp: empirical method; Bayes: Bayesian parametric method.

Figure 3.1: **A Meta-analysis of Sagittal Plane SPV:** Mean (95% CI) and 95% prediction interval for a new individual for each study; Overall is the 95% CI for pooled mean estimated by the fixed effects model; 95% reference ranges are estimated from different methods under the normal distribution.

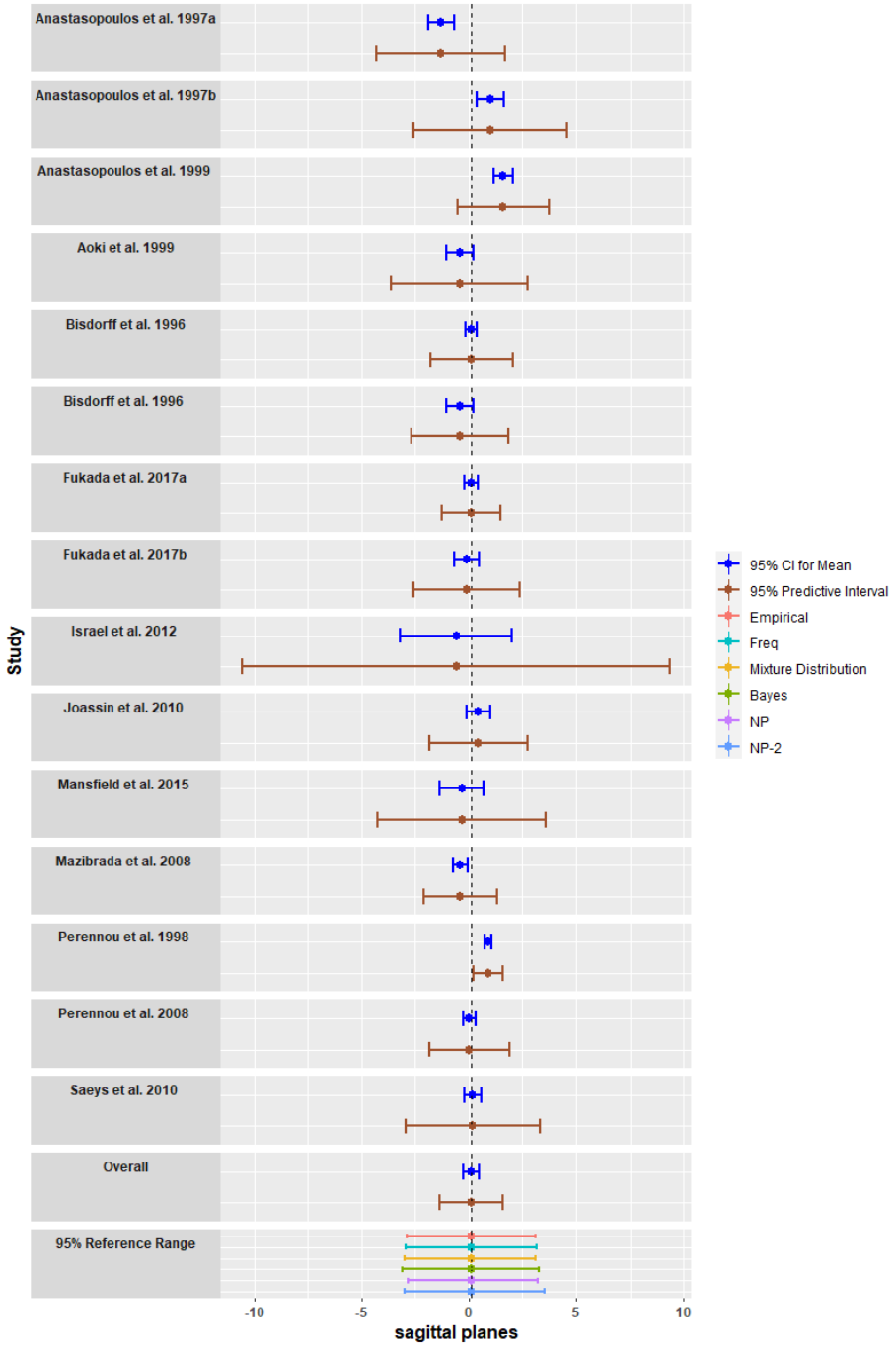
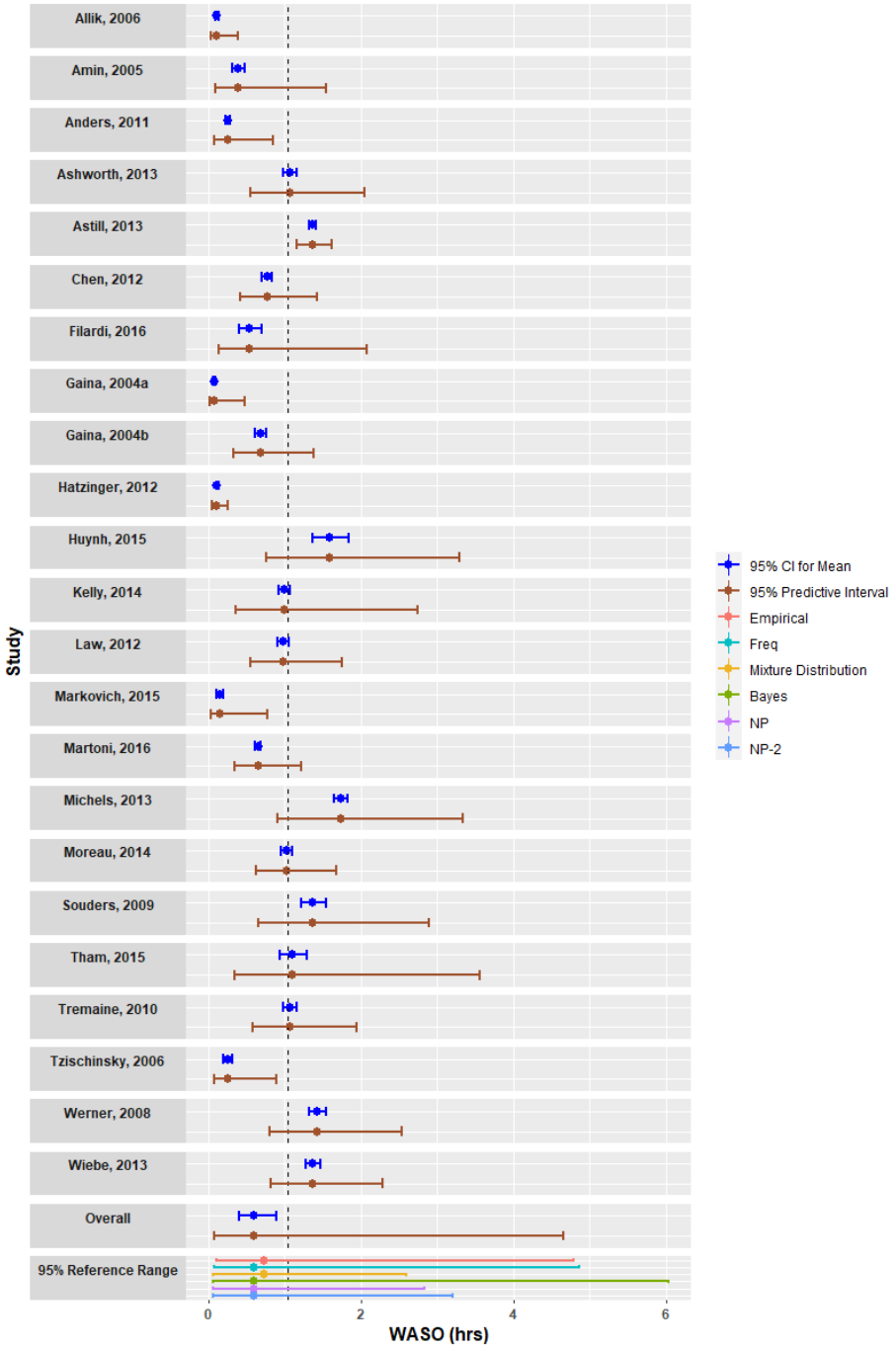


Figure 3.2: **A Meta-analysis of wake time after sleep onset:** Mean (95% CI) and 95% prediction interval for a new individual for each study; Overall is the 95% CI for a pooled mean estimated by the fixed effects model; 95% reference ranges are estimated from different methods under the lognormal distribution.



Chapter 4

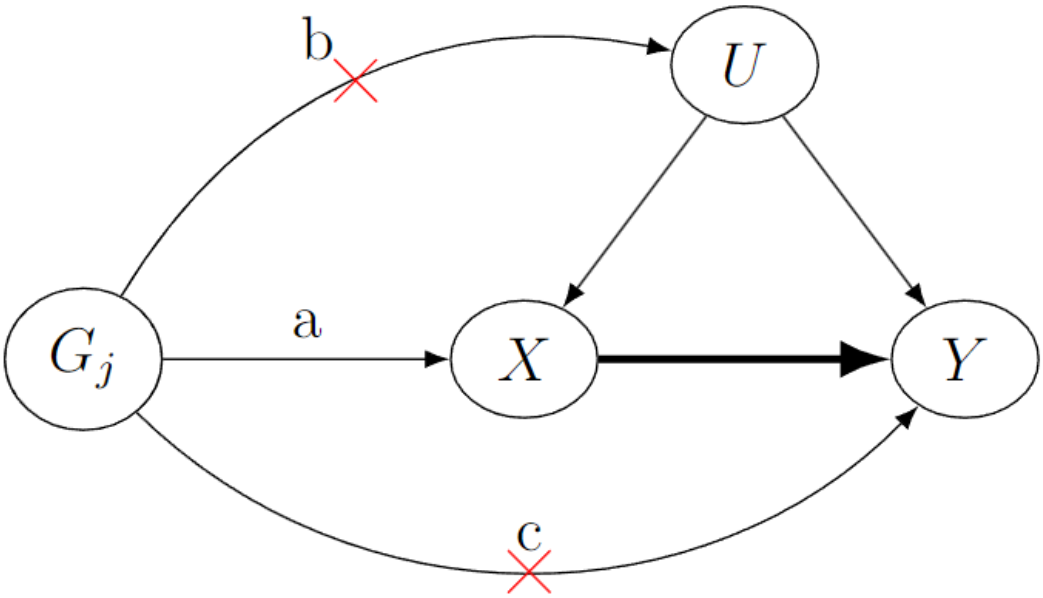
A random effect model-based method of moments estimation of causal effect in Mendelian randomization studies

4.1 Introduction

Inferring the causal direction between correlated variables is a pervasive issue in biology that cannot be assessed through simple association testing or regression analysis. Mendelian randomization (MR) is a powerful tool for estimating the causal effect of an exposure variable X on the outcome Y by utilizing genetic variants as instrumental variables G for exposure [79–81]. To date, MR has been successfully applied to a wide range of observational associations to assess the causal effects of biomarkers on disease, and to understand the causal basis for correlations between physiological measures and various behavioral traits and mental health disorders [82–86]. With the availability of an increasing number of well-powered genome-wide association studies (GWASs) on a growing number of traits, there has been tremendous interest in using genetic variants as IVs. The validity of MR depends on three key assumptions as shown in Fig 4.1: (a) Relevance: the IVs are associated with the exposure factor; (b) Independence: there are no unmeasured

confounders of the associations between IVs and outcome; and (c) Exclusion restriction: the IVs affect outcome only through their effect on the exposure factor. MR can be implemented using either individual-level data or summary statistics, employing methods such as the two-stage least squares (2SLS) method [87] or the ratio of coefficients method [88]. These methods can be applied using a single IV and can also be extended to accommodate multiple IVs [89, 90]. However, there are many ways in which the key assumptions may be violated, which have been recently discussed [91–93].

Figure 4.1: A causal model illustrating the three assumptions on a valid IV



With the proliferation of GWA studies, there is a growing tendency to employ a large number of genetic variants as IVs in such investigations [94]. Nevertheless, larger sets of genetic variants are more prone to contain invalid IVs due to horizontal pleiotropy [95], wherein certain IVs may be associated with the outcome variable, thereby violating the exclusion assumption. Another challenge in utilizing genetic variants as instruments is their modest association with the exposure variable of interest, which restricts the power to test

causal hypotheses and the precision of causal effects. In many MR analyses, a commonly employed approach involves incorporating multiple instruments, collectively accounting for a greater portion of the variance in the exposure variable, even if the effect size of each individual variant is weak [92, 96]. However, the inclusion of a weak IV can introduce bias in MR studies, even when the instrument satisfies the three assumptions and a large sample size is employed [97]. When the effect of the IV on the exposure variable, X, is weak, the instrument explains only a small amount of the variation in the exposure variable, while confounders may account for a larger portion of the difference in the exposure variable compared to the instrument. The bias introduced by the weak IVs aligns with the confounded observational association between the exposure factor and outcome [98], resulting in estimates with wide confidence intervals. An instrument weakly associated with the exposure variable yields a small denominator in the ratio estimator that is commonly used for MR [91], thereby amplifying bias caused by any minor violations of the independence assumption and exclusion restriction [99].

Several methods have been developed have developed methods, such as MR-Egger [93], Simple Median [100], Weighted Median [100], MR-Lasso [101], that address the challenge of invalid IVs with exclusion restriction violation. These methods typically involve the preselection of IVs with strong effects to eliminate weak IVs. However, this selection process relies on prior knowledge, and using measured F-statistics to select significant IVs can potentially exacerbate bias [102]. We call those methods pleiotropy-IV MR methods. To handle the issue of weak IVs under the assumption of no systematic exclusion restriction violation, alternative methods such as the limited information maximum likelihood (LIML) estimator, or the continuously updating estimator (CUE), have been proposed [92, 103]. These methods, which we call weak-IV MR methods, aim to mitigate the impact of weak IVs on causal inference. In the presence of both weak IVs and horizontal pleiotropy, some authors have proposed pleiotropy-weak-IV robust methods that allow for the inclusion of pleiotropic IVs and multiple weak IVs, such as debiased IVW (dIVW) [104] and genius MR for many weak invalid instruments (GENIUS-MAWII [105]). However, these proposed MR techniques for multiple IVs still require a large sample size to ensure robustness. With increasing representation of global populations in a GWAS generally with small sample sizes, there is a strong need to develop approaches that can perform robust MR

estimation in such study populations. It is important to note that with a small sample size, the systematic finite sample bias can be substantial [102] in the above-mentioned existing approaches. In this article, we aim to address this specific scenario by proposing a robust approach that performs valid MR estimation in small sample studies by allowing a large number of instrumental variables. We relax the exclusion restriction and allow to accommodate many weak IVs that belong to the pleiotropy-weak-IV methods category.

In recent MR studies, there has been a notable focus on utilizing the first-order moment, which represents the average effects of multiple IVs on the exposure variable. However, relying on data-driven selection criteria for strong IVs, such as F-statistics, can introduce significant bias. Moreover, the limited sample size can undermine the reliability of MR methods, particularly in high-dimensional models involving numerous genetic variants as IVs, such as MR Lasso [106]. To address these challenges in small-scale studies, we propose a novel two-stage approach to MR called TS-RE (Two-Stage with Random Effects). Our method shifts the focus from estimating the mean effect sizes of individual IVs to modeling the second-order moment, encompassing the variance and covariance components of the effects of multiple IVs on both the exposure and outcome variables. A distinguishing feature of our approach is the inclusion of many weak IVs, with the simple requirement that the variance of the IVs' effects is non-zero. In fact, our proposed method allows for all IVs to be weak. The second-order moment estimator in our approach utilizes individual-level data to calculate the genetic correlation matrix (GRM) using the genetic IVs. The IVs included in the analysis are genetic variants that explain a significant proportion of the variance in the exposure variable. By adopting this framework, our proposed approach offers a robust alternative to standard MR analysis, effectively addressing the challenges posed by a multitude of weak IVs and pleiotropic IVs.

In Section 2, we provide a comprehensive review of the 2SLS method, along with two commonly used ratio estimators, namely the inverse variance weighting and Egger methods. We highlight the strengths and limitations of these approaches in causal effect estimation. Next, we introduce our proposed method, TS-RE, and discuss its ability to relax the exclusion restriction assumption. We present theoretical derivations that demonstrate the advantages of TS-RE, particularly in scenarios involving weak IVs and small sample sizes. We emphasize how TS-RE overcomes the limitations of other MR methods and

provides more reliable causal effect estimates. In Section 4, we conduct extensive simulations to compare the performance of our proposed methods across various scenarios with four commonly used MR methods, MR-Inverse Variance Weight (IVW), MR-Egger, MR-Simple Median, and MR-Weighted Median The four methods we compare belong to the pleiotropy-IV methods category. Through these simulations, we illustrate the superiority of TS-RE in terms of bias reduction, efficiency, and robustness under different conditions. Furthermore, in Section 5, we apply our proposed methods to investigate the effects of body mass index (BMI) on systolic blood pressure (SBP) using data on the Black British population in the UK Biobank. Finally, Section 6 concludes the article with a discussion of our findings and the practical implications of our proposed methods.

4.2 Materials and methods

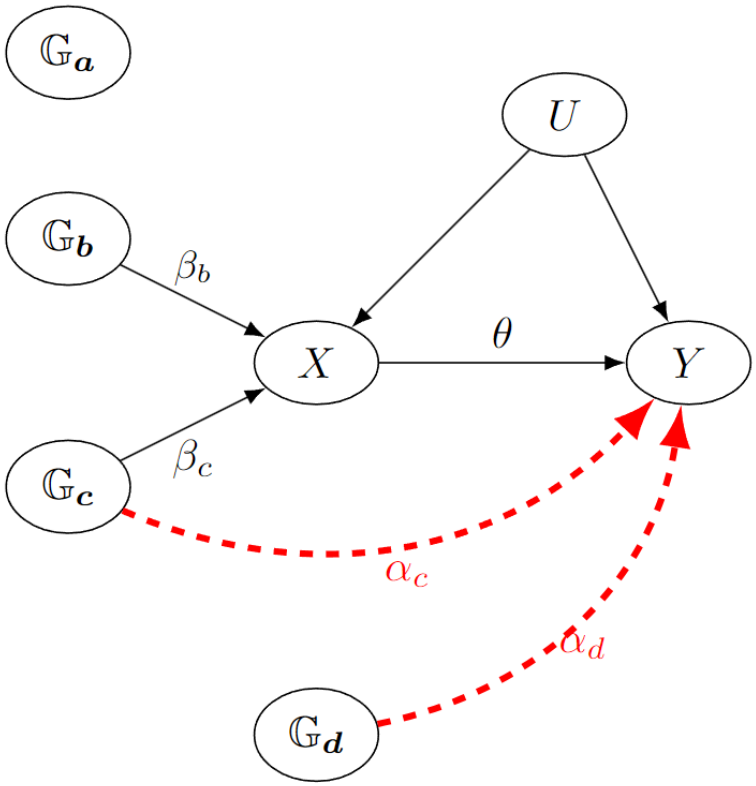
4.2.1 Overview of Existing Methods

Suppose we have an exposure variable X, and an outcome Y that are causally related, with X affecting Y according to the linear model:

$$Y = \theta X + \epsilon, \quad \epsilon = U + \epsilon', \tag{4.1}$$

The parameter of interest is θ, which represents the causal effect of X on Y. The model error ϵ consists of two components: unmeasured confounders U and a residual error term ϵ'. Standard regression estimators like ordinary least squares (OLS) or generalized least squares cannot provide consistent and unbiased estimates for θ due to the presence of unmeasured confounders U that are associated with both X and Y. To address this issue, MR studies utilize genetic variants, such as Single Nucleotide Polymorphisms (SNPs), as IVs. However, it is important to note that not all selected SNPs are assumed to be valid IVs that satisfy the necessary assumptions aforementioned. There are four possible relationships between the IVs and the phenotypes X and Y: (1) $\mathbb{G}_{\boldsymbol{a}}$ is not related to either X or Y; (2) $\mathbb{G}_{\boldsymbol{b}}$ has a direct effect on X and an indirect effect on Y; (3) $\mathbb{G}_{\boldsymbol{c}}$ has direct effects on both X and Y; (4) $\mathbb{G}_{\boldsymbol{d}}$ has a direct effect on Y but no relationship with X. These different relationships between the SNPs and the phenotypes are illustrated in Fig 4.2.

Figure 4.2: A more general Mendelian Randomization model: we are interested in the causal effect θ. Four potential relationships considered: (1) $\mathbb{G}_a$ related to neither X nor Y; (2) $\mathbb{G}_b$ with direct effect on X and indirect effect on Y; (3) $\mathbb{G}_c$ with direct effects both on X and Y; (4) $\mathbb{G}_d$ with direct effect on Y but no relationship with X.



Let's denote M as the total number of IVs, and M_a, M_b, M_c, M_d as the number of IVs in each respective group. The direct effect of the k-th IV (G_k) on the exposure and outcome variables is represented by β_k and α_k, where $k = 1, \ldots, M$. Therefore, we can express the overall effect of the k-th IV on the exposure variable as γ_{xk}, and the overall effect on the

outcome variable as γ_{yk}:

$$\begin{cases} \gamma_{xk} = 0, \gamma_{yk} = 0, \text{if } G_k \in \mathbb{G}_{\boldsymbol{a}}; \\ \gamma_{xk} = \beta_{bk}, \gamma_{yk} = \theta\beta_{bk}, \text{if } G_k \in \mathbb{G}_{\boldsymbol{b}}; \\ \gamma_{xk} = \beta_{ck}, \gamma_{yk} = \alpha_{ck} + \theta\beta_{ck}, \text{if } G_k \in \mathbb{G}_{\boldsymbol{c}}; \\ \gamma_{xk} = 0, \gamma_{yk} = \alpha_{dk}, \text{if } G_k \in \mathbb{G}_{\boldsymbol{d}}. \end{cases} \tag{4.2}$$

where different IVs are assumed independent. Here, the generative model for X, Y, G is

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\gamma}_{\boldsymbol{x}}^{\boldsymbol{T}} \boldsymbol{G} + \mathbf{e}_x \\ &= \boldsymbol{\beta}_{\boldsymbol{b}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{b}} + \boldsymbol{\beta}_{\boldsymbol{c}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{c}} + \mathbf{e}_x, \\ \mathbf{Y} &= \boldsymbol{\gamma}_{\boldsymbol{y}}^{\boldsymbol{T}} \boldsymbol{G} + \mathbf{e}_y \\ &= \theta\mathbf{X} + \boldsymbol{\alpha}_{\boldsymbol{c}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{c}} + \boldsymbol{\alpha}_{\boldsymbol{d}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{d}} + \boldsymbol{\epsilon}, \end{aligned} \tag{4.3}$$

where $\boldsymbol{G} = (\mathbb{G}_{\boldsymbol{a}}, \mathbb{G}_{\boldsymbol{b}}, \mathbb{G}_{\boldsymbol{c}}, \mathbb{G}_{\boldsymbol{d}})$ is the $n \times M$ matrix of IVs, and $\boldsymbol{\beta}, \boldsymbol{\alpha}$ are vector of corresponding coefficients, $\mathbf{e}_x, \mathbf{e}_y$ are error terms.

First, assuming all IVs are from $\mathbb{G}_{\boldsymbol{b}}$, then the 2SLS method can be used if the individual-level data are available:

$$\hat{\theta}_{2SLS} = (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{P}_{\mathbb{G}_{\boldsymbol{b}}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{P}_{\mathbb{G}_{\boldsymbol{b}}} \boldsymbol{Y} \tag{4.4}$$

where $\boldsymbol{P}_{\mathbb{G}_{\boldsymbol{b}}} = \mathbb{G}_{\boldsymbol{b}} (\mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{b}})^{-\mathbf{1}} \mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}}$ is the projection matrix. Let $\boldsymbol{G} = \mathbb{G}_{\boldsymbol{b}}$ in Eq 4.3, we can obtain the estimates of coefficients $\hat{\boldsymbol{\gamma}}_{\boldsymbol{x}} = (\mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{b}})^{-\mathbf{1}} \mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}} \boldsymbol{X}$, $\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}} = (\mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{b}})^{-\mathbf{1}} \mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}} \boldsymbol{Y}$, and variance $se(\hat{\boldsymbol{\gamma}}_{\boldsymbol{x}})^2, se(\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}})^2$ are the diagonal elements of the matrix of $(\mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{b}})^{-1} \sigma_{e_x^*}^2$,$(\mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{b}})^{-1} \sigma_{e_y^*}^2$, where $\sigma_{e_x^*}, \sigma_{e_y^*}$ is residual standard error. If the IVs are perfectly uncorrelated and the effects $\beta_{b1}, \ldots, \beta_{bM_b}$ are independent, the off-diagonal elements of $(\mathbb{G}_{\boldsymbol{b}}^{\boldsymbol{T}} \mathbb{G}_{\boldsymbol{b}})^{-1}$ are all zero. This means that the 2SLS estimator can be viewed as a weighted average of multiple ratios $\hat{\gamma}yk/\hat{\gamma}xk$ for $G_k \in \mathbb{G}_{\boldsymbol{b}}$. This equivalence allows us to use the IVW method with summary statistics, which is given by:

$$\begin{aligned} \hat{\theta}_{IVW} &= \frac{\sum_k \hat{\gamma}_{yk} \hat{\gamma}_{xk} se(\hat{\gamma}_{yk})^{-2}}{\sum_k (\hat{\gamma}_{xk})^2 se(\hat{\gamma}_{yk})^{-2}} \\ &= (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{P}_{\mathbb{G}_{\boldsymbol{b}}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{P}_{\mathbb{G}_{\boldsymbol{b}}} \boldsymbol{Y} \end{aligned} \tag{4.5}$$

The IVW estimator will slightly differ from the 2SLS estimator in finite samples, as the correlation between independent genetic variants will not exactly equal zero [90], but the two estimates will be equal asymptotically [107]. The IVW estimator can also be regarded as a regression model as follows:

$$\hat{\gamma}_{yk} = \theta\hat{\gamma}_{xk} + \epsilon_k, \quad \epsilon_k \sim N(0, \phi_I^2 se(\hat{\gamma}_{yk})^2).$$

If IVs are all from $\mathbb{G}_{\boldsymbol{b}}$, then each of the ratios estimates $\hat{\gamma}_{yk}/\hat{\gamma}_{xk}$ will be a consistent estimate of the causal effect θ, thus the 2SLS and IVW (a weighted mean of multiple ratio estimates) will be a consistent estimate of θ. $\phi_I^2 = 1$ is specified under the fixed-effect(FE) IVW without any random intercept if all IVs are valid from $\mathbb{G}_{\boldsymbol{b}}$; if some IVs are invalid from $\mathbb{G}_{\boldsymbol{c}}$, but the average direct effect of G on Y is zero (referred to as "balanced pleiotropy"), then the model will have a random intercept $\alpha_0 \sim N(0, \tau^2)$ and $\phi_I^2 > 1$ is assumed under the random-effect IVW. To combat heterogeneity that some IVs are invalid, a random-effects (RE) IVW is used in this paper.

When IVs are from $\mathbb{G}_{\boldsymbol{c}}$ with pleiotropic effect, the ratio γ_{yk}/γ_{xk} becomes $\theta + \alpha_{ck}/\beta_{ck}$. If the 2SLS and IVW estimators mistakenly include IVs from $\mathbb{G}_{\boldsymbol{c}}$ as part of $\mathbb{G}_{\boldsymbol{b}}$, they will be biased toward:

$$\theta + \frac{\sum_{k=1}^{M_c} \gamma_{xk} se(\gamma_{yk})^{-2} \alpha_{ck}}{\sum_k (\gamma_{xk})^2 se(\gamma_{yk})^{-2}} = \theta + Bias(\boldsymbol{\alpha_c}, \boldsymbol{\gamma_c}). \tag{4.6}$$

This implies that if the exclusion restriction is satisfied all $a_{ck} = 0$, 2SLS and IVW estimates are unbiased. However, this will not be universally plausible. To address this issue, the Egger method assumes an average pleiotropic effect for all IVs [107], which assumes that the effects β_k, α_k are all random variables. It estimates the average direct effect α_0 as part of the analysis, which is assumed to be zero in the IVW method. for Egger. Using the same weights in IVW, the Egger estimator is:

$$\begin{aligned} \hat{\gamma}_{yk} &= \alpha_0 + \theta\hat{\gamma}_{xk} + \epsilon_k, \quad \epsilon_k \sim N(0, \phi_E^2 se(\hat{\gamma}_{yk})^2) \\ \hat{\theta}_{Egger} &= \frac{Cov_w(\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{x}})}{Var_w(\hat{\boldsymbol{\gamma}}_{\boldsymbol{x}})} = \theta + \frac{Cov_w(\hat{\boldsymbol{\alpha}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{c}})}{Var_w(\hat{\boldsymbol{\beta}})} \end{aligned} \tag{4.7}$$

Due to the potential overdispersion resulting from the pleiotropic effects of IVs, it is recommended to employ a random intercept assuming $\phi_E^2 > 1$. In the Egger method, the weighted covariance (Cov_w) and weighted variance (Var_w) are computed using the inverse-variance weights ($se(\hat{\gamma}_{yk})^{-2}$) and the vector of IV coefficients ($\boldsymbol{\beta} = (\boldsymbol{\beta_b}, \boldsymbol{\beta_c})$). To satisfy the necessary condition for Egger, it is required that the correlation between the effects of IVs on the exposure and the direct effects of IVs on the outcome is zero, denoted as $Cov(\boldsymbol{\alpha_c}, \boldsymbol{\beta_c}) = 0$. This condition, known as the InSIDE (Instrument Strength Independent of Direct Effect) assumption, can be viewed as a weakened version of the exclusion restriction assumption. Under the InSIDE assumption and as the sample size increases, the weighted covariance $Cov_w(\hat{\boldsymbol{\alpha}}_{\boldsymbol{c}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{c}})$ converges to zero as n tends to infinity, which implies that $Cov_w(\boldsymbol{\alpha_c}, \boldsymbol{\beta_c}) \xrightarrow{n\to\infty} 0$. Consequently, the Egger estimate becomes a consistent estimate of θ. In Eq 4.7, the intercept term α_0 represents the average pleiotropic effect of the genetic variants included in the analysis [108]. If $\alpha_0 = 0$, the estimates obtained by Egger and IVW methods will be the same. However, the standard error of the Egger method will be larger than that of IVW. Instead of using summary statistics for Egger, we can expand Eq 4.7 to show the difference between Egger and 2SLS with individualized data:

$$\begin{aligned}\hat{\theta}_{Egger} &= \frac{\sum_k se(\hat{\gamma}_{yk})^{-2} \sum_k \hat{\gamma}_{yk}\hat{\gamma}_{xk} se(\hat{\gamma}_{yk})^{-2} - \sum_k se(\hat{\gamma}_{yk})^{-2}\hat{\gamma}_{yk} \sum_k \hat{\gamma}_{xk} se(\hat{\gamma}_{yk})^{-2}}{\sum_k se(\hat{\gamma}_{yk})^{-2} \sum_k (\hat{\gamma}_{xk})^2 se(\hat{\gamma}_{yk})^{-2} - (\sum_k \hat{\gamma}_{xk} se(\hat{\gamma}_{yk})^{-2})^2} \\ &= \frac{\boldsymbol{1^T G^T G 1 X^T P_G Y - X^T G 1 Y^T G 1}}{\boldsymbol{1^T G^T G 1 X^T P_G X - X^T G 1 X^T G 1}}\end{aligned} \tag{4.8}$$

where $\mathbf{1}$ is a $(M_b + M_c) \times 1$ vector with all elements equal to 1 and $G = (\mathbb{G}_b, \mathbb{G}_c)$. However, it is unrealistic to assume that the InSIDE assumption always holds all IVs. The limitation of Egger (and related methods) has been discussed [109], which depends on the orientation of SNPs to get an average pleiotropic effect α_0

Assuming all IVs effects are random variables, we derived the bias term for IVW (2SLS) and Egger if all IVs have a direct effect on X (see Supporting information Eq A.6) and Eq A.7: $Bias_{IVW} = \frac{M_c E(\beta_{ck}\alpha_{ck})}{M_b E(\beta_{bk}^2) + M_c E(\beta_{ck}^2)}$ and $Bias_{Egger} = \frac{M_c}{M} \frac{E(\beta_c k \alpha_c k) - E(\beta_c k)E(\alpha_c k)}{Var(\beta)}$. To get an unbiased estimate, Egger requires the InSIDE assumption that $Var(\alpha_c, \beta_c) = 0$, while IVW needs both InSIDE assumption $Var(\alpha_c, \beta_c) = 0$ and balanced pleiotropic assumption $E(\alpha_c) = 0$. A selection to avoid IVs with weak effect is required for 2SLS, IVW and Egger,

and including IVs from $\mathbb{G}_a$ and $\mathbb{G}_d$ will lead to a large bias to all three methods.

To address the problem that includes many invalid IVs, the other two commonly used methods Simple Median [100] and Weighted Median [100] focus on using the median of M ordered ratio estimator for each IV as the estimate of θ and allow up to 50% IVs to violate the exclusion restriction. However, the two median estimators are low-powered and sometimes biased when the proportion of invalid IVs is greater than 50%. Furthermore, those biases can be exaggerated under a finite small sample size. Details about these and other existing approaches are listed in Table 4.1.

4.2.2 Our Proposed Approach

The methods of 2SLS, IVW, and Egger aim to estimate the causal effect of variable X on variable Y by utilizing the weighted mean of multiple ratios γ_{yk}/γ_{xk}. However, these methods rely on the assumption that the selected IVs have significant and strong effects on the exposure variable X. If the selected IVs have weak effects, such as in cases with small sample sizes, these methods may not provide precise estimates of the causal effect. Moreover, including IVs from groups $\mathbb{G}_{\boldsymbol{a}}$ and $\mathbb{G}_{\boldsymbol{d}}$ can lead to misspecification in these existing methods. Table 4.1 compares our proposed TS-RE method with popular existing

methods.

Method	Invalid (c)	Weak IV	Comment
TSLS [87]	No	No	individual data, $M \leq n$
IVW (FE) [90]	No	No	all IVs are valid
IVW (RE) [90]	Yes	No	balanced pleiotropy, InSIDE
Egger [107]	Yes	No	InSIDE Assumption, large SE
Lasso [101]	Yes	No	need choose tuning parameter
Simple Median [100]	Yes	No	$\leq 50\%$ invalid IVs
Weighted Median [100]	Yes	No	$\leq 50\%$ invalid IVs
LIML [92]	No	Yes	individual data, $M \leq n$
CUE [92]	No	Yes	individual data, $M \leq n$
Genus-MAWII [105]	Yes	Yes	individual data, $M \leq n$
debiased-IVW [104]	Yes	Yes	balanced pleiotropy, InSIDE
TS-RE	Yes	Yes	$E(\beta_{ck}\alpha_{ck}) = 0$

Table 4.1: Comparison of different MR methods, including whether IVs violated exclusion restriction and weak IVs are allowed

We introduce a new method that models the variance of multiple variants instead of estimating individual effect sizes. Our proposed TS-RE approach assumes a random effects model for the IVs' effect sizes. It can accommodate a large number of IVs and is less sensitive to the presence of weak instruments. In this framework, the variance components can be used to estimate the causal effect θ. The TS-RE method allows for the inclusion of IVs from all four groups in Fig 4.2, which makes it more flexible than Egger. However, it still requires at least two IVs to come from either $\mathbb{G}_{\boldsymbol{b}}$ or $\mathbb{G}_{\boldsymbol{c}}$ to estimate the variance. The generative model for the TS-RE approach is listed in Equation 4.9.

$$\begin{aligned} X &= \mathbb{G}_{\boldsymbol{b}}\boldsymbol{\beta_b} + \mathbb{G}_{\boldsymbol{b}}\boldsymbol{\beta_c} + e_x, \\ Y &= \theta X + \mathbb{G}_{\boldsymbol{c}}\boldsymbol{\alpha_c} + \mathbb{G}_{\boldsymbol{b}}\boldsymbol{\alpha_d} + e_y. \end{aligned} \tag{4.9}$$

where

$$\begin{aligned}
\beta_{bk} &\sim N(\mu_{g_{\boldsymbol{b}}}, \sigma^2_{g_{\boldsymbol{b}}}), k = 1, \ldots, M_b \\
\begin{pmatrix} \beta_{ck} \\ \alpha_{ck} \end{pmatrix} &\sim N\left[\boldsymbol{\mu_{g_c}}, \begin{pmatrix} \sigma^2_{g^x_{\boldsymbol{c}}} & \rho_{g_{\boldsymbol{c}}}\sigma_{g^x_{\boldsymbol{c}}}\sigma_{g^y_{\boldsymbol{c}}} \\ \rho_{g_{\boldsymbol{c}}}\sigma_{g^y_{\boldsymbol{c}}}\sigma_{g^x_{\boldsymbol{c}}} & \sigma^2_{g^y_{\boldsymbol{c}}} \end{pmatrix}\right], k = 1, \ldots, M_c \\
\alpha_{dk} &\sim N(\mu_{g_d}, \sigma^2_{g_{\boldsymbol{d}}}), k = 1, \ldots, M_d
\end{aligned} \tag{4.10}$$

Here $\mu_{g(\cdot)}$ is the mean effect and $\rho_{\mathbb{G}_{\boldsymbol{c}}}$ is the genetic correlation coefficient and quantifies how the InSIDE assumption is violated. If the InSIDE assumption holds, then $\rho_{\mathbb{G}_{\boldsymbol{c}}} = 0$. Suppose that (1) residual terms e_y, e_x, and $\boldsymbol{G}$ are independent of each other; (2) different G_k are independent and the effects of G_k are independent (no linkage disequilibrium or interactions). Then the genetic variances and covariance of X and Y explained by the included IVs can be written as:

$$\begin{aligned}
Var_g(\mathbf{X}) =& \mathbb{G}_{\boldsymbol{b}}\mathbb{G}^{\boldsymbol{T}}_{\boldsymbol{b}}E(\beta^2_{bk}) + \mathbb{G}_{\boldsymbol{c}}\mathbb{G}^{\boldsymbol{T}}_{\boldsymbol{c}}E(\beta^2_{ck}), \\
Var_g(\mathbf{Y}) =& \theta^2 Var_g(\mathbf{X}) + \mathbb{G}_{\boldsymbol{c}}\mathbb{G}^{\boldsymbol{T}}_{\boldsymbol{c}}[\theta E(\alpha_{ck}\beta_{ck}) + E(\alpha^2_{ck})] + \mathbb{G}_{\boldsymbol{d}}\mathbb{G}^{\boldsymbol{T}}_{\boldsymbol{d}}E(\alpha^2_{dk}), \\
Cov_g(\mathbf{X}, \mathbf{Y}) =& \theta Var_g(\mathbf{X}) + \mathbb{G}_{\boldsymbol{c}}\mathbb{G}^{\boldsymbol{T}}_{\boldsymbol{c}}E(\alpha_{ck}\beta_{ck}) \\
\xrightarrow{E(\alpha_{ck}\beta_{ck})=0}& \theta Var_g(\mathbf{X}).
\end{aligned} \tag{4.11}$$

If the assumption that $E(\alpha_{ck}\beta_{ck}) = 0$ is valid, then the genetic covariance of X and Y is $\theta Var_g(\mathbf{X}|G)$, which is the causal effect θ times the genetic variance of X. Thus, we considered using a second-moment estimator for the genetic variance and covariance of phenotype variables explained by IVs, instead of means, for the causal estimation where weak IVs are included. After taking the cross-product, the original model in Eq 4.1 can be written as

$$Y_i X_j = \theta X_i X_j + e_{yi} X_j, \quad i, j = 1, \ldots, n. \tag{4.12}$$

Here the OLS of θ is still biased since the residual component $\epsilon_{yi}X_j$ is associated with X_iX_j. We involve the SNPs as IVs in this cross-product model to address this problem. In addition, since we can assume that different observations are independent, using the covariance of X_i and X_j $(i < j)$ to estimate the genetic effect of G on X can allow us

to eliminate the variance residual term X_iX_i. After including all IVs, we perform the following model:

$$\begin{aligned} X_iX_j &= \eta A_{ij} + e_{ij}^{xx}, \\ Y_iX_j &= \delta A_{ij} + e_{ij}^{yx}. \end{aligned} \tag{4.13}$$

where $i < j;\ i, j = 1, \ldots, n,\ A_{ij}$ is the ij-th element of the genetic relationship matrix (GRM), $\mathbf{A} = \boldsymbol{GG^T}/M$, and there are $N = \frac{n(n-1)}{2}$ observations in the regression model. Here the genetic data are standardized and $E(A_{ij}) = 0$. The GRM $\mathbf{A}$ is used to construct a second-moment estimator for estimating the variance and covariance of phenotypes captured by the included SNPs.

As a two-stage estimator, the first stage for our TS-RE is to estimate $\hat{\eta}$ and $\hat{\delta}$ in Eq 4.13. The second stage is to calculate the ratio $\hat{\theta}_{TS-RE} = \frac{\hat{\eta}}{\hat{\delta}}$. Similar to the TSLS [87], our TS-RE estimator is also equivalent to a generalized method of moment (GMM) estimator. Denote $\boldsymbol{Vec(A), Vec(X \otimes X), Vec(X \otimes Y)}$ to be the $N = n(n-1)/2$ dimensional vectors in Eq 4.13. Given certain independence conditions in the Supporting information A.2.1, we can prove $E[A_{ij}e_{ij}^{yx}))] = 0$. Then, the generalized method of moment estimator is given by solving

$$g(\hat{\theta}) = \frac{1}{N}\boldsymbol{Vec(A)}^T(\boldsymbol{Vec(X \otimes Y)} - \hat{\theta}\boldsymbol{Vec(X \otimes X)}) = 0,$$

which leads to the following estimator:

$$\begin{aligned} \hat{\theta}_{GMM} &= [\boldsymbol{Vec(A)^T Vec(X \otimes X)}]^{-1}\boldsymbol{Vec(A)^T Vec(X \otimes Y)} \\ &= \widehat{Cov}\left[A_{ij}, Y_iX_j\right] / \widehat{Cov}\left[A_{ij}, X_iX_j\right] \end{aligned} \tag{4.14}$$

Through our proof in the Supporting information A.2.2, the bias of our TS-RE estimator $\hat{\theta}_{TS-RE}$ is $\frac{M_cE(\beta_{ck}\alpha_{ck})}{M_bE(\beta_{bk}^2)+M_cE(\beta_{ck}^2)}$. When $E(\beta_{ck}\alpha_{ck}) = 0$, the estimator will be an unbiased consistent estimator of θ

$$\sqrt{N}[\hat{\theta}_{TS-RE} - \theta] \xrightarrow{\mathcal{D}} N(0, \tau^2)$$

, and

$$\tau^2 = \frac{M[M_b E(\beta_{bk}^2) + M_c E(\beta_{ck}^2) + \sigma_{e_x}^2][M_c E(\alpha_{ck}^2) + M_d E(\alpha_{dk}^2) + \sigma_{e_y}^2]}{[M_b E(\beta_{bk}^2) + M_c E(\beta_{ck}^2)]^2} \tag{4.15}$$

where $E(\beta_{g(\cdot)}^2) = \mu_{g(\cdot)}^2 + \sigma_{g(\cdot)}^2$. The key strength of our TS-RE method is that we use SNPs from four groups without selection. IVs from $\mathbb{G}_a, \mathbb{G}_d$ do not contribute to the genetic variance of X and genetic covariance of X, Y, thus do not impact the estimation of δ and η. The TS-RE method can still have an unbiased estimator even including those invalid IVs.

First, assuming all IVs are from $\mathbb{G}_b$, then the bias term will be 0 and the asymptotic variance term in Eq 4.15 can be written as

$$\tau^2 = (\frac{1}{E(\beta_{bk}^2)} + \frac{\sigma_{e_x}^2}{M_b E(\beta_{bk}^2)^2})\sigma_{e_y}^2.$$

This indicates for a finite sample that fixing N, when all IVs are valid from $\mathbb{G}_b$, stronger effect μ_{gb}, larger variance σ_{gb}^2, and a larger number of IVs M_b can lead to smaller asymptotic variance. Hence even when the IVs are weak with $mu_{gb} \approx 0$, including a large number of weak IVs that explains a large proportion of the overall exposure variance could give us an efficient estimator. Then, assuming IVs are from $\mathbb{G}_b, \mathbb{G}_c$ and $E(\beta_{bk}^2) = E(\beta_{ck}^2)$, Eq 4.15 can be written as

$$\tau^2 = (\frac{1}{E(\beta_{bk}^2)} + \frac{\sigma_{e_x}^2}{(M_b + M_c)E(\beta_{bk}^2)^2})M_c E(\alpha_{ck}^2) + \sigma_{e_y}^2.$$

Including a large number of IVs with direct effects on exposure variable $M_b + M_c$ can control the variance of the TS-RE estimator even if the IVs' effects are weak. Including IVs with directional pleiotropic effects $\mu_{\mathbb{G}_c^y} \neq 0$ leads to a larger τ^2. Among IVs with direct effects on X from $\mathbb{G}_b, \mathbb{G}_c$, the variance will increase if the proportion of IVs from $\mathbb{G}_c$ increases. When IVs from $\mathbb{G}_c, \mathbb{G}_d$ is also included, a higher proportion of null IVs from groups $\mathbb{G}_a, \mathbb{G}_d$ will lead to a larger SE.

4.3 Simulation

4.3.1 Simulation set-ups

We performed a large number of simulations to study the performance of our proposed method in comparison with different existing methods for MR analysis. In our simulations, we compared the TS-RE method with five pleiotropy-IV MR approaches based on summary statistics: Simple Median, Weighted Median, IVW, Egger, and Lasso. We also included one pleiotropy-weak-IV method, dIVW. We did not include TSLS and the weak-IV methods in Table 4.1 using individual-level data since those methods require the sample size to be larger than the number of IVs.

In our simulations, we considered $M = M_a + M_b + M_c + M_d$ variants belong to four groups: (1) M_a SNPs related to neither X nor Y; (2) M_b SNPs with direct effect on X and indirect effect on Y; (3) M_c SNPs with both direct effects on X and Y; (4)M_d SNPs in $\mathbb{G}_{\boldsymbol{d}}$ with only direct effect on Y but no relationship with X. Our simulation model is described in Fig 4.2. Note that variants belonging to $\mathbb{G}_{\boldsymbol{b}}$ will be valid IVs. For the summary statistics-based approaches, we performed a simple linear regression for each IV to get the summarized statistics. We considered using all IVs for the TS-RE method and the top 20 significant (based on p-value) IVs for other methods. Considering top significant variants will generally eliminate the weak IVs and invalid IVs in $\mathbb{G}_{\boldsymbol{a}}$ and $\mathbb{G}_{\boldsymbol{d}}$ which is required for other methods that use the summary statistics. The supplementary file **??** includes both all IVs and top 20 IVs results for all methods. For each simulation setup, we generated 100 datasets and estimated the causal effects. We reported the bias and the standard error of θ from these 100 simulations. to check the bias and standard error (SE).

We followed the following procedure for each simulation. First, the minor allele frequency (MAF) f_k of each SNP k ($k = 1, \ldots, M$) was independently generated from a uniform distribution $U(0.2, 0.3)$ and the corresponding genotypes were simulated from a binomial distribution $Bin(2, f_k)$. Second, we generate the effects of IVs from Eq 4.10. We used different $\mu_{(\cdot).g}, \sigma^2_{(\cdot)g}$ to generate IVs effects. Assuming the total variance for X and Y are σ^2_X and σ^2_Y, where $\sigma^2_X = \sigma^2_{G_x} + \sigma^2_{e_x}, \sigma^2_Y = \sigma^2_{G_y} + \sigma^2_{e_y}$, where $\sigma^2_{G_x}, \sigma^2_{G_y}$ are the total genetic variances of included SNPs as IVs. If only IVs from $\mathbb{G}_b$ were included, then $\sigma^2_{G_x} = M_b \sigma^2_{gb}$,

$\sigma^2_{G_y} = \theta^2 M_b \sigma^2_{gb}$, and the residual variance component are $\sigma^2_{e_y}, \sigma^2_{e_x}$. The proportion of total variance explained by included genetic IVs is defined as conditional heritability (Her) $Her_X = \sigma^2_{G_x}/\sigma^2_X$. Then, X and Y were generated from Eq 4.9. The primary objective was to estimate the causal effect θ.

We mainly compared our TS-RE method with existing MR methods given a small sample size of $n = 1000$ with different numbers of IVs. We considered three scenarios: (1) only M_b valid IVs from $\mathbb{G}_b$; (2) a mixture of IVs from $\mathbb{G}_b$ and IVs with pleiotropic effect from $\mathbb{G}_c$; (3) a mixture of IVs from all four groups. For IVs from $\mathbb{G}_b, \mathbb{G}_c$ that have direct effects on X, we varied the proportion of IVs with weak effects on X. Weak IVs effect on X were generated from a normal distribution with a mean 0 and strong IVs were generated with a mean 0.2. For IVs with pleiotropic effects from $\mathbb{G}_c$, the effects were generated from the following four sub-scenarios: (a) Balanced pleiotropy ($\mu_{\alpha_c} = 0$), InSIDE assumption satisfied($\rho_{g_c} = 0$); (b) Directional pleiotropy ($\mu_{\alpha_c} = 0.1$), InSIDE assumption satisfied($\rho_{g_c} = 0$); (c) Balanced pleiotropy ($\mu_{\alpha_c} = 0$), InSIDE assumption violated ($\rho_{g_c} = 0.6$); (d) Directional pleiotropy ($\mu_{\alpha_c} = 0.1$), InSIDE assumption violated ($\rho_{g_c} = 0.6$).

4.3.2 Weak IVs from $\mathbb{G}_b$: impact of number of IVs and genetic variance

In this and following two simulation studies, we considered $M_a = 0, M_c = 0, M_d = 0$. The simulation model was

$$\begin{aligned} X &= \mathbb{G}_{\boldsymbol{b}} \boldsymbol{\beta}_{\boldsymbol{b}} + e_x, \\ Y &= \theta X + e_y. \end{aligned} \tag{4.16}$$

IVs effects $\beta_{bk} \sim N(0, \sigma^2_{gb}), k = 1, \ldots, M_b$. We varied the parameter σ_{gb} to be 0.01, 0.03, 0.05, and the number of included IVs $M_b = 100, 1000, 5000$. Specifically, we varied the causal effect θ between 0.1 and 0.3. The residual variance parameter $\sigma^2_{e_x}$ was fixed at 2. Additionally, for $M_b = 5000$, we used $\sigma^2_{e_x} = 17$ while keeping the genetic variance constant but modifying the heritability to be small. Fig 4.3, Fig 4.4 show the results for TS-RE using all IVs and the other MR methods using selected top 20 most significant IVs.

Figure 4.3: Empirical distributions of the estimates of the causal effect $\theta = 0.1$ by the methods with different numbers of IVs and different genetic variances. TS-RE used all IVs while other MR methods used the selected top 20 most significant IVs.

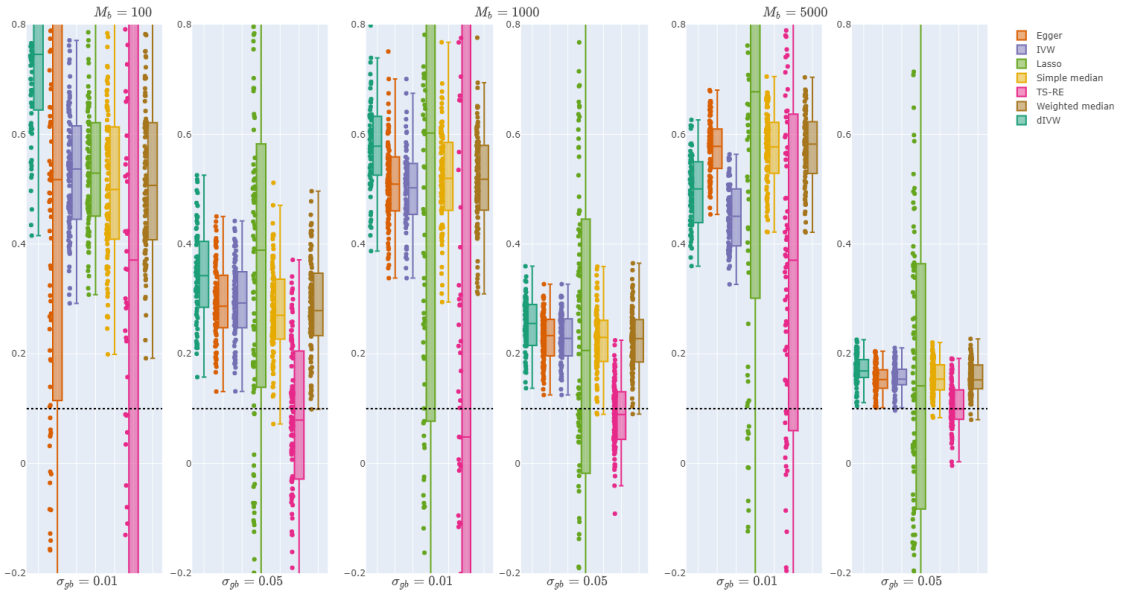
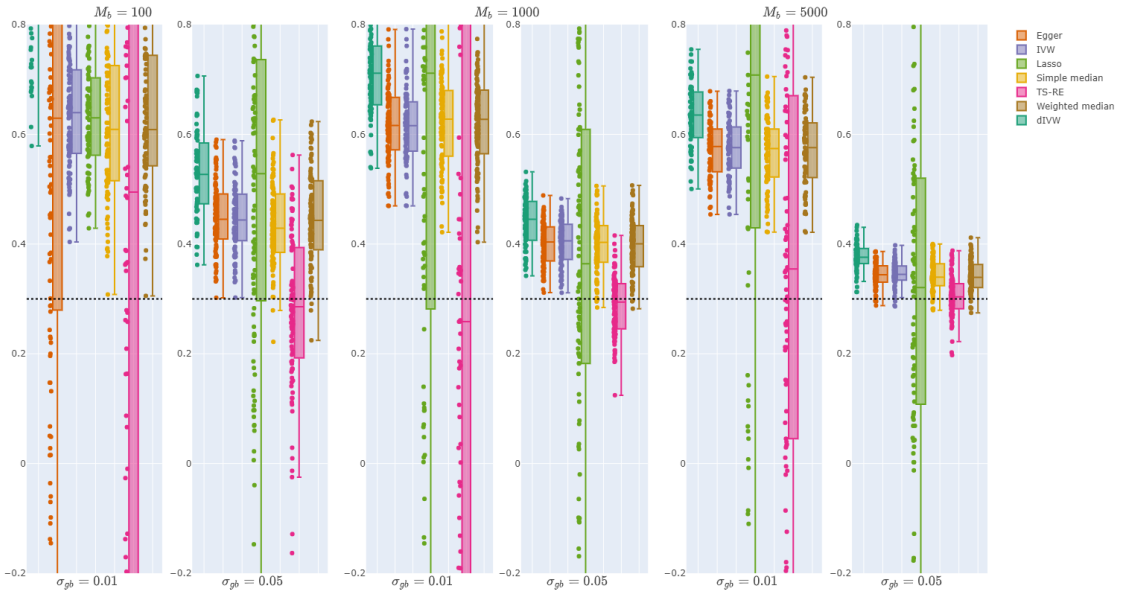


Figure 4.4: Empirical distributions of the estimates of the causal effect $\theta = 0.3$ by the methods with different numbers of IVs and different genetic variances. TS-RE used all IVs while other MR methods used the selected top 20 most significant IVs.



The observations revealed that the bias decreased for our proposed TE-RE method as Her increased and all IVs were utilized. The bias for the TS-RE method remained consistently small as long as the Her values were greater than 0.1. The increase in Her could be achieved through a larger $\sigma^2_{g_b}$ or a larger M_b. When the variances of the residuals increased, the Her decreased from 0.86 to 0.42 for $M_b = 5000$. Despite this change, the TE-RE method still provided an unbiased estimate, albeit with a larger SE, while other methods' bias increased a lot. On the other hand, when the top 20 significant IVs were used for the TS-RE method, it resulted in a substantial overestimation of θ. This overestimation occurred because the effects of the selected 20 IVs were extreme values (e.g., 0.5, -0.5), leading to an overestimation of the genetic variance. In addition, 20 IVs is not enough for the TS-RE since this method focuses on the second moment, in other words, the variant of the IVs to estimate the causal effect. However, biases for TS-RE with 20 selected IVs were at the same level as other MR methods. This indicated that our method was not worse than other methods, even though a large number of IVs is preferred for our TS-RE when

the sample size is small. Additionally, all the other methods produced larger biases when the true causal effect was small ($\theta = 0.1$).

Our proposed method demonstrated a notable advantage by consistently delivering unbiased results even when all IVs are weak. In contrast, other MR methods tend to exhibit significant biases even when limited to the top 20 significant IVs. This robustness in the face of weak IVs is a key strength of our approach. The Egger method, in particular, exhibited greater variation compared to the other methods, primarily due to the estimation of the average pleiotropic effect. The selection of the top 20 significant IVs in the Egger method significantly increased SE compared to using all IVs (as shown in Supplementary Table S1). This may be attributed to the fact that a larger number of IVs is required to accurately estimate the average pleiotropic effect.

Interestingly, increasing the number of IVs, whether using all IVs or just the top 20, did not reduce the biases observed in the other MR methods (as shown in Table S1). However, increasing the genetic variance parameter $\sigma^2_{g_b}$ resulted in lower bias, likely because larger effect values β could be more easily generated, reducing the impact of weak IVs on the results. These findings underscore that a larger Her, which could be caused by a larger $\sigma^2_{g_b}$ or a larger M_b, does not always guarantee better performance for other MR methods, as previously suggested by Freeman et al. [96]. Increasing Her through the inclusion of numerous weak IVs did not consistently improve the performance of other MR methods. Instead, the efficiency of the TS-RE method appears to benefit more when the variants collectively explain a larger proportion of the variance in the exposure, indicating its potential advantage under such circumstances.

4.3.3 IVs from $\mathbb{G}_b$ with 20% IVs having strong effects on X

Considering that other methods require some of the IVs to have strong effects, we introduced a scenario in which 20% of the IVs had strong effects generated from a normal distribution $N(0.2, 0.05^2)$, while the remaining IVs had weak effects from $N(0, 0.05^2)$. We explored different numbers of IVs, denoted as M_b, with values of 100, 500, 1000, 2000, and 5000, while setting the causal effect to $\theta = 0.3$. The residual variance parameter $\sigma^2_{e_x}$ was fixed at 2 and Her was 0.56. The results are depicted in Fig 4.5 for a scenario where all IVs were weak and Fig 4.6 when 80% of the IVs were weak. The inclusion of IVs with

strong effects led to improvements in bias for the other methods, albeit some bias still remained. Notably, the TS-RE method exhibited the least bias across all configurations and approached an unbiased estimate when the number of IVs exceeded 500.

Furthermore, incorporating some IVs with strong effects also contributed to a reduction in the SE of the TS-RE method, consistent with our proof in Eq 4.15. In Supplementary Table S2, which provides more detailed results, it can be observed that the TS-RE estimates, even when using only 20 selected IVs, still exhibited bias when IVs with strong effects were included. It should be noted that the inclusion of some IVs with strong effects did not notably enhance the performance of the other methods due to the limited sample size.

Among all the MR methods, it was observed that the dIVW method was particularly sensitive to the small sample size, displaying significant bias when all IVs were weak. This sensitivity might be attributed to dIVW's reliance on a large sample size to yield a precise estimate for $\hat{\gamma}_x$ and $se(\hat{\gamma}_x)$, which are crucial for adjusting the bias introduced by IVW.

Figure 4.5: Empirical distributions of the estimates of the causal effect $\theta = 0.3$ by the methods with different numbers of IVs and all IVs are weak. TS-RE used all IVs while other MR methods used the selected top 20 most significant IVs.

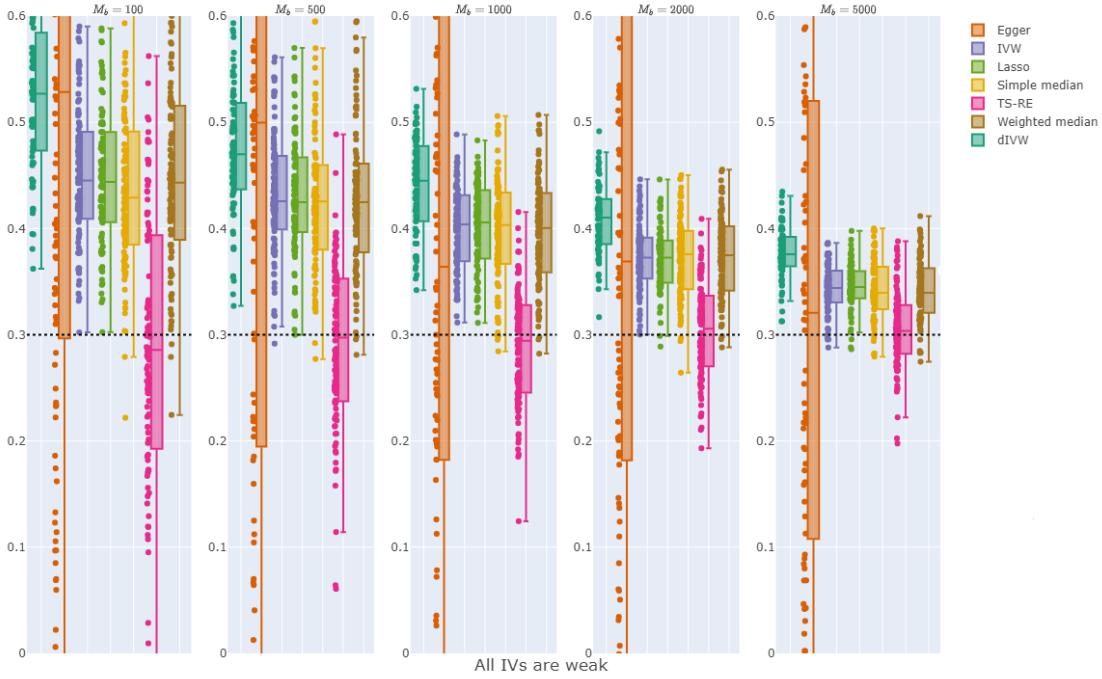
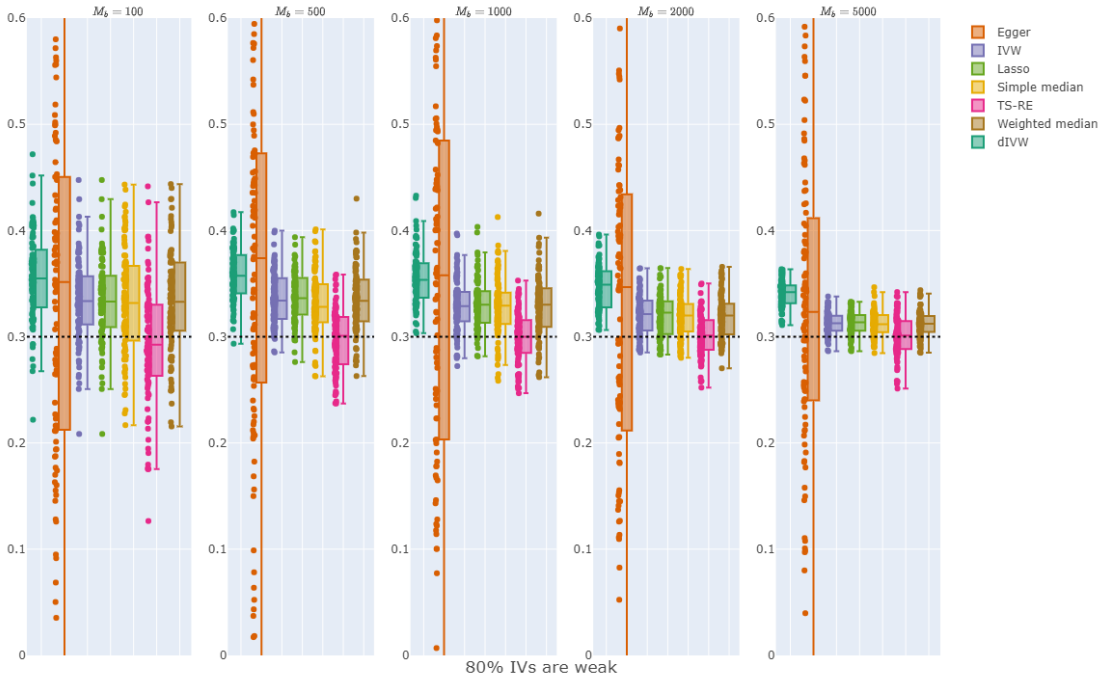


Figure 4.6: Empirical distributions of the estimates of the causal effect $\theta = 0.3$ by the methods with different numbers of IVs and 80% IVs are weak. TS-RE used all IVs while other MR methods used the selected top 20 most significant IVs.



4.3.4 IVs from $\mathbb{G}_b$ under large sample sizes

Given the necessity for large sample sizes in other MR methods, we conducted an investigation involving different sample sizes ranging from 1000 to 10000. We held certain parameters constant, specifically setting $M_b = 1000$, $\sigma_{g_b} = 0.03$, and a causal effect of $\theta = 0.3$. The residual variance parameter $\sigma^2_{e_x}$ was maintained at a fixed value of 2, and the Her remained at 0.31. In our presentation of results, Fig 4.7 portrays the outcomes when all instrumental variables (IVs) were weak, while Fig 4.8 showcases the results when 80% of the IVs were weak. Supplementary Table S3 included more detailed results.

In cases where all IVs lacked strength ($\mu_{g_b} = 0$), it was evident that enlarging the sample size led to some reduction in bias for other MR methods. Nevertheless, these methods still yielded biased estimates. In stark contrast, our TS-RE consistently produced

unbiased estimates.

Under configurations where 20% of the IVs possessed strong effects ($\mu_{g_b} = 0.2$), our TS-RE estimates remained unbiased across all sample sizes. In contrast, the other MR methods required sample sizes of $n \geq 5000$ to achieve unbiased estimates. Notably, when the top 20 significant IVs were used, Egger exhibited the largest standard error SE compared to the other methods. For the weak-IV MR method dIVW, increasing the sample size significantly reduced bias, but it still displayed the largest bias when all IVs were included. This may be attributed to the fact that the consistency of the dIVW estimator relies on a very large "effective sample size" as defined in [104], a condition not guaranteed in our simulation setting.

Figure 4.7: Empirical distributions of the estimates of the causal effect $\theta = 0.3$ by the methods with different sample sizes, all IVs effects are weak. TS-RE used all IVs while other MR methods used the selected top 20 most significant IVs.

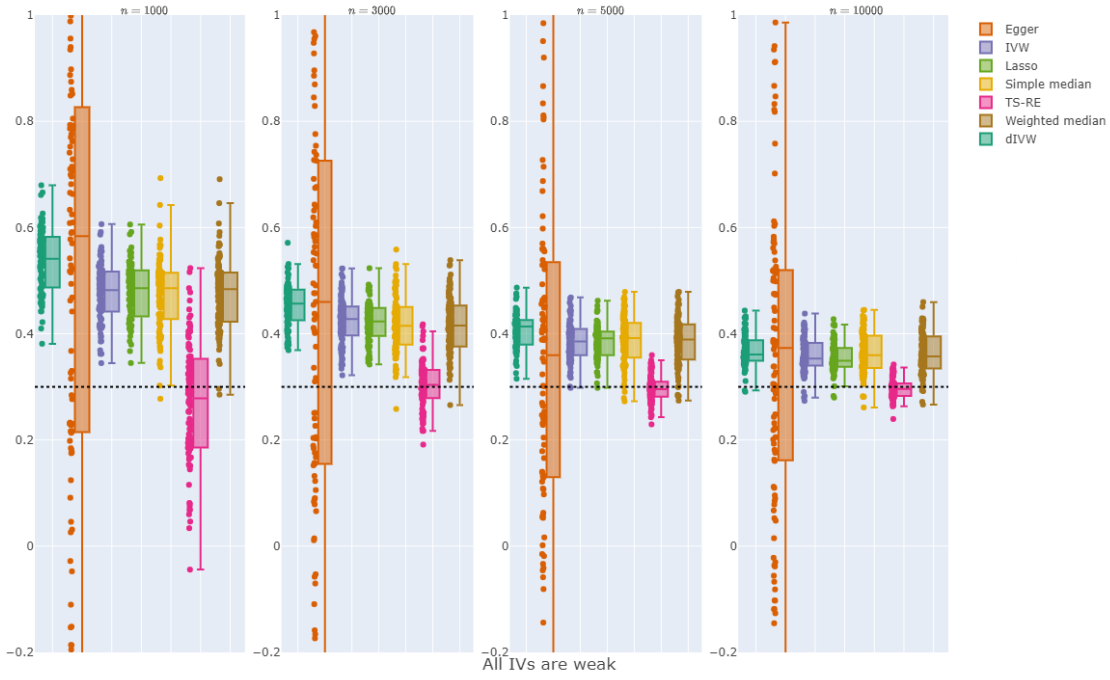
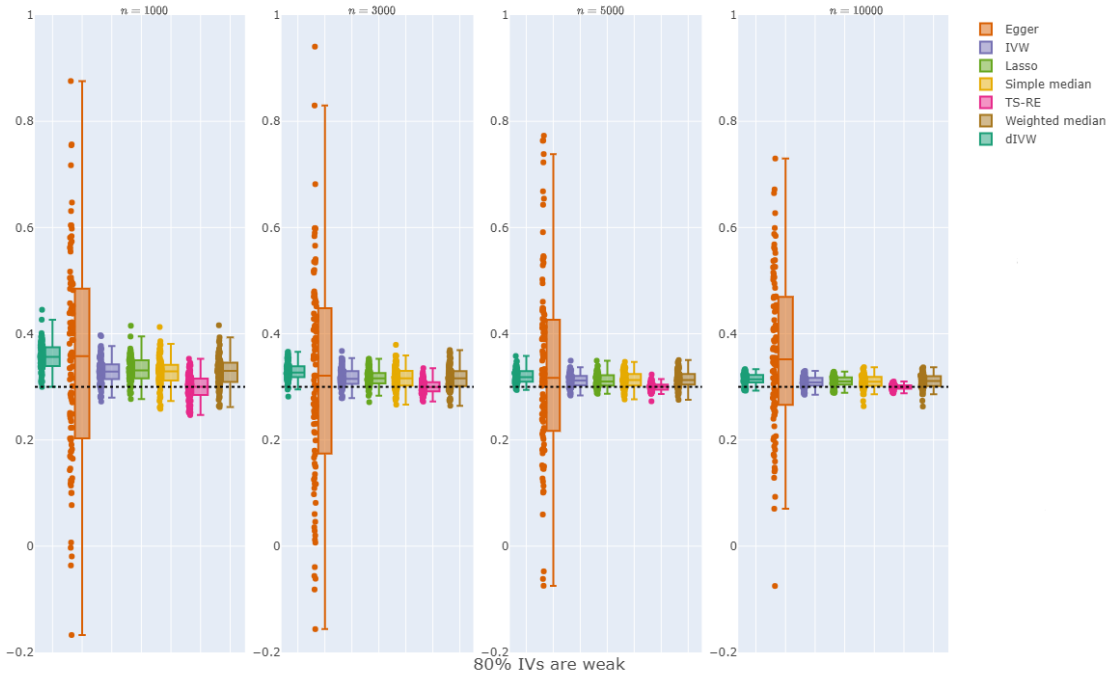


Figure 4.8: Empirical distributions of the estimates of the causal effect $\theta = 0.3$ by the methods with different sample sizes, 80% IVs effects are weak. TS-RE used all IVs while other MR methods used the selected top 20 most significant IVs.



4.3.5 Pleiotropic IVs from $\mathbb{G}_c$: large sample sizes and large effects

In this and the next two sections, the simulation model was

$$\begin{aligned} X &= \mathbb{G}_{\boldsymbol{b}}\boldsymbol{\beta_b} + \mathbb{G}_{\boldsymbol{c}}\boldsymbol{\beta_c} + e_x, \\ Y &= \theta X + \mathbb{G}_{\boldsymbol{c}}\boldsymbol{\alpha_c} + e_y. \end{aligned} \tag{4.17}$$

In our initial simulation, we aimed to evaluate the performance of the included MR methods under ideal conditions with a sample size of $n = 10000$. We deliberately chose a small number of IVs ($M_b = M_c = 100$) and set genetic variances as $\sigma^2_{g^x_b} = \sigma^2_{g^x_c} = \sigma^2_{g^y_c} = 0.03$, along with residual variances of $\sigma^2_{e_x} = \sigma^2_{e_y} = 2$. The causal effect was $\theta = 0.3$. We explored two pleiotropy scenarios: balanced ($\mu_\alpha = 0$) and directional ($\mu_\alpha = 0.1$) while ensuring that the InSIDE assumption held true ($\rho_{g_c} = 0$). We also considered three scenarios to

investigate the impact of IVs with weak effects: (1) All IVs Have Weak Effects: In this setting, we assigned weak effects to all IVs, with $\mu_{\beta_b} = \mu_{\beta_c} = 0$. (2) 20% IVs Have Strong Effects: Here, only 20% of the IVs were endowed with strong effects, with $\mu_{\beta_b} = \mu_{\beta_c} = 0.2$. (3) All IVs Have Strong Effects: In this particular case, we assumed that all IVs exhibited strong effects, with $\mu_{\beta_b} = \mu_{\beta_c} = 0.2$.

The results of this simulation are detailed in Supplementary Table S4. Notably, when all IVs had weak effects, only the TS-RE method yielded unbiased results, and its standard error (SE) was comparable to that of other MR methods. However, when large-effect IVs were introduced into the analysis, the other MR methods were also able to produce unbiased estimates.

4.3.6 Pleiotropic IVs from $\mathbb{G}_c$: directional pleiotropy effect and InSIDE assumption

Based on the results of the previous simulation, we excluded the MR-dIVW method due to its poor performance in the context of a small sample size ($n = 1000$). We checked the scenarios that all IVs were weak and 80% were weak. The causal effect was $\theta = 0.3$ and the variance parameters were the same as in the previous simulation. Additionally, we omitted the Lasso method as its reliable performance depended on additional information about how to select the tuning parameter.

Table S3 illustrates the impact of directional pleiotropy and the validity of the InSIDE assumption when $\mathbb{G}$ IVs were introduced into the model. Here are the key findings from the analysis: (Scenario 1) Balanced Pleiotropy and InSIDE Satisfied: In this scenario, where balanced pleiotropy and the InSIDE assumption were met, TS-RE consistently yielded nearly unbiased estimates, outperforming the other MR methods when all IVs were used. (Scenario 2) - Directional Pleiotropy and InSIDE Satisfied: When directional pleiotropy was introduced while still satisfying the InSIDE assumption, the performance of TS-RE declined, particularly when 20% of the M_c IVs had strong effects. This violated the unbiasedness requirement of TS-RE, which assumes that $E(\beta_c\alpha_c) = 0$. However, when all M_c IVs were weak, TS-RE's bias remained smaller than that of the other MR methods. It's worth noting that, due to the small sample size, Egger could not provide unbiased estimates, even though it was designed to handle pleiotropy. (Scenarios 3 and 4) Invalid InSIDE

Assumption: In these scenarios, the InSIDE assumption was violated ($Var(\beta_c, \alpha_c) \neq 0$), thus $E(\beta_c\alpha_c) \neq 0$, which is a key assumption for TS-RE. None of the methods, including TS-RE, were able to provide unbiased estimates under this condition.

In summary, the performance of TS-RE was generally robust when the InSIDE assumption was satisfied and pleiotropy was balanced. However, it was sensitive to directional pleiotropy when strong IV effects were introduced. Violations of the InSIDE assumption led to biases in all methods, including TS-RE.

	p_w	SM	WM	IVW	Egger	TS-RE
BP,	0.8	0.32 (0.06)	0.33 (0.07)	0.33 (0.05)	0.40 (0.30)	0.29 (0.05)
InSIDE	1	0.43 (0.11)	0.44 (0.11)	0.43 (0.09)	0.42 (0.61)	0.29 (0.16)
DP,	0.8	0.70 (0.07)	0.70 (0.07)	0.70 (0.06)	0.63 (0.34)	0.68 (0.07)
InSIDE	1	0.42 (0.16)	0.41 (0.17)	0.42 (0.15)	0.32 (1.00)	0.27 (0.24)
BP,	0.8	0.76 (0.12)	0.78 (0.12)	0.77 (0.11)	0.82 (0.48)	0.73 (0.09)
No InSIDE	1	0.81 (0.13)	0.80 (0.13)	0.79 (0.12)	0.74 (0.49)	0.78 (0.10)
DP,	0.8	0.80 (0.12)	0.83 (0.13)	0.84 (0.11)	0.93 (0.51)	0.79 (0.09)
No InSIDE	1	0.89 (0.14)	0.89 (0.14)	0.89 (0.12)	0.87 (0.67)	0.87 (0.10)

Table 4.2: Mean and SE of different methods for $M_b = 100, M_c = 100$ IVs: under different pleiotropic effect and InSIDE assumption conditions. The directional effect is 0.1 and the genetic correlation is 0.6 when the InSIDE assumption is invalid. The strong effect for some valid M_b IVs is 0.2.

4.3.7 Pleiotropic IVs from $\mathbb{G}_c$: different proportion of valid IVs

In our investigation, we explored the influence of the proportion of valid IVs on the estimates while maintaining a fixed sample size of $n = 1000$ and a total of $M = 1000$ IVs. We varied the number of valid IVs, denoted as M_b, ranging from 0 to 1000. The causal effect was set to be $\theta = 0.1, 0.3$. The genetic variances were set to $\sigma_{g_b^x} = \sigma_{g_c^x} = \sigma_{g_c^y} = 0.03$, and residual variances were $\sigma^2_{e_x} = \sigma^2_{e_y} = 2$. For both balanced pleiotropy ($E(\alpha_c) = 0$) and directional pleiotropy ($E(\alpha_c) = 0.1$), we ensured that the InSIDE assumption remained valid. All IVs were considered weak ($E(\beta_b) = E(\beta_c) = 0$), and the heritability was set to $Her = 0.31$. In Fig 4.9, solid lines represent bias, while dashed lines represent SE for each method.

When all IVs were used, TS-RE consistently exhibited much smaller biases compared

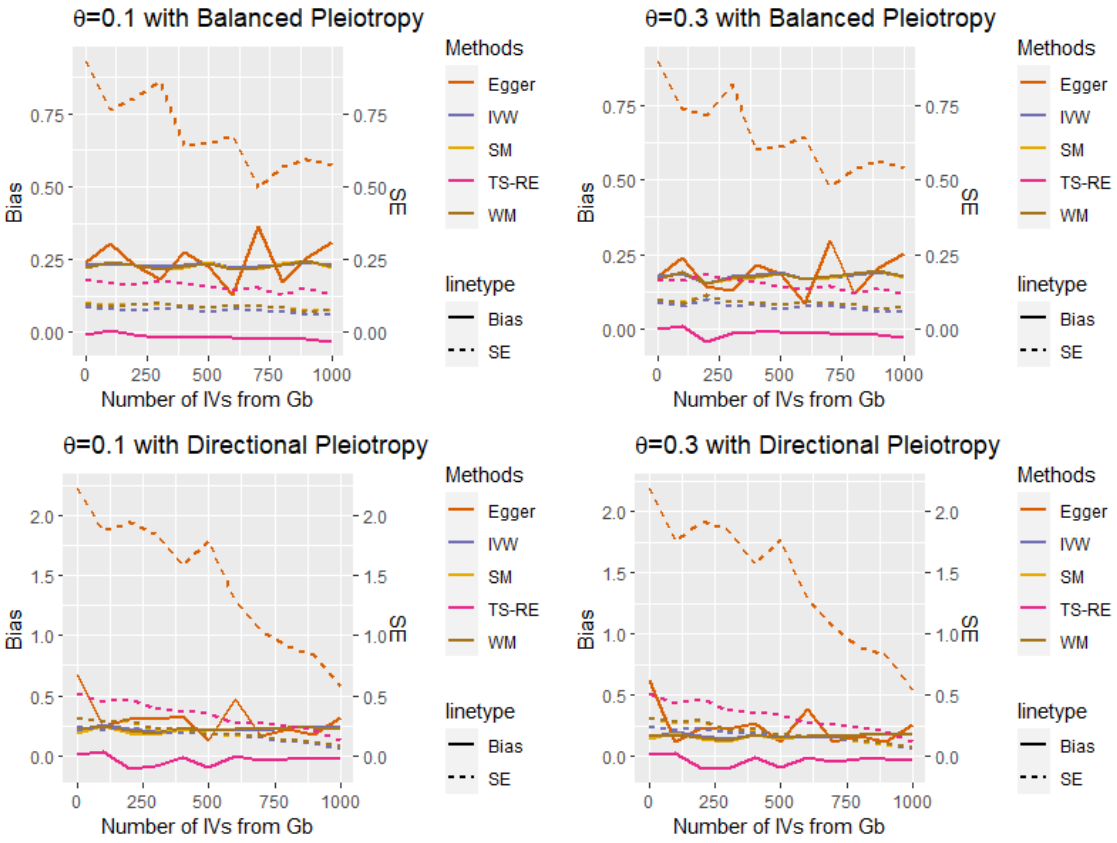
to other methods, regardless of whether pleiotropy was balanced or directional. Selecting the top 20 significant IVs substantially increased the bias and SE for the TS-RE method.

Effect of Proportion of Invalid IVs on SE with Directional Pleiotropy: Under directional pleiotropy ($E(\alpha_c) = 0.1$), a higher proportion of invalid IVs resulted in higher SE for the TS-RE estimator. This was primarily due to the increased contribution of IVs from $\mathbb{G}_c$, which significantly increased the value of the term $M_c E(\alpha_k^2) = M_c E(\alpha_c)^2 + M_c Var(\alpha_c)$ in the numerator of τ^2. Notably, TS-RE exhibited a larger bias when a moderate proportion of IVs were valid (around $30 - 50\%$).

Effect of Proportion of Invalid IVs on SE with Balanced Pleiotropy: In contrast, under balanced pleiotropy ($E(\alpha_c) = 0$), the impact of the proportion of valid IVs on SE was considerably reduced compared to the scenario with directional pleiotropy. This was because, for balanced pleiotropy, the value $M_c E(\alpha_k^2) = 0.03^2 M_c$ was much smaller than the value for directional pleiotropy, where $M_c E(\alpha_k^2) = (0.1^2 + 0.03^2) M_c$.

In summary, TS-RE consistently exhibited lower bias than other methods when all IVs were used, regardless of the type of pleiotropy. However, the choice to select only the top 20 significant IVs for TS-RE significantly increased bias and SE. The impact of the proportion of invalid IVs on SE was more pronounced under directional pleiotropy compared to balanced pleiotropy, where the effect on SE was mitigated.

Figure 4.9: Bias and standard error (SE) of the estimates the causal effect $\theta = 0.1, 0.3$, the solid lines are biases and dashed lines are SEs. For the balanced pleiotropy $E(\alpha_c) = 0$ and for the directional pleiotropy $E(\alpha_c) = 0.1$, the InSIDE assumption is valid. Here total number of IVs is $M = 1000$, the sample size is $n = 1000$, $E(\beta_b) = E(\beta_c) = 0$ and $\sigma_{\mathbb{G}_b} = \sigma_{\mathbb{G}_c} = 0.03$, $Her = 0.31$. TS-RE used all IVs while other MR methods used the selected top 20 most significant IVs.



In Supplementary Tables S5 and S6, we calculated various performance metrics, including Bias, SE, and Mean Squared Error (MSE), for all methods. Notably, our TS-RE method consistently had the smallest MSE among all methods for scenarios involving balanced pleiotropy when all IVs were utilized. This outcome suggests that when the condition $E(\alpha_c\beta_c) = 0$ holds, our method performed well. However, under directional pleiotropy, the MSE for TS-SE was larger than that of other methods due to its larger SE. For the other

four methods, selecting only the top 20 significant IVs did not lead to improvements in bias because the sample size was too small. Although methods like Egger, simple median, and weighted median were designed to handle IVs from $\mathbb{G}_c$ with pleiotropic effects, the limited sample size in this scenario constrained their performance.

While the SEs for other methods were smaller than that of our TS-RE, the biases of our TS-RE were consistently much smaller than those of the other methods across all scenarios. This suggests that, in these particular conditions, TS-RE offers a favorable trade-off between bias and precision, resulting in superior overall performance.

4.3.8 Null IVs from group $\mathbb{G}_a$ and $\mathbb{G}_d$

In this simulation, we included four groups of IVs, the simulation model was

$$\begin{aligned} X &= \mathbb{G}_{\boldsymbol{b}}\boldsymbol{\beta_b} + \mathbb{G}_{\boldsymbol{c}}\boldsymbol{\beta_c} + e_x, \\ Y &= \theta X + \mathbb{G}_{\boldsymbol{c}}\boldsymbol{\alpha_c} + \mathbb{G}_{\boldsymbol{d}}\boldsymbol{\alpha_d} + e_y. \end{aligned} \tag{4.18}$$

Weak IV effects β_b, β_con X were generated $N(0, 0.03^2)$, while strong effects were generated $N(0.2, 0.03^2)$. Effects α_c, α_d on Y were generated $N(0, 0.03^2)$. The causal effect was $\theta = 0.3$, and the residual variances were $\sigma^2_{e_x} = \sigma^2_{e_y} = 2$. The IVs from group $\mathbb{G}$ satisfied both the balanced pleiotropic assumption and the InSIDE assumption. The number of IVs from each group was set to 100, 200, and 500, while the total number of IVs was 400, 800, and 2000. Our methods consistently provided more unbiased estimates than other methods when all IVs were included. The complete results are available in Supplementary Table S7.

For our TS-RE method, the SE was larger than that of the other four methods, while the Mean Squared Error (MSE) was similar to other methods. However, the results for our TS-RE method, as presented in Table 4.3, revealed that including too many null IVs from $\mathbb{G}_a$ which had no effects on X resulted in increased bias for TS-RE. We increased the number of null IVs in group $\mathbb{G}_a$ from 1000 to 50000 while maintaining the number of IVs in the other three groups at 1000. The results showed a significant increase in SE as the number of null IVs increased, particularly considering that the sample size was much smaller than the number of IVs. This observation aligns with the theoretical result, as the

asymptotic variance $\tau^2 \propto \frac{M}{M_b+M_c}$.

M_a	Estimate	SE
1000	0.3047400	0.09359366
2000	0.2962185	0.12384585
5000	0.2859627	0.26879603
10000	0.2915771	0.21386340
20000	0.2859134	0.38467096
50000	0.1124109	2.40895732

Table 4.3: Change the number of null IVs where the number of each other three groups is fixed to be 1000.

4.4 Real data analysis: causal effect of BMI on SBP

We estimated the causal effect of BMI on systolic blood pressure(SBP) for black British individuals from the UK-biobank dataset, with a total sample size of 3396. Following quality control procedures and linkage disequilibrium pruning, we retained 151442 SNPs with an allele frequency of 0.05, HWE of 0.000001, MAF of 0.01, and LD window of 1000, the step of 50, and r^2 of 0.1. Our sample size was 2802 after removing the related individuals with GRM cutoff ≥ 0.05. We then applied our TS-RE method and four other MR methods to estimate the causal effect of BMI on SBP.

Initially, TS-RE using all SNPs without selection yielded an imprecise result with a large standard error: $\theta = 0.31, SE = 0.41$. Here $\theta = 0.31$ means a one kg/m^2unit increase in BMI increases SBP by $0.31mmHg$. This is consistent with previous studies using MR methods with large samples [110].

SM	WM	IVW	Egger
-0.195 (0.871)	0.748 (0.819)	-0.093 (0.639)	0.949 (0.892)

Table 4.4: Causal effect of BMI on SBP for independent black British with selected 56 SNPs based on an external study.

Subsequently, we selected the top 20 significant SNPs for BMI, resulting in nonsensical results for all methods, as shown in Table 4.5. To avoid including an extremely large number of null IVs in $\mathbb{G}_a$ and $\mathbb{G}_d$, we selected 7580 SNPs with a $p_{value} < 0.005$, which reduced the SE of the TS-RE to 0.01 and yielded an estimate of $\theta = 0.33$. In Table4.4, we attempted to use 56 significant SNPs as IVs for this small study, where those SNPs were identified by the previous study focused on the white population [111]. Other MR methods failed to estimate the causal effect (large SE and $p-value > 0.05$) due to the small sample size.

TS-RE	SM	WM	IVW	Egger
0.31 (0.41)	2.42 (0.25)	2.42 (0.25)	2.27 (0.17)	2.63 (0.42)

Table 4.5: Causal effect of BMI on SBP for independent black British: TS-RE used all SNPs and the other MR methods used the selected top 20 significant SNPs

4.5 Discussion

The proposed TS-RE estimator offers a promising solution to address the challenges faced in MR analyses, particularly in small-scale studies. These challenges often include issues related to weak IVs and pleiotropic effects, which traditionally require a large sample size to effectively detect and estimate. In contrast to existing MR approaches that primarily rely on first-order moments for estimating causal effects, our TS-RE method leverages second-order moments to estimate the causal effect of the exposure variable on the outcome variable. By utilizing a substantial number of genetic variants, TS-RE enables the estimation of genetic variances for both the exposure and the outcome variables, ultimately providing an unbiased estimate of the causal effect. Remarkably, this can be achieved even with a small sample size or when dealing with a minority population.

In practice, GWAS often generate data with a large number of SNPs through advanced DNA sequencing techniques, regardless of the sample size. Our novel TS-RE method capitalizes on this wealth of genetic information, allowing for the application of MR to estimate causal effects in small studies or within specific sub-populations of interest within a larger dataset. Furthermore, real data analysis has demonstrated the limitations of

conventional MR methods, even when utilizing significant SNPs identified by large external studies. This is particularly evident when attempting to rectify the issue of small sample sizes, primarily due to the heterogeneity between the smaller, specific population of interest and the larger external population. TS-RE offers a robust alternative in such scenarios, where commonly used MR methods relying on large sample sizes may not be feasible or effective.

The TS-RE method offers several key advantages supported by both theoretical insights and empirical simulations. These advantages make it a valuable tool for MR analyses. First, TS-RE does not require strict selection criteria for IVs, such as a specific p-value threshold (e.g., p-value $\leq 5 \times 10^{-8}$). It can handle weak IVs without significantly amplifying biases resulting from the violation of the exclusion restriction. This means that even including some null IVs is acceptable for TS-RE, providing greater flexibility in IV selection. Second, unlike many other MR methods that require a very large sample size for consistency, TS-RE achieves consistency even with a small sample size by incorporating a large number of IVs. This sets it apart from first-moment-based MR methods that depend on large sample sizes for reliable estimates. TS-RE excels in small studies and performs comparably to other MR methods in larger studies. Third, theoretical analysis shows that TS-RE is equivalent to the IVW method when all IVs have a direct effect on the exposure variable. After obtaining the Genetic Relationship Matrix (GRM), the application of TS-RE is as straightforward as the IVW estimator, simplifying the estimation process. Although the TS-RE prefers a large number IVs to estimate the genetic variance and covariance, as we showed in the simulation study, the performance of our TS-RE was not worse than other methods using the selected top 20 significant IVs.

We acknowledge the certain limitations of TS-RE. First, TS-RE tends to have larger standard errors compared to other MR methods, primarily because it employs a second-moment estimator, which introduces more uncertainty. Despite this, TS-RE significantly reduces bias and achieves similar MSE compared to other MR methods. This makes it a valuable alternative for obtaining more unbiased estimates in practical settings. In addition, while TS-RE relaxes the strict exclusion restriction, it still assumes that the expectation of the product of pleiotropic effects is zero, as specified by conditions like InSIDE. This assumption may have a limited biological basis, as it restricts unknown

pleiotropic effects of SNPs. However, when a large number of weak IVs are included, the average of the product approximates zero. Future developments of TS-RE may explore ways to further relax both the independence and exclusion restrictions, potentially aligning with approaches like MR-Genius [112].

In summary, TS-RE presents a robust and flexible approach to MR analyses, particularly suited for small studies, weak IVs, and scenarios where other MR methods may struggle. Its theoretical foundation and performance in simulations demonstrate its potential as a valuable tool for causal inference in a variety of research contexts. Future work may focus on refining and extending TS-RE to address additional challenges and broaden its applicability in MR analyses. The proposed framework for small studies can also be extended to integrate information across multiple studies [102, 113]. Using a meta-analysis to combine the estimates of genotype-phenotype association from different studies can give more precise estimates of the IVs effect. Similarly, the current analysis of a small sub-population can be extended to multiple sub-populations to investigate the causal effect in the presence of population sub-structure [114].

Chapter 5

Conclusion and Discussion

5.1 Summary of current findings

For the normative data, our proposed methods for in Chapter 2 and 3 extended the application of using meta-analysis for estimating reference ranges. Current methods proposed by Siegel st al. [2] are based on a random effects model with normal distribution and equal within-study variance assumptions. We showed that the fixed effects model-based method in Chapter 2 can be an alternative when the number of available studies is small and the normal assumption is improper. The Bayesian nonparametric methods in Chapter 3 used DP instead of a parametric (e.g. normal) distribution under a random effects model. With such a flexible distribution assumption, the random effects model, which is the most commonly used model in meta-analysis, can be applied to a wider range of studies. The simulation studies in Chapter 3 showed that the mixture method and DP method are similar since they both make a "mixture" distribution assumption, while the number of mixture components in the DP method is unknown. When the number of included studies is large enough ($N = 10$), the coverage and the estimated 95% reference ranges with different methods were similar. None of those methods showed a general advantage over other methods, thus the recommendation is carefully making assumptions and choosing the appropriate method based on the scientific question and data type. We can provide the following guidance for choosing the appropriate method: (1) when the number of studies is small and it is hard to examine the random effects model assumptions, the mixture

distribution method is recommended (fixed effects meta-analysis model-based); (2) when the number of included studies is large and there is evidence showing the study means are non-normal distributed, the NP methods will be a good choice; (3) When choosing NP and NP-2, the decision about whether the within-study variances should be assumed equal or not will not have a big impact on the estimated reference intervals, but the goodness of fit will be different.

For Mendelian randomization, our proposed method provides a completely new perspective for estimating the causality with genetic data. Current existing methods use the means of the effects of IVs on the exposure and outcome variable to construct a ratio estimator and a large sample size N is required to get a consistent estimate of the causal effect. However, our method uses the variance of the effects of IVs, which can be regarded as the heritability of the exposure and outcome variables, to build up the ratio estimator. This method addresses the weak effect problem and the small sample size problem faced by many other MR methods. The proof in Chapter 4 shows that this new method actually is a more generative method of MR-IVW. MR-IVW has been criticized for being too conservative that it assumes that all IVs are without pleiotropy or with balanced pleiotropy. Our method allowed the inclusion of IVs with directional pleiotropic effects and weak IVs. Particularly, our TS-RE method is the only method that can handle the weak and pleiotropic IVs under a small sample size, while existing methods always require a large sample size to conduct the selection or bias reduction.

5.2 Future work

The findings described in the previous section lead to many opportunities for future work. The first part will discuss the extension of the methods used in Chapter 2 and 3 for meta-analysis. The remaining part of this section stems from the work on Mendelian randomization in Chapter 4.

5.2.1 Borrowing information from large external data.

Meta-analytic methods in Chapter 2 and 3 can be used to combine evidence from different studies to estimate a reference range for the overall population. However, even a meta-analysis that combines multiple studies can still have small sample issues. For example, in the real data analysis in Chapter 3 for pediatric sleeping time, most of the included studies have very small sample sizes (less than 50). The information for those small studies can lead to imprecise reference intervals and inaccurate conclusions for a subject's measurement. Such pediatric studies often face challenges, including economic, logistical, technical, and ethical barriers, in collecting sufficient data. The same modeling framework in the meta-analysis may also be used to not only derive a combined estimate but also to borrow information for a particular study from another [115]. A better alternative is partially borrowing useful information from external data for the primary meta-analysis.

Established methods focus on borrowing from external data for a single study, it will be attractive to conduct Bayesian borrowing for a meta-analysis (BB-MA), to borrow external data for a meta-analysis including multiple studies. Particularly, we considered the internal studies to have individual-level participant data that included the covariates information. The challenges under this borrowing framework are determining: (1) which part of the external data information should be borrowed; and (2) how much the external information should be borrowed. For the first question, the external study might not contain all the covariates in the internal studies, which can be regarded as a reduced model. Thus, we first need to align the external summary statistics with the parameters for the internal study covariates, e.g., by using the method proposed by Taylor et. al. [116] Then, the next step is using the aligned external data information to construct a Bayesian prior for the internal data and determining how much to borrow. One way to do it is to estimate the commensurability of the external data and internal data [117]. Then the commensurability will be used as a weight for mixing the external data information with a non-informative distribution to get a commensurate prior distribution. The commensurate prior distribution will be integrated with the internal data for estimating the posterior distribution of interested parameters. Under a Bayesian framework, the prior information on the pediatric data is derived from the adult data. Several Bayesian methods for borrowing information to construct the prior have been proposed in recent years [118, 119].

In addition, DP used in Chapter 3 to relax the distribution assumptions in the Bayesian framework can also be extended for borrowing information across data from different sources, including mixtures of DPs, dependent DPs, hierarchical DPs, and nested DPs [57, 120–122]. We will use those methods to get the prior distribution when building the Bayesian framework for estimating the reference range with the normative pediatric data.

5.2.2 Mendelian Randomization with correlated pleiotropic effects

For the methodology of Mendelian randomization, we have developed the TS-RE allowing weak IVs and pleiotropic IVs with a small sample size. However, the InSIDE assumption is still required for our method as well as all other MR methods. A violation of the InSIDE assumption (correlated pleiotropic effect) can cause huge bias, but this phenomenon that the effects of a gene variant on different phenotypes are related is very common in genetic research. To address this problem, we will adjust the current TS-RE model by adding a new group of IVs with correlated pleiotropic effects. Then, the next step is to estimate this coefficient of correlation. Recently, there are Bayesian methods have been developed for MR that incorporate a prior distribution for this correlation pleiotropy parameter. [123,124] We can use a similar technique by putting our proposed TS-RE into a Bayesian framework and relaxing the current assumption.

5.2.3 Multivariate Mendelian Randomization

A genetic variant may be associated with multiple exposure variables so long as any association with the outcome is via the measured exposure variables. In addition, those multiple exposure variables can also correlate with each other. An univariate model may result in biased causal estimates and inappropriate inferences [125, 126]. Including more exposure variables can also address the problem of unmeasured confounders since they can be regarded as measured confounders. In addition, the correlated pleiotropic effect from one exposure factor to the outcome variable might be caused by other exposure variables. Thus, extending the current method to multivariate cases will multiple exposure variables is attractive, [127] especially for some complex exposure variables such as brain image data since different brain surfaces are connected and thus correlated. We will also investigate the robustness of the model by including mediators, colliders, and reverse causation, which

are very common in causal inference.

References

[1] Paul S Horn and Amadeo J Pesce. A robust approach to reference interval estimation and evaluation. *Clin Chim Acta*, 334(1-2):5–23, 2003.

[2] Lianne Siegel, M Hassan Murad, and Haitao Chu. Estimating the Reference Range from a Meta-Analysis. *Res Synth Methods*, 12(2):148–160, 2021.

[3] Paul S Horn, Amadeo J Pesce, and Bradley E Copeland. A robust approach to reference interval estimation and evaluation. *Clin Chem*, 44(3):622–631, 1998.

[4] Kenneth Rice, Julian PT Higgins, and Thomas Lumley. A re-evaluation of fixed effect(s) meta-analysis. *J R Stat Soc Ser A Stat Soc*, 181(1):205–227, 2018.

[5] Julian PT Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc*, 172(1):137–159, 2009.

[6] Anna-Bettina Haidich. Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29–37, 2010.

[7] Faraz Pathan, Nicholas D'Elia, Mark T Nolan, Thomas H Marwick, and Kazuaki Negishi. Normal ranges of left atrial strain by speckle-tracking echocardiography: a systematic review and meta-analysis. *J Am Soc Echocardiogr*, 30(1):59–70, 2017.

[8] Philip T Levy, Aliza Machefsky, Aura A Sanchez, Meghna D Patel, Sarah Rogal, Susan Fowler, Lauren Yaeger, Angela Hardi, Mark R Holland, Aaron Hamvas, et al. Reference ranges of left ventricular strain measures by two-dimensional speckle-tracking echocardiography in children: a systematic review and meta-analysis. *J Am Soc Echocardiogr*, 29(3):209–225, 2016.

[9] Allison A Venner, Patricia K Doyle-Baker, Martha E Lyon, and Tak S Fung. A meta-analysis of leptin reference ranges in the healthy paediatric prepubertal population. *Ann Clin Biochem*, 46(1):65–72, 2009.

[10] Jean F Wyman, Jincheng Zhou, D Yvette LaCoursiere, Alayne D Markland, Elizabeth R Mueller, Laura Simon, Ann Stapleton, Carolyn RT Stoll, Haitao Chu, and Siobhan Sutcliffe. Normative noninvasive bladder function measurements in healthy women: a systematic review and meta-analysis. *Neurourol Urodyn*, 39(2):507–522, 2020.

[11] Ali Reza Khoshdel, Ammarin Thakkinstian, Shane L Carney, and John Attia. Estimation of an age-specific reference interval for pulse wave velocity: a meta-analysis. *J Hypertens*, 24(7):1231–1237, 2006.

[12] Fateh Bazerbachi, Samir Haffar, Zhen Wang, Joaquín Cabezas, Maria Teresa Arias-Loste, Javier Crespo, Sarwa Darwish-Murad, M Arfan Ikram, John K Olynyk, Eng Gan, et al. Range of normal liver stiffness and factors associated with increased stiffness measurements in apparently healthy individuals. *Clin Gastroenterol Hepatol*, 17(1):54–64, 2019.

[13] Barbara C Galland, Michelle A Short, Philip Terrill, Gabrielle Rigney, Jillian J Haszard, Scott Coussens, Mistral Foster-Owens, and Sarah N Biggs. Establishing normal values for pediatric nighttime sleep measured by actigraphy: a systematic review and meta-analysis. *Sleep*, 41(4):zsy017, 2018.

[14] PDA Benfica, Larissa Tavares Aguiar, SAF Brito, Luane Helena Nunes Bernardino, Luci Fuscaldi Teixeira-Salmela, and CDCM Faria. Reference values for muscle strength: a systematic review with a descriptive meta-analysis. *Braz J Phys Ther*, 22(5):355–369, 2018.

[15] Laila B Conceição, Jussara AO Baggio, Suleimy C Mazin, Dylan J Edwards, and Taiza EG Santos. Normative data for human postural vertical: a systematic review and meta-analysis. *PLoS One*, 13(9):e0204122, 2018.

[16] Jan A Staessen, Robert H Fagard, Paul J Lijnen, Lutgarde Thijs, Roger Van Hoof, and Antoon K Amery. Mean and range of the ambulatory pressure in normotensive subjects from a meta-analysis of 23 studies. *Am J Cardiol*, 67(8):723–727, 1991.

[17] Joanna IntHout, John PA Ioannidis, Maroeska M Rovers, and Jelle J Goeman. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, 6(7):e010247, 2016.

[18] Arnaud Chiolero, Valérie Santschi, Bernard Burnand, Robert W Platt, and Gilles Paradis. Meta-analyses: with confidence or prediction intervals? *Eur J Epidemiol*, 27(10):823–825, 2012.

[19] Dan Jackson and Ian R White. When should meta-analysis avoid making hidden normality assumptions? *Biom J*, 60(6):1040–1058, 2018.

[20] Nan M Laird and Frederick Mosteller. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care*, 6(1):5–30, 1990.

[21] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*, 1(2):97–111, 2010.

[22] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Control Clin Trials*, 7(3):177–188, 1986.

[23] Dankmar Böhning, Uwe Malzahn, Ekkehart Dietz, Peter Schlattmann, Chukiat Viwatwongkasem, and Annibale Biggeri. Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. *Biostatistics*, 3(4):445–457, 2002.

[24] John E Cornell, Cynthia D Mulrow, Russell Localio, Catharine B Stack, Anne R Meibohm, Eliseo Guallar, and Steven N Goodman. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med*, 160(4):267–270, 2014.

[25] Larry V Hedges and Jack L Vevea. Fixed- and random-effects models in meta-analysis. *Psychol Methods*, 3(4):486–504, 1998.

[26] John E Hunter and Frank L Schmidt. *Methods of meta-analysis: Correcting error and bias in research findings*. SAGE Publications, Ltd, 55 City Road, London, 3rd edition, 2015.

[27] Fulgencio Marín-Martínez and Julio Sánchez-Meca. Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educ Psychol Meas*, 70(1):56–73, 2010.

[28] Hui Quan and Ji Zhang. Estimate of standard deviation for a log-transformed variable using arithmetic means and standard deviations. *Stat Med*, 22(17):2723–2736, 2003.

[29] Julien Barra, Adélaïde Marquer, Roxane Joassin, Céline Reymond, Liliane Metge, Valérie Chauvineau, and Dominic Pérennou. Humans use internal models to construct and update a sense of verticality. *Brain*, 133(12):3552–3563, 2010.

[30] DA Pérennou, G Mazibrada, V Chauvineau, R Greenwood, J Rothwell, MA Gresty, and AM Bronstein. Lateropulsion, pushing and verticality perception in hemisphere stroke: a causal relationship? *Brain*, 131(9):2401–2413, 2008.

[31] Theo Stijnen, Taye H Hamza, and Pinar Özdemir. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med*, 29(29):3046–3067, 2010.

[32] Sean McGrath, XiaoFei Zhao, Russell Steele, Brett D Thombs, Andrea Benedetti, and DEPRESSion Screening Data (DEPRESSD) Collaboration. Estimating the sample mean and standard deviation from commonly reported quantiles in meta-analysis. *Stat Methods Med Res*, 29(9):2520–2537, 2020.

[33] Dehui Luo, Xiang Wan, Jiming Liu, and Tiejun Tong. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. *Stat Methods Med Res*, 27(6):1785–1805, 2018.

[34] Xiang Wan, Wenqian Wang, Jiming Liu, and Tiejun Tong. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol*, 14(1):1–13, 2014.

[35] J. Shi, D. Luo, H. Weng, X. T. Zeng, L. Lin, H. Chu, and T. Tong. Optimally estimating the sample standard deviation from the five-number summary. *Res Synth Methods*, 11(5):641–654, 2020.

[36] Yves Dominicy and David Veredas. The method of simulated quantiles. *J Econom*, 172(2):235–247, 2013.

[37] Nikolaos Sgouropoulos, Qiwei Yao, and Claudia Yastremiz. Matching a distribution by matching quantiles estimation. *J Am Stat Assoc*, 110(510):742–759, 2015.

[38] Lifeng Lin, Haitao Chu, Mohammad Hassan Murad, Chuan Hong, Zhiyong Qu, Stephen R Cole, and Yong Chen. Empirical comparison of publication bias tests in meta-analysis. *J Gen Intern Med*, 33(8):1260–1267, 2018.

[39] Lifeng Lin and Haitao Chu. Quantifying publication bias in meta-analysis. *Biometrics*, 74(3):785–794, 2018.

[40] L Lin, L Shi, Haitao Chu, and Mohammad Hassan Murad. The magnitude of small-study effects in the Cochrane Database of Systematic Reviews: an empirical study of nearly 30,000 meta-analyses. *BMJ Evid Based Med*, 25(1):27–32, 2020.

[41] Eileen M Wright and Patrick Royston. Calculating reference intervals for laboratory measurements. *Statistical Methods in Medical Research*, 8(2):93–112, 1999.

[42] Lianne Siegel, M Hassan Murad, Richard D Riley, Fateh Bazerbachi, Zhen Wang, and Haitao Chu. A guide to estimating the reference range from a meta-analysis using aggregate or individual participant data. *American journal of epidemiology*, 191(5):948–956, 2022.

[43] Wenhao Cao, Lianne Siegel, Jincheng Zhou, Motao Zhu, Tiejun Tong, Yong Chen, and Haitao Chu. Estimating the reference interval from a fixed effects meta-analysis. *Research synthesis methods*, 12(5):630–640, 2021.

[44] Mariel M Finucane, Christopher J Paciorek, Gretchen A Stevens, and Majid Ezzati. Semiparametric bayesian density estimation with disparate data sources: a meta-analysis of global childhood undernutrition. *Journal of the American Statistical Association*, 110(511):889–901, 2015.

[45] Dan Jackson and Ian R White. When should meta-analysis avoid making hidden normality assumptions? *Biom J*, 60(6):1040–1058, 2018.

[46] Roser Bono, María J Blanca, Jaume Arnau, and Juana Gómez-Benito. Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in psychology*, 8:1602, 2017.

[47] Michael D Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

[48] Ken P Kleinman and Joseph G Ibrahim. A semiparametric bayesian approach to the random effects model. *Biometrics*, pages 921–938, 1998.

[49] Deborah Burr and Hani Doss. A bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, 100(469):242–251, 2005.

[50] Minjung Kyung, Jeff Gill, and George Casella. Estimation in dirichlet random effects models. *The Annals of Statistics*, 38(2):979–1009, 2010.

[51] Saman Muthukumarana and Ram C Tiwari. Meta-analysis using dirichlet process. *Statistical Methods in Medical Research*, 25(1):352–365, 2016.

[52] Christopher A Bush and Steven N MacEachern. A semiparametric bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

[53] Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

[54] Abel Rodriguez, David B Dunson, and Jack Taylor. Bayesian hierarchically weighted finite mixture models for samples of distributions. *Biostatistics*, 10(1):155–171, 2009.

[55] Jessica Gurevitch and Larry V Hedges. Statistical issues in ecological meta-analyses. *Ecology*, 80(4):1142–1149, 1999.

[56] Hisashi Noma, Kengo Nagashima, Shogo Kato, Satoshi Teramukai, and Toshi A. Furukawa. Meta-analysis using flexible random-effects distribution models. *Journal of Epidemiology*, 32(10):441–448, 2022.

[57] David B Dunson. Nonparametric bayes applications to biostatistics. *Bayesian nonparametrics*, 28:223–273, 2010.

[58] Katherine J Lee and Simon G Thompson. Flexible parametric models for random-effects distributions. *Statistics in medicine*, 27(3):418–434, 2008.

[59] Chia-Chun Wang and Wen-Chung Lee. Evaluation of the Normality Assumption in Meta-Analyses. *American Journal of Epidemiology*, 189(3):235–242, 11 2019.

[60] Adam J. Branscum and Timothy E. Hanson. Bayesian nonparametric meta-analysis using polya tree mixture models. *Biometrics*, 64(3):825–833, 2008.

[61] M West. On scale mixtures of normal distributions. *Biometrika*, 74:646–648, 1987.

[62] Ralph B D'Agostino. Transformation to normality of the null distribution of g_1. *Biometrika*, 57:679–681, 1970.

[63] Zhongxue Chen, Guoyi Zhang, and Jing Li. Goodness-of-fit test for meta-analysis. *Scientific reports*, 5(1):16983, 2015.

[64] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[65] David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[66] Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[67] Perry de Valpine, Daniel Turek, Christopher Paciorek, Cliff Anderson-Bergman, Duncan Temple Lang, and Ras Bodik. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–413, 2017.

[68] Martyn Plummer. rjags: Bayesian graphical models using mcmc. 2022. R package version 4-13.

[69] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[70] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[71] John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

[72] Sumio Watanabe and Manfred Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

[73] Wei Liu, Frank Bretz, and Mario Cortina-Borja. Reference range: Which statistical intervals to use? *Statistical methods in medical research*, 30(2):523–534, 2021.

[74] Michael Daniel Lucagbo and Thomas Mathew. Rectangular tolerance regions and multivariate normal reference regions in laboratory medicine. *Biometrical Journal*, 65(3):2100180, 2023.

[75] Stefan Wellek and Christine Jennen-Steinmetz. Reference ranges: Why tolerance intervals should not be used. comment on liu, bretz and cortina-borja, reference range: Which statistical intervals to use? smmr, 2021, vol. 30 (2) 523-534. *Statistical methods in medical research*, 31(11):2255–2256, 2022.

[76] David B Dunson, Ya Xue, and Lawrence Carin. The matrix stick-breaking process: Flexible bayes meta-analysis. *Journal of the American Statistical Association*, 103(481):317–327, 2008.

[77] Deborah Burr and Hani Doss. A bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, 100(469):242–251, 2005.

[78] George Karabatsos, Elizabeth Talbott, and Stephen G Walker. A bayesian nonparametric meta-analysis model. *Research Synthesis Methods*, 6(1):28–44, 2015.

[79] Connor A Emdin, Amit V Khera, and Sekar Kathiresan. Mendelian randomization. *Jama*, 318(19):1925–1926, 2017.

[80] George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*, 23(R1):R89–R98, 2014.

[81] Nuala A Sheehan, Vanessa Didelez, Paul R Burton, and Martin D Tobin. Mendelian randomisation and causal inference in observational epidemiology. *PLoS medicine*, 5(8):e177, 2008.

[82] Suzanne H Gage, Hannah J Jones, Stephen Burgess, Jack Bowden, G Davey Smith, Stanley Zammit, and Marcus R Munafo. Assessing causality in associations between cannabis use and schizophrenia risk: a two-sample mendelian randomization study. *Psychological medicine*, 47(5):971–980, 2017.

[83] Henning Jansen, Nilesh J Samani, and Heribert Schunkert. Mendelian randomization studies in coronary artery disease. *European heart journal*, 35(29):1917–1924, 2014.

[84] Daniel I Swerdlow, Karoline B Kuchenbaecker, Sonia Shah, Reecha Sofat, Michael V Holmes, Jon White, Jennifer S Mindell, Mika Kivimaki, Eric J Brunner, John C Whittaker, et al. Selecting instruments for mendelian randomization in the wake of genome-wide association studies. *International journal of epidemiology*, 45(5):1600–1616, 2016.

[85] Michael V Holmes, Mika Ala-Korpela, and George Davey Smith. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nature Reviews Cardiology*, 14(10):577–590, 2017.

[86] Fernando Pires Hartwig, Maria Carolina Borges, Bernardo Lessa Horta, Jack Bowden, and George Davey Smith. Inflammatory biomarkers and risk of schizophrenia: a 2-sample mendelian randomization study. *JAMA psychiatry*, 74(12):1226–1233, 2017.

[87] Christopher F Baum, Mark E Schaffer, and Steven Stillman. Instrumental variables and gmm: Estimation and testing. *The Stata Journal*, 3(1):1–31, 2003.

[88] Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, 16(4):309–330, 2007.

[89] Jeffrey Wooldridge. Instrumental variables estimation and two stage least squares. *Introductory Econometrics: A Modern Approach. Nashville, TN: South-Western*, 2009.

[90] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665, 2013.

[91] Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

[92] Neil M Davies, Stephanie von Hinke Kessler Scholder, Helmut Farbmacher, Stephen Burgess, Frank Windmeijer, and George Davey Smith. The many weak instruments problem and mendelian randomization. *Statistics in Medicine*, 34(3):454–468, 2015.

[93] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

[94] Eric AW Slob and Stephen Burgess. A comparison of robust mendelian randomization methods using summary data. *BioRxiv*, page 577940, 2019.

[95] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):nrg3461, 2013.

[96] Guy Freeman, Benjamin J Cowling, and C Mary Schooling. Power and sample size calculations for mendelian randomization studies using one genetic instrument. *International journal of epidemiology*, 42(4):1157–1163, 2013.

[97] John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450, 1995.

[98] Michal Kolesár, Raj Chetty, John Friedman, Edward Glaeser, and Guido W Imbens. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484, 2015.

[99] Miguel A Hernán and James M Robins. *Causal Inference: What If*. CRC Boca Raton, FL, 2020.

[100] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.

[101] Jessica MB Rees, Angela M Wood, Frank Dudbridge, and Stephen Burgess. Robust methods in mendelian randomization via penalization of heterogeneous causal estimates. *PloS one*, 14(9):e0222362, 2019.

[102] Stephen Burgess, Simon G Thompson, and CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in mendelian randomization studies. *International journal of epidemiology*, 40(3):755–764, 2011.

[103] Sheng Wang and Hyunseung Kang. Weak-instrument robust tests in two-sample summary-data mendelian randomization. *Biometrics*, 78(4):1699–1713, 2022.

[104] Ting Ye, Jun Shao, and Hyunseung Kang. Debiased inverse-variance weighted estimator in two-sample summary-data mendelian randomization. *The Annals of statistics*, 49(4):2079–2100, 2021.

[105] Ting Ye, Zhonghua Liu, Baoluo Sun, and Eric Tchetgen Tchetgen. Genius-mawii: For robust mendelian randomization with many weak invalid instruments. *arXiv preprint arXiv:2107.06238*, 2021.

[106] Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019.

[107] Stephen Burgess and Simon G Thompson. *Mendelian randomization: methods for using genetic variants in causal estimation*. CRC Press, 2015.

[108] Stephen Burgess and Simon G Thompson. Use of allele scores as instrumental variables for mendelian randomization. *International journal of epidemiology*, 42(4):1134–1144, 2013.

[109] Zhaotong Lin, Isaac Pan, and Wei Pan. A practical problem with egger regression in mendelian randomization. *PLoS genetics*, 18(5):e1010166, 2022.

[110] Donald M Lyall, Carlos Celis-Morales, Joey Ward, Stamatina Iliodromiti, Jana J Anderson, Jason MR Gill, Daniel J Smith, Uduakobong Efanga Ntuk, Daniel F Mackay, Michael V Holmes, et al. Association of body mass index with cardiometabolic disease in the uk biobank: a mendelian randomization study. *JAMA cardiology*, 2(8):882–889, 2017.

[111] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[112] Eric Tchetgen Tchetgen, BaoLuo Sun, and Stefan Walter. The genius approach to robust mendelian randomization inference. *Statistical Science*, 36(3):443–464, 2021.

[113] Jack Bowden and Michael V Holmes. Meta-analysis and mendelian randomization: A review. *Research synthesis methods*, 10(4):486–496, 2019.

[114] Zhaotong Lin, Souvik Seal, and Saonli Basu. Estimating snp heritability in presence of population substructure in biobank-scale datasets. *Genetics*, 220(4):iyac015, 2022.

[115] David A Schoenfeld, Hui Zheng, and Dianne M Finkelstein. Bayesian design using adult data to augment pediatric trials. *Clinical Trials*, 6(4):297–304, 2009.

[116] Jeremy MG Taylor, Kyuseong Choi, and Peisong Han. Data integration: exploiting ratios of parameter estimates from a reduced external model. *Biometrika*, 110(1):119–134, 2023.

[117] Brian P Hobbs, Bradley P Carlin, Sumithra J Mandrekar, and Daniel J Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011.

[118] Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

[119] Christian Röver and Tim Friede. Dynamically borrowing strength from another study through shrinkage estimation. *Statistical Methods in Medical Research*, 29(1):293–308, 2020.

[120] Mario Medvedovic, Ka Yee Yeung, and Roger Eugene Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

[121] Abel Rodriguez, David B Dunson, and Alan E Gelfand. The nested dirichlet process. *Journal of the American statistical Association*, 103(483):1131–1154, 2008.

[122] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

[123] Andrew J Grant and Stephen Burgess. A bayesian approach to mendelian randomization using summary statistics in the univariable and multivariable settings with correlated pleiotropy. *bioRxiv*, pages 2023–05, 2023.

[124] Carlo Berzuini, Hui Guo, Stephen Burgess, and Luisa Bernardinelli. A bayesian approach to mendelian randomization with multiple pleiotropic variants. *Biostatistics*, 21(1):86–101, 2020.

[125] Stephen Burgess and Simon G Thompson. Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American journal of epidemiology*, 181(4):251–260, 2015.

[126] Yangqing Deng, Dongsheng Tu, Chris J O'Callaghan, Derek J Jonker, Christos S Karapetis, Jeremy Shapiro, Geoffrey Liu, and Wei Xu. A bayesian approach for two-stage multivariate mendelian randomization with mixed outcomes. *Statistics in Medicine*, 2023.

[127] Zhaotong Lin, Haoran Xue, and Wei Pan. Robust multivariable mendelian randomization based on constrained maximum likelihood. *The American Journal of Human Genetics*, 110(4):592–605, 2023.

Appendix A

Supplementary materials

A.1 Appendix for NP methods

A.1.1 Figures

Means of estimated reference interval limits: study means generated from a mixture of normal distributions: $\mu = (8, 10, 11)$, $\tau^2 = (1.5^2, 0.8^2, 0.5^2)$ and $p = (0.4, 0.4, 0.3)$. The solid lines represent the true 2.5th and 97.5th percentiles of the marginal distribution of measurements. NP: nonparametric model using one DP for study means; NP-2: nonparametric model using two DPs for study means and within-study variances; Mix: mixture distribution method; Freq: frequentist method; Emp: empirical method; Bayes: Bayesian parametric method.

Figure A.1

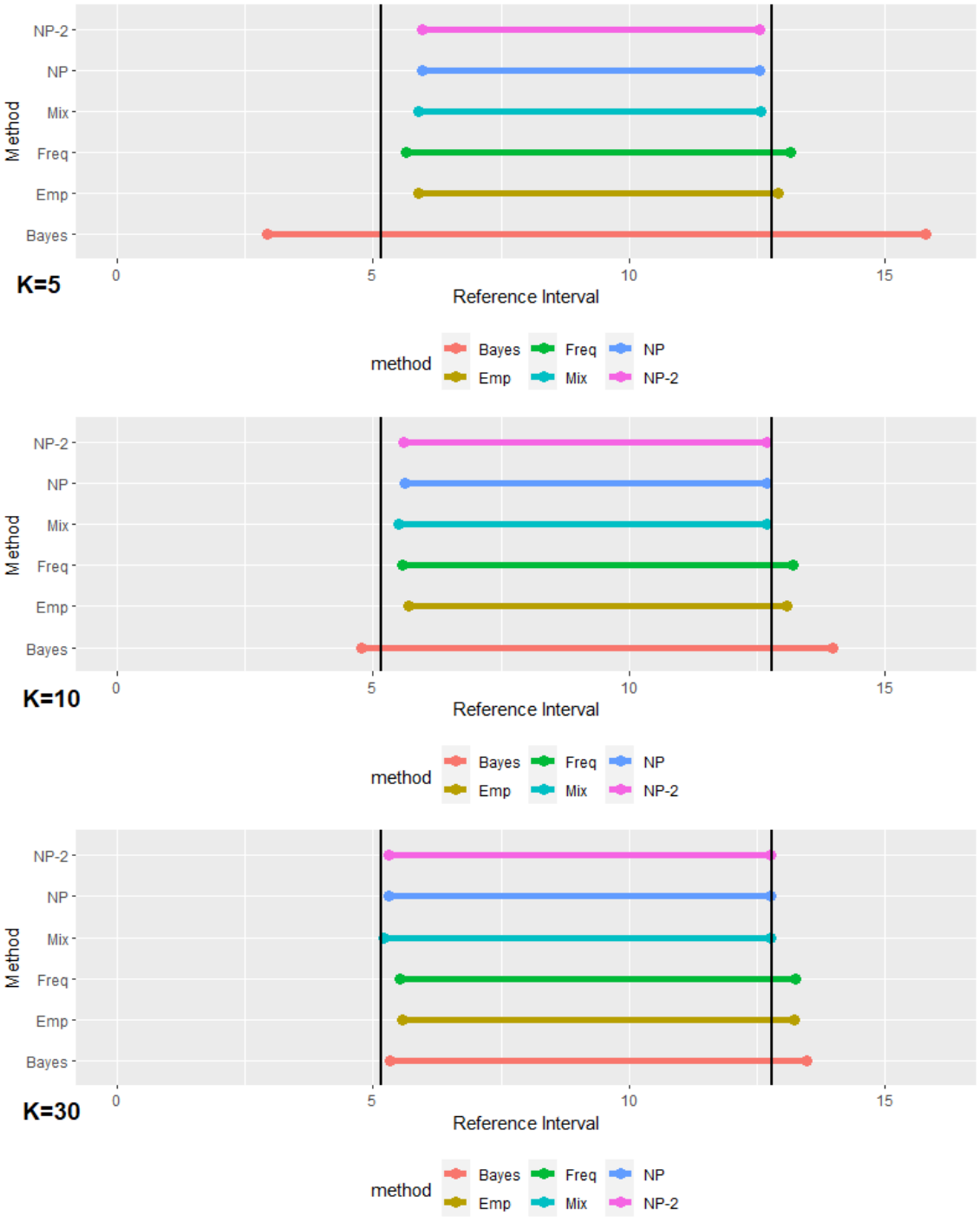

NP-2
NP
Mix
Freq
Emp
Bayes
Method
0
5
10
15
Reference Interval
K=5
K=10
K=30
method
Bayes
Freq
NP
Emp
Mix
NP-2

Figure A.2: Mean of estimated reference intervals: study means generated from a log-normal distribution (mean = 8, SD = 3.5). The solid lines represent the true 2.5th and 97.5th percentiles of the marginal distribution of measurements. NP: nonparametric model using one DP for study means; NP-2: nonparametric model using two DPs for study means and within-study variances; Mix: mixture distribution method; Freq: frequentist method; Emp: empirical method; Bayes: Bayesian parametric method.

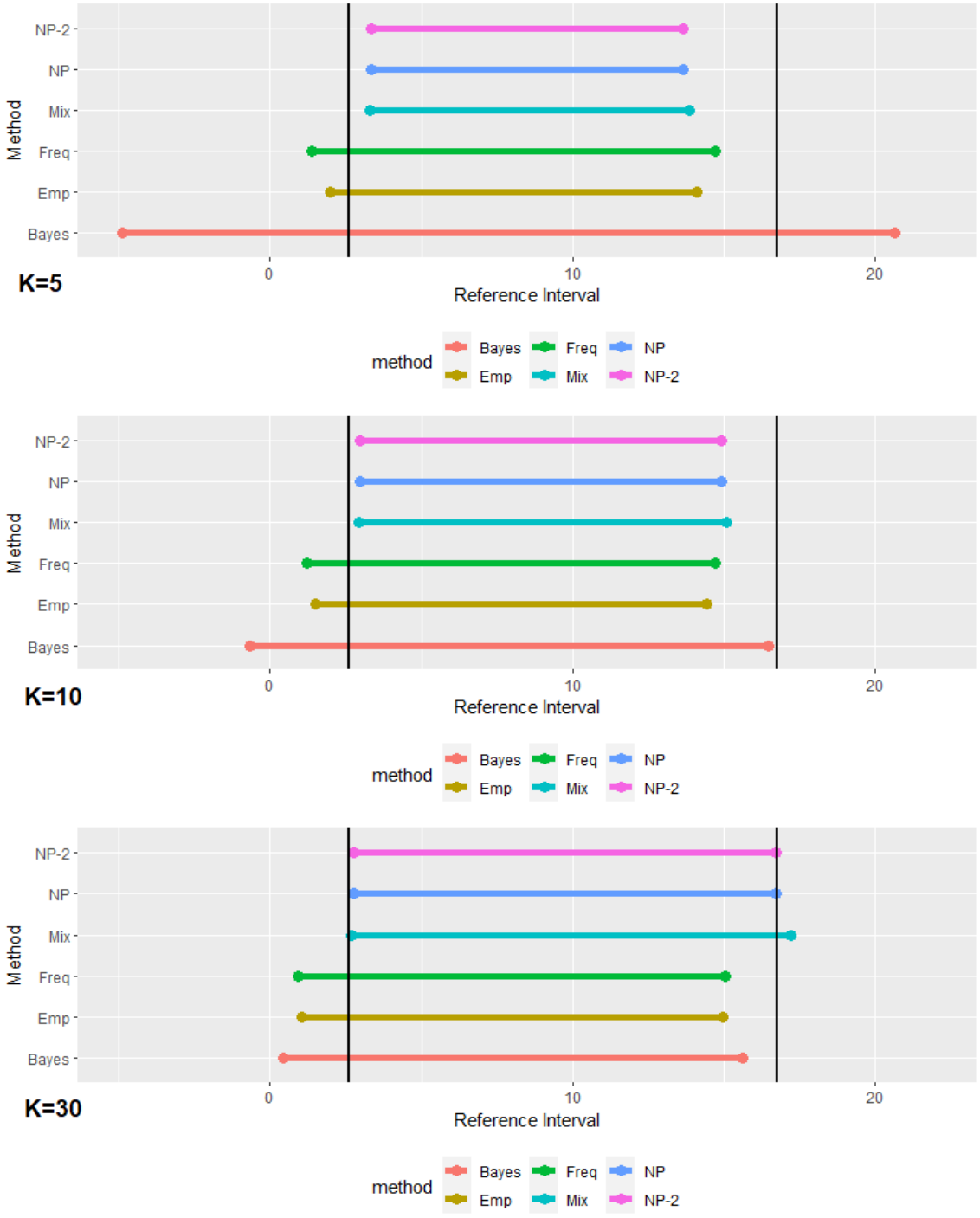


Figure A.3: Means of estimated reference interval limits: study means generated from a gamma distribution (mean = 5, SD = 3.5). The solid lines represent the true 2.5th and 97.5th percentiles of the marginal distribution of measurements. NP: nonparametric model using one DP for study means; NP-2: nonparametric model using two DPs for study means and within-study variances; Mix: mixture distribution method; Freq: frequentist method; Emp: empirical method; Bayes: Bayesian parametric method.

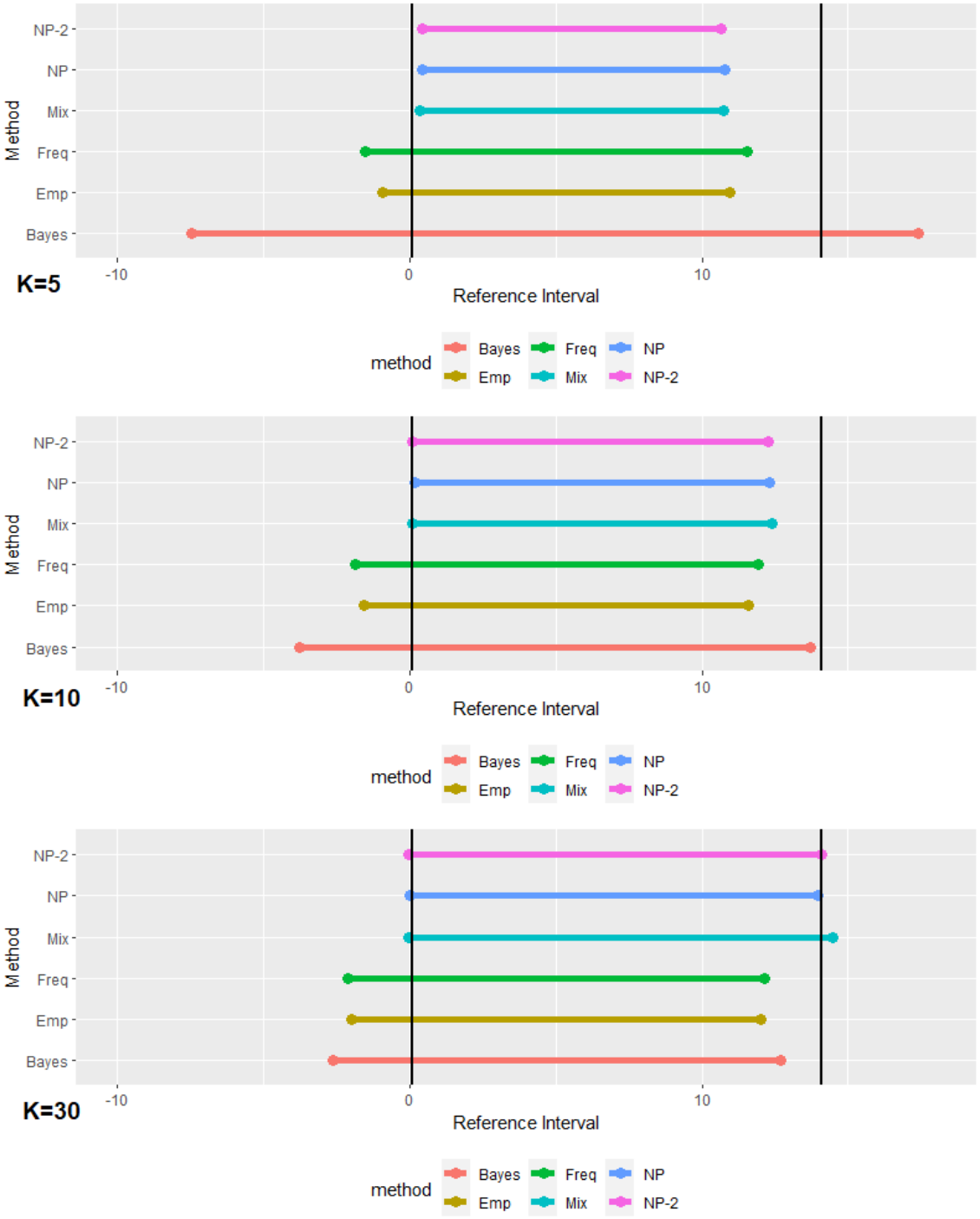
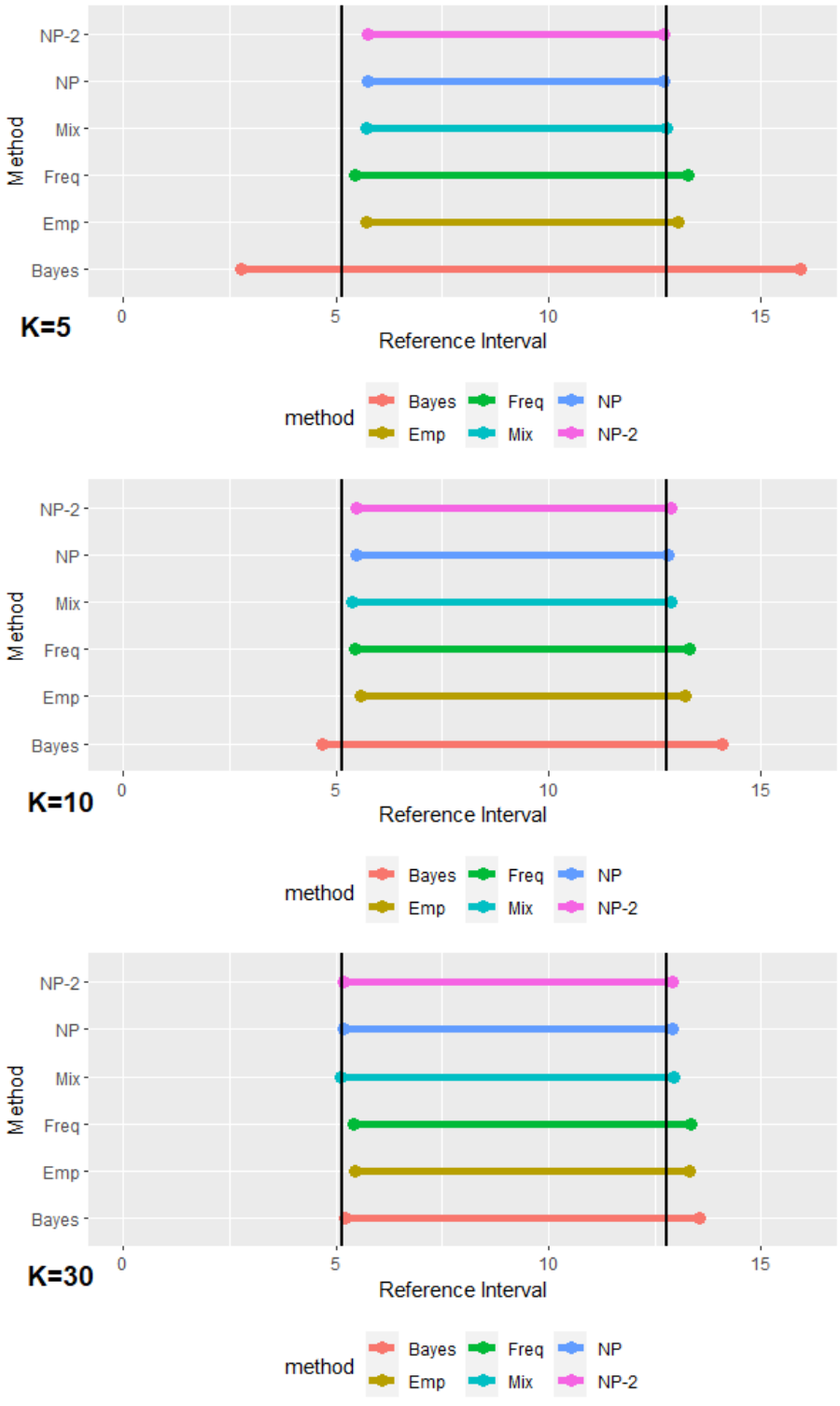


Figure A.4: Simulation Results for Mixed Normal adding outliers: the outliers defined values smaller than $Q1 - 1.5 \times IQR$ or larger than $Q3 + 1.5 \times IQR$. The overall proportion of outliers was close to 2.5%. Mean of Estimated reference intervals: the true effects generated from a mixture of normal distributions: $\mu = (8, 10, 11)$, $\tau^2 = (1.5^2, 0.8^2, 0.5^2)$ and $p = (0.4, 0.4, 0.3)$. The solid lines represent the true 95% reference intervals. NP: nonparametric model using one DP for study means; NP-2: nonparametric model using two DPs for study means and within-study variances; Mix: mixture distribution method; Freq: frequentist method; Emp: empirical method; Bayes: Bayesian parametric method.



A.2 Appendix C

A.2.1 Proof for $E[A_{ij}e_{ij}^{yx}))] = 0$

$$\begin{aligned}
E[A_{ij}e_{ij}^{yx}))] =& E[\boldsymbol{G_i^T G_j}(\boldsymbol{G_{ic}^T\alpha_c} + \boldsymbol{G_{id}^T\alpha_d} + e_{y_i})(\boldsymbol{G_{jb}^T\beta_b} + \boldsymbol{G_{jc}^T\beta_c} + e_j^x)] \\
\xrightarrow[e_{y_i}\perp e_{x_j}]{e\perp G}& E[\boldsymbol{G_i^T G_j}(\boldsymbol{G_{ic}^T\alpha_c} + \boldsymbol{G_{id}^T\alpha_d})(\boldsymbol{G_{jb}^T\beta_b} + \boldsymbol{G_{jc}^T\beta_c})] \\
\xrightarrow[e_{y_i}\perp e_{x_j}, e\perp G]{\mathbb{G}_{\boldsymbol{a}}\perp\mathbb{G}_{\boldsymbol{b}}\perp\mathbb{G}_{\boldsymbol{c}}\perp\mathbb{G}_{\boldsymbol{d}}}& E[\boldsymbol{G_i^T G_j}(\boldsymbol{G_{ic}^T\alpha_c} + \boldsymbol{G_{id}^T\alpha_d})(\boldsymbol{G_{jb}^T\beta_b} + \boldsymbol{G_{jc}^T\beta_c})] \\
\xrightarrow[\alpha_c\perp\beta_b]{\alpha_d\perp\beta_b,\beta_c}& E[\boldsymbol{G_i^T G_j G_{ic}^T\alpha_c G_{jc}^T\beta_c}] \\
\xrightarrow{E(\alpha_c\beta_c)}& 0.
\end{aligned} \tag{A.1}$$

A.2.2 Bias and asymptotic variance of the TS-RE

In the following proof, we used r to denote terms with expectation 0 ($E(r) = 0$). Here the genotype data are assumed standardized that $E(G_{ik} = 0)$,$Var(G_{ik}) = 1, E(A_{ij} = 0)$. Let $\delta = E\left[A_{ij}X_iX_j\right], \eta = E\left[A_{ij}X_iY_j\right]$, then $\hat{\theta}_{TS-RE}$ can be regarded as a ratio of η/δ.

$$\begin{aligned}
\hat{\delta} =& \frac{1}{N}\sum_{i<j} E(A_{ij}X_iX_j) \\
\xrightarrow[e_{x_j}\perp Y_i]{e_{y_i}\perp X_j}& E[A_{ij}(\boldsymbol{G_{ib}^T\beta_b} + \boldsymbol{G_{ic}^T\beta_c})(\boldsymbol{G_{jb}^T\beta_b} + \boldsymbol{G_{jc}^T\beta_c}) + r] \\
\xrightarrow[p\neq q]{\beta p\perp\beta_q}& \frac{1}{M}E[\sum_{k=1}^{M} G_{ik}G_{jk}(\boldsymbol{G_{ib}^T G_{jb}}\beta_{bk}^2 + \boldsymbol{G_{ic}^T G_{jc}}\beta_{ck}^2)] \\
\xrightarrow[p\neq q]{G_{\cdot p}\perp G_{\cdot q}}& \frac{1}{M}E[\sum_{k=1}^{M_b} G_{ik}^2G_{jk}^2\beta_{bk}^2 + \sum_{k=1}^{M_c} G_{ik}^2G_{jk}^2\beta_{ck}^2] \\
=& \frac{M_b}{M}E(\beta_{bk}^2) + \frac{M_c}{M}E(\beta_{ck}^2)
\end{aligned} \tag{A.2}$$

Similarly, using Eq A.1 and Eq A.2

$$\begin{aligned}\hat{\eta} =& E(\widehat{Cov}\,[A_{ij}, Y_i X_j]) \\ =& \frac{1}{M} E[\theta \sum_{k=1}^{M_b} G_{ik}^2 G_{jk}^2 \beta_{bk}^2 + \sum_{k=1}^{M_c} G_{ik}^2 G_{jk}^2 (\theta \beta_{ck}^2 + \beta_{ck}\alpha_{ck}))] \\ =& \frac{M_b}{M}\theta E(\beta_{bk}^2) + \frac{M_c}{M}\theta[E(\beta_{ck}^2) + E(\beta_{ck}\alpha_{ck})]\end{aligned} \tag{A.3}$$

The ratio of $\hat{\eta}$ and $\hat{\delta}$ is the TS-RE estimator

$$\hat{\theta}_{TS-RE} = \frac{\hat{\eta}}{\hat{\delta}} = \theta + \frac{M_c E(\beta_{ck}\alpha_{ck})}{M_b E(\beta_{bk}^2) + M_c E(\beta_{ck}^2)}. \tag{A.4}$$

The bias will tend to zero if $E(\beta_c \alpha_c) = 0$, e.g., the Inside assumption holds, $\rho_{\mathbb{G}_c} = 0$, and either $\mu_{\mathbb{G}_c^x} = 0$ or $\mu_{\mathbb{G}_c^y} = 0$, t. Given the $\theta = \frac{\eta}{\delta}$, we can use the multivariate Delta method to derive the asymptotic property for $\hat{\theta}_{GMM}$,

$$\sqrt{N}[\hat{\theta}_{TS-RE} - \theta] \xrightarrow{\mathcal{D}} N(0, \tau^2),$$

where

$$\begin{aligned}\tau^2 =& \frac{\eta^2}{\delta^2}\left(\frac{Var(A_{ij}Y_iX_j)}{\eta^2} + \frac{Var(A_{ij}X_iX_j)}{\delta^2} - 2\frac{Cov(A_{ij}Y_iX_j, A_{ij}X_iX_j)}{\eta\delta}\right) \\ =& \frac{M[M_b E(\beta_{bk}^2) + M_c E(\beta_{ck}^2) + \sigma_{e_x}^2][M_c E(\alpha_{ck}^2) + M_d E(\alpha_{dk}^2) + \sigma_{e_y}^2]}{[M_b E(\beta_{bk}^2) + M_c E(\beta_{ck}^2)]^2}\end{aligned} \tag{A.5}$$

Thus $Var(\hat{\theta}_{TS-RE}) = \frac{2M}{n(n-1)} \frac{[M_b E(\beta_{bk}^2)+M_c E(\beta_{ck}^2)+\sigma_{e_x}^2][M_c E(\alpha_{ck}^2)+M_d E(\alpha_{dk}^2)+\sigma_{e_y}^2]}{[M_b E(\beta_{bk}^2)+M_c E(\beta_{ck}^2)]^2}$. To show the relationship between our method with IVW and Egger, we consider the situation that all included IVs have a direct effect on X, which are from $\mathbb{G}_b$ and $\mathbb{G}_c$, then

$$\begin{aligned}E(\hat{\theta}_{IVW}) &= E((\boldsymbol{X^T P_G X})^{-1}\boldsymbol{X^T P_G Y}) \\ &\xrightarrow[G_{ik}^2=1]{\beta p \perp \beta_q, p \neq q} \theta + \frac{M_c E(\beta_{ck}\alpha_{ck})}{M_b E(\beta_{bk}^2) + M_c E(\beta_{ck}^2)}\end{aligned} \tag{A.6}$$

$$
\begin{aligned}
E(\hat{\theta}_{Egger}) &= E(\frac{\boldsymbol{I^T G^T G I X^T P_G Y - X^T G I Y^T G I}}{\boldsymbol{I^T G^T G I X^T P_G X - X^T G I X^T G I}}) \\
&\xrightarrow[G_{ik}^2=1]{\beta p \perp \beta_q, p \neq q} \theta + \frac{M_c}{M} \frac{E(\beta_{ck}\alpha_{ck}) - E(\beta_{ck})E(\alpha_{ck})}{Var(\beta)}
\end{aligned}
\tag{A.7}
$$

Thus, if there are only two groups of IVs $\mathbb{G}, \mathbb{G}$, the bias of TS-RE is equivalent to the IVW.

A.2.3 Simulation results

M	p_w	IVs	SM	WM	IVW	Egger	Lasso	dIVW	TS-RE
$M_b = 100$	0.8	All	0.36 (0.06)	0.34 (0.04)	0.35 (0.04)	0.33 (0.05)	0.35 (0.04)	0.45 (0.05)	0.29 (0.06)
		Top20	0.33 (0.05)	0.33 (0.05)	0.33 (0.04)	0.33 (0.17)	0.33 (0.04)	0.36 (0.04)	0.32 (0.05)
	1	All	0.44 (0.08)	0.44 (0.07)	0.44 (0.05)	0.45 (0.10)	0.44 (0.06)	0.87 (0.16)	0.27 (0.17)
		Top20	0.44 (0.08)	0.45 (0.09)	0.45 (0.06)	0.53 (0.34)	0.45 (0.07)	0.53 (0.08)	0.42 (0.09)
$M_b = 500$	0.8	All	0.34 (0.02)	0.34 (0.02)	0.34 (0.02)	0.33 (0.02)	0.34 (0.02)	0.58 (0.03)	0.30 (0.03)
		Top20	0.33 (0.03)	0.33 (0.03)	0.34 (0.02)	0.37 (0.19)	0.34 (0.02)	0.36 (0.03)	0.33 (0.03)
	1	All	0.42 (0.04)	0.42 (0.03)	0.42 (0.03)	0.43 (0.04)	0.42 (0.03)	0.98 (0.10)	0.29 (0.08)
		Top20	0.43 (0.07)	0.43 (0.07)	0.43 (0.06)	0.43 (0.42)	0.43 (0.06)	0.48 (0.07)	0.42 (0.07)
$M_b = 1000$	0.8	All	0.33 (0.02)	0.33 (0.02)	0.33 (0.01)	0.33 (0.02)	0.33 (0.1)	0.73 (0.04)	0.30 (0.02)
		Top20	0.33 (0.03)	0.33 (0.03)	0.33 (0.02)	0.33 (0.20)	0.33 (0.02)	0.36 (0.03)	0.33 (0.03)
	1	All	0.40 (0.03)	0.40 (0.02)	0.40 (0.02)	0.40 (0.03)	0.40 (0.02)	1.13 (0.10)	0.29 (0.05)
		Top20	0.40 (0.05)	0.40 (0.05)	0.40 (0.04)	0.35 (0.36)	0.40 (0.04)	0.44 (0.05)	0.40 (0.05)
$M_b = 2000$	0.8	All	0.32 (0.01)	0.32 (0.01)	0.32 (0.01)	0.32 (0.01)	0.32 (0.01)	1.02 (0.06)	0.30 (0.02)
		Top20	0.32 (0.02)	0.32 (0.02)	0.32 (0.02)	0.32 (0.17)	0.32 (0.02)	0.35 (0.02)	0.32 (0.02)
	1	All	0.38 (0.03)	0.38 (0.02)	0.37 (0.02)	0.37 (0.03)	0.37 (0.02)	1.42 (0.15)	0.30 (0.05)
		Top20	0.37 (0.04)	0.37 (0.04)	0.37 (0.03)	0.40 (0.36)	0.37 (0.03)	0.41 (0.03)	0.37 (0.04)
$M_b = 5000$	0.8	All	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	1.93 (0.19)	0.30 (0.02)
		Top20	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.32 (0.14)	0.31 (0.01)	0.34 (0.01)	0.31 (0.01)
	1	All	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	2.39 (0.24)	0.30 (0.04)
		Top20	0.34 (0.03)	0.34 (0.03)	0.34 (0.02)	0.33 (0.30)	0.35 (0.02)	0.37 (0.03)	0.34 (0.03)

Table S1: Mean and SE of different methods: 20% of the IVs having strong effects on X. The large effect for 20% of the IVs is 0.2. The variance parameter $\sigma_{\mathbb{G}_b}$ was 0.05 and the corresponding heritability values are $0.02, 0.11, 0.38, 0.56, 0.71, 0.86$. The true causal effect is $\theta = 0.3$

n	p_w	IVs	SM	WM	IVW	Egger	Lasso	dIVW	TS-RE
1000	0.8	All	0.33 (0.02)	0.33 (0.02)	0.33 (0.01)	0.33 (0.02)	0.34 (0.01)	0.75 (0.05)	0.30 (0.02)
		20	0.33 (0.03)	0.33 (0.03)	0.33 (0.02)	0.35 (0.20)	0.33 (0.03)	0.36 (0.02)	0.33 (0.03)
	1	All	0.49 (0.04)	0.49 (0.03)	0.49 (0.03)	0.48 (0.04)	0.48 (0.03)	2.01 (0.32)	0.28 (0.12)
		20	0.47 (0.07)	0.48 (0.07)	0.48 (0.05)	0.55 (0.54)	0.47 (0.06)	0.54 (0.06)	0.47 (0.07)
3000	0.8	All	0.33 (0.01)	0.31 (0.01)	0.32 (0.01)	0.31 (0.01)	0.32 (0.01)	0.45 (0.01)	0.30 (0.01)
		20	0.32 (0.02)	0.32 (0.02)	0.32 (0.02)	0.32 (0.20)	0.32 (0.02)	0.33 (0.02)	0.32 (0.02)
	1	All	0.42 (0.03)	0.42 (0.02)	0.42 (0.02)	0.42 (0.03)	0.42 (0.02)	0.88 (0.07)	0.30 (0.05)
		20	0.42 (0.05)	0.42 (0.05)	0.42 (0.04)	0.43 (0.41)	0.42 (0.04)	0.45 (0.04)	0.42 (0.06)
5000	0.8	All	0.32 (0.01)	0.31 (0.01)	0.31 (0.01)	0.30 (0.01)	0.31 (0.01)	0.39 (0.01)	0.30 (0.01)
		20	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.33 (0.20)	0.31 (0.01)	0.32 (0.01)	0.31 (0.02)
	1	All	0.39 (0.02)	0.39 (0.02)	0.39 (0.01)	0.39 (0.02)	0.39 (0.02)	0.64 (0.03)	0.30 (0.02)
		20	0.39 (0.04)	0.39 (0.05)	0.39 (0.04)	0.34 (0.37)	0.38 (0.03)	0.41 (0.04)	0.38 (0.05)
10000	0.8	All	0.31 (0.007)	0.30 (0.006)	0.31 (0.004)	0.30 (0.005)	0.30 (0.005)	0.30 (0.005)	0.30 (0.005)
		20	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.36 (0.15)	0.30 (0.01)	0.31 (0.01)	0.31 (0.01)
	1	All	0.36 (0.02)	0.36 (0.02)	0.36 (0.01)	0.36 (0.02)	0.36 (0.01)	0.47 (0.02)	0.30 (0.02)
		20	0.36 (0.04)	0.36 (0.05)	0.36(0.04)	0.35 (0.29)	0.35 (0.03)	0.37 (0.03)	0.35 (0.04)

Table S2: Mean and SE of different methods under different sample sizes, $M_b = 1000$

	p_w	IVs	SM	WM	IVW	Egger	Lasso	dIVW	TS-RE
BP	0	All	0.30 (0.01)	0.30 (0.01)	0.30 (0.01)	0.33 (0.02)	0.30 (0.01)	0.30 (0.01)	0.30 (0.01)
		Top20	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.36 (0.02)	0.31 (0.01)	0.31 (0.01)	0.30 (0.01)
	0.8	All	0.32 (0.02)	0.31 (0.01)	0.31 (0.01)	0.30 (0.01)	0.31 (0.01)	0.32 (0.01)	0.30 (0.01)
		Top20	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.36 (0.16)	0.31 (0.01)	0.31 (0.01)	0.30 (0.01)
	1	All	0.36 (0.02)	0.36 (0.01)	0.36 (0.01)	0.36 (0.02)	0.36 (0.01)	0.47 (0.02)	0.30 (0.02)
		Top20	0.36 (0.03)	0.36 (0.03)	0.36 (0.03)	0.40 (0.34)	0.36 (0.03)	0.37 (0.03)	0.35 (0.03)
DP	0	All	0.30 (0.01)	0.30 (0.01)	0.30 (0.01)	0.33 (0.02)	0.30 (0.01)	0.31 (0.01)	0.30 (0.01)
		Top20	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.34 (0.19)	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)
	0.8	All	0.32 (0.02)	0.31 (0.01)	0.31 (0.01)	0.30 (0.01)	0.31 (0.01)	0.32 (0.01)	0.30 (0.01)
		Top20	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)	0.35 (0.16)	0.31 (0.01)	0.31 (0.01)	0.31 (0.01)
	1	All	0.36 (0.04)	0.36 (0.03)	0.36 (0.03)	0.36 (0.04)	0.36 (0.03)	0.45 (0.03)	0.30 (0.04)
		Top20	0.36 (0.04)	0.36 (0.05)	0.36 (0.04)	0.40 (0.25)	0.36 (0.04)	0.38 (0.04)	0.35 (0.05)

Table S3: Performance of different methods under different proportion of weak IVs, $M_b = M_c = 100$

M_b	Method	Balanced Pleiropy						Directional Pleiropy					
		All IVs			Top 20 IVs			All IVs			Top 20 IVs		
		Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
0	SM	0.24	0.06	0.06	0.22	0.1	0.06	0.27	0.2	0.11	0.19	0.32	0.14
	WM	0.24	0.06	0.06	0.22	0.1	0.06	0.27	0.17	0.1	0.22	0.31	0.14
	IVW	0.24	0.05	0.06	0.23	0.09	0.06	0.26	0.13	0.09	0.21	0.24	0.1
	Egger	0.24	0.07	0.06	0.24	0.93	0.93	0.25	0.16	0.09	0.67	2.23	5.43
	TS-RE	-0.01	0.18	0.03	0.22	0.11	0.06	0.01	0.52	0.27	0.2	0.29	0.12
100	SM	0.24	0.06	0.06	0.24	0.09	0.07	0.27	0.19	0.11	0.25	0.28	0.14
	WM	0.24	0.05	0.06	0.24	0.09	0.07	0.26	0.16	0.09	0.24	0.28	0.14
	IVW	0.24	0.05	0.06	0.24	0.08	0.06	0.26	0.13	0.08	0.25	0.21	0.11
	Egger	0.24	0.06	0.06	0.3	0.76	0.67	0.26	0.16	0.09	0.25	1.88	3.58
	TS-RE	0.01	0.17	0.03	0.23	0.1	0.06	0.03	0.44	0.2	0.25	0.27	0.14
200	SM	0.23	0.06	0.06	0.23	0.1	0.06	0.22	0.16	0.07	0.19	0.28	0.12
	WM	0.23	0.05	0.06	0.23	0.09	0.06	0.21	0.14	0.07	0.21	0.29	0.13
	IVW	0.23	0.04	0.06	0.23	0.08	0.06	0.21	0.12	0.06	0.22	0.24	0.1
	Egger	0.23	0.06	0.05	0.23	0.8	0.69	0.19	0.15	0.06	0.3	1.95	3.88
	TS-RE	-0.01	0.16	0.03	0.22	0.09	0.06	-0.11	0.47	0.24	0.22	0.29	0.13
300	SM	0.23	0.06	0.06	0.22	0.1	0.06	0.23	0.14	0.07	0.17	0.21	0.08
	WM	0.23	0.05	0.06	0.22	0.1	0.06	0.2	0.12	0.06	0.19	0.22	0.08
	IVW	0.23	0.05	0.06	0.23	0.08	0.06	0.22	0.12	0.06	0.2	0.19	0.08
	Egger	0.23	0.06	0.06	0.18	0.86	0.78	0.2	0.15	0.06	0.3	1.85	3.51
	TS-RE	-0.02	0.17	0.03	0.21	0.1	0.05	-0.09	0.38	0.16	0.17	0.23	0.08
400	SM	0.23	0.06	0.06	0.22	0.09	0.06	0.23	0.14	0.07	0.23	0.21	0.1
	WM	0.23	0.05	0.06	0.23	0.09	0.06	0.24	0.13	0.07	0.23	0.22	0.1
	IVW	0.23	0.04	0.06	0.23	0.08	0.06	0.24	0.11	0.07	0.23	0.19	0.09
	Egger	0.24	0.06	0.06	0.28	0.64	0.49	0.24	0.13	0.08	0.33	1.59	2.62
	TS-RE	-0.01	0.17	0.03	0.23	0.1	0.06	-0.02	0.36	0.13	0.23	0.23	0.11
500	SM	0.24	0.05	0.06	0.24	0.09	0.06	0.22	0.12	0.06	0.2	0.18	0.08
	WM	0.24	0.05	0.06	0.24	0.09	0.06	0.22	0.11	0.06	0.21	0.18	0.07
	IVW	0.24	0.04	0.06	0.24	0.07	0.06	0.22	0.1	0.06	0.21	0.17	0.07
	Egger	0.24	0.06	0.06	0.23	0.65	0.47	0.21	0.13	0.06	0.12	1.79	3.22
	TS-RE	-0.02	0.16	0.02	0.23	0.09	0.06	-0.1	0.35	0.13	0.19	0.2	0.08
600	SM	0.24	0.05	0.06	0.22	0.09	0.05	0.24	0.11	0.07	0.22	0.17	0.08
	WM	0.23	0.05	0.06	0.22	0.09	0.05	0.24	0.09	0.07	0.22	0.17	0.08
	IVW	0.24	0.04	0.06	0.22	0.08	0.05	0.24	0.09	0.07	0.22	0.16	0.07
	Egger	0.23	0.06	0.06	0.13	0.68	0.47	0.23	0.11	0.06	0.47	1.3	1.9
	TS-RE	-0.02	0.14	0.02	0.2	0.1	0.05	-0.01	0.27	0.07	0.2	0.18	0.07
700	SM	0.25	0.05	0.06	0.22	0.09	0.06	0.24	0.1	0.07	0.22	0.16	0.07
	WM	0.24	0.05	0.06	0.22	0.09	0.06	0.22	0.09	0.06	0.22	0.16	0.07
	IVW	0.24	0.04	0.06	0.22	0.08	0.06	0.23	0.09	0.06	0.21	0.15	0.07
	Egger	0.23	0.05	0.06	0.36	0.5	0.38	0.21	0.11	0.06	0.16	1.04	1.1
	TS-RE	-0.02	0.15	0.02	0.21	0.09	0.05	-0.05	0.27	0.07	0.19	0.19	0.07
800	SM	0.24	0.05	0.06	0.23	0.09	0.06	0.24	0.09	0.06	0.23	0.13	0.07
	WM	0.24	0.04	0.06	0.23	0.09	0.06	0.23	0.08	0.06	0.23	0.13	0.07
	IVW	0.23	0.04	0.06	0.23	0.07	0.06	0.23	0.07	0.06	0.23	0.12	0.07
	Egger	0.24	0.05	0.06	0.17	0.57	0.35	0.23	0.09	0.06	0.22	0.91	0.87
	TS-RE	-0.03	0.13	0.02	0.22	0.09	0.05	-0.03	0.24	0.06	0.22	0.14	0.07
900	SM	0.24	0.04	0.06	0.25	0.08	0.07	0.24	0.07	0.06	0.23	0.11	0.07
	WM	0.24	0.04	0.06	0.25	0.07	0.07	0.24	0.06	0.06	0.23	0.11	0.07
	IVW	0.24	0.04	0.06	0.24	0.06	0.06	0.24	0.06	0.06	0.24	0.11	0.07
	Egger	0.24	0.04	0.06	0.26	0.59	0.42	0.23	0.07	0.06	0.17	0.84	0.74
	TS-RE	-0.02	0.15	0.02	0.23	0.08	0.06	-0.02	0.21	0.04	0.23	0.13	0.07
1000	SM	0.24	0.04	0.06	0.22	0.08	0.06	0.24	0.04	0.06	0.22	0.08	0.06
	WM	0.24	0.03	0.06	0.23	0.07	0.06	0.24	0.03	0.06	0.23	0.07	0.06
	IVW	0.24	0.03	0.06	0.23	0.06	0.06	0.24	0.03	0.06	0.23	0.06	0.06
	Egger	0.23	0.04	0.06	0.31	0.57	0.43	0.23	0.04	0.06	0.31	0.57	0.43
	TS-RE	-0.03	0.13	0.02	0.22	0.08	0.06	-0.03	0.13	0.02	0.22	0.08	0.06

Table S4: Simulation results for a mixture of IVs from $\mathbb{G}_b$ and $\mathbb{G}_c$. The total number of IVs is $1000 = M_b + M_c$ and the number M_b is varied from 0 to 1000. For the balanced pleiotropy $E(\alpha_c) = 0$, and for the directional pleiotropy $E(\alpha_c) = 0.1$, the InSIDE assumption is valid that $\rho_{\mathbb{G}_c} = 0$. All IVs have weak effect $N(\mu = 0, \sigma^3 = 0.03^3)$ and the true causal effect is $\theta = 0.1$

M_b	Method	Balanced Pleiropy						Directional Pleiropy					
		All IVs			Top 20 IVs			All IVs			Top 20 IVs		
		Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
0	SM	0.18	0.06	0.04	0.17	0.1	0.04	0.23	0.19	0.09	0.14	0.32	0.12
	WM	0.18	0.06	0.04	0.17	0.09	0.04	0.22	0.17	0.08	0.17	0.31	0.12
	IVW	0.18	0.05	0.04	0.18	0.08	0.04	0.2	0.14	0.06	0.15	0.24	0.08
	Egger	0.18	0.07	0.04	0.17	0.9	0.84	0.19	0.16	0.06	0.62	2.19	5.21
	TS-RE	0	0.16	0.03	0.17	0.11	0.04	0.01	0.51	0.26	0.14	0.29	0.11
100	SM	0.19	0.06	0.04	0.19	0.09	0.04	0.21	0.19	0.08	0.19	0.27	0.11
	WM	0.19	0.05	0.04	0.19	0.08	0.04	0.2	0.16	0.07	0.18	0.28	0.11
	IVW	0.19	0.05	0.04	0.18	0.07	0.04	0.21	0.12	0.06	0.2	0.21	0.08
	Egger	0.19	0.06	0.04	0.24	0.74	0.61	0.2	0.16	0.06	0.12	1.76	3.12
	TS-RE	0.01	0.16	0.03	0.18	0.1	0.04	0.03	0.43	0.19	0.2	0.27	0.11
200	SM	0.15	0.08	0.03	0.15	0.12	0.04	0.18	0.16	0.06	0.14	0.29	0.1
	WM	0.15	0.08	0.03	0.15	0.11	0.04	0.17	0.14	0.05	0.16	0.3	0.11
	IVW	0.16	0.07	0.03	0.15	0.1	0.03	0.16	0.12	0.04	0.17	0.24	0.09
	Egger	0.15	0.08	0.03	0.14	0.72	0.53	0.14	0.15	0.04	0.24	1.92	3.76
	TS-RE	-0.05	0.18	0.03	0.13	0.12	0.03	-0.11	0.47	0.23	0.17	0.29	0.11
300	SM	0.18	0.05	0.03	0.17	0.09	0.04	0.17	0.14	0.05	0.12	0.21	0.06
	WM	0.18	0.05	0.03	0.17	0.09	0.04	0.15	0.12	0.04	0.13	0.21	0.06
	IVW	0.18	0.04	0.03	0.17	0.08	0.04	0.17	0.12	0.04	0.14	0.19	0.06
	Egger	0.18	0.06	0.04	0.13	0.82	0.7	0.15	0.14	0.04	0.22	1.86	3.5
	TS-RE	-0.02	0.16	0.03	0.16	0.1	0.04	-0.1	0.38	0.15	0.12	0.23	0.07
400	SM	0.18	0.05	0.04	0.17	0.08	0.04	0.18	0.14	0.05	0.18	0.21	0.08
	WM	0.18	0.05	0.03	0.18	0.08	0.04	0.19	0.12	0.05	0.18	0.22	0.08
	IVW	0.18	0.04	0.03	0.18	0.08	0.04	0.19	0.11	0.05	0.18	0.19	0.07
	Egger	0.18	0.06	0.04	0.22	0.6	0.41	0.19	0.13	0.05	0.27	1.57	2.55
	TS-RE	-0.01	0.16	0.03	0.18	0.1	0.04	-0.01	0.36	0.13	0.18	0.23	0.08
500	SM	0.18	0.05	0.04	0.19	0.08	0.04	0.16	0.11	0.04	0.15	0.18	0.05
	WM	0.19	0.05	0.04	0.18	0.08	0.04	0.17	0.11	0.04	0.15	0.17	0.05
	IVW	0.19	0.04	0.04	0.19	0.07	0.04	0.17	0.09	0.04	0.15	0.17	0.05
	Egger	0.19	0.05	0.04	0.18	0.61	0.41	0.16	0.12	0.04	0.12	1.77	3.14
	TS-RE	-0.01	0.14	0.02	0.18	0.08	0.04	-0.1	0.34	0.13	0.14	0.2	0.06
600	SM	0.19	0.05	0.04	0.17	0.09	0.04	0.19	0.1	0.05	0.17	0.16	0.05
	WM	0.18	0.05	0.03	0.17	0.09	0.04	0.19	0.1	0.04	0.17	0.16	0.05
	IVW	0.18	0.04	0.04	0.17	0.08	0.03	0.19	0.09	0.04	0.16	0.16	0.05
	Egger	0.18	0.06	0.03	0.08	0.65	0.43	0.17	0.11	0.04	0.39	1.29	1.81
	TS-RE	-0.02	0.13	0.02	0.15	0.1	0.03	-0.01	0.27	0.07	0.15	0.18	0.06
700	SM	0.19	0.05	0.04	0.17	0.08	0.04	0.18	0.09	0.04	0.16	0.16	0.05
	WM	0.18	0.05	0.04	0.17	0.08	0.04	0.17	0.09	0.04	0.16	0.16	0.05
	IVW	0.19	0.04	0.04	0.17	0.07	0.04	0.18	0.08	0.04	0.16	0.15	0.05
	Egger	0.18	0.05	0.04	0.3	0.48	0.32	0.16	0.11	0.04	0.11	1.06	1.14
	TS-RE	-0.02	0.14	0.02	0.16	0.09	0.03	-0.04	0.26	0.07	0.14	0.19	0.05
800	SM	0.18	0.04	0.04	0.18	0.08	0.04	0.19	0.08	0.04	0.18	0.12	0.05
	WM	0.18	0.04	0.04	0.18	0.08	0.04	0.18	0.07	0.04	0.18	0.12	0.05
	IVW	0.18	0.04	0.03	0.18	0.07	0.04	0.18	0.07	0.04	0.18	0.11	0.05
	Egger	0.18	0.05	0.04	0.12	0.53	0.3	0.18	0.09	0.04	0.16	0.89	0.81
	TS-RE	-0.02	0.12	0.01	0.16	0.09	0.03	-0.03	0.23	0.05	0.17	0.14	0.05
900	SM	0.19	0.04	0.04	0.19	0.07	0.04	0.19	0.07	0.04	0.18	0.1	0.04
	WM	0.19	0.04	0.04	0.19	0.06	0.04	0.18	0.05	0.04	0.18	0.1	0.04
	IVW	0.19	0.03	0.04	0.19	0.06	0.04	0.19	0.06	0.04	0.19	0.1	0.04
	Egger	0.18	0.04	0.04	0.21	0.56	0.36	0.18	0.06	0.04	0.12	0.83	0.71
	TS-RE	-0.02	0.14	0.02	0.18	0.08	0.04	-0.02	0.2	0.04	0.18	0.13	0.05
1000	SM	0.19	0.04	0.04	0.17	0.07	0.03	0.19	0.04	0.04	0.17	0.07	0.03
	WM	0.18	0.03	0.03	0.18	0.07	0.04	0.18	0.03	0.03	0.18	0.07	0.04
	IVW	0.18	0.03	0.03	0.18	0.05	0.03	0.18	0.03	0.03	0.18	0.05	0.03
	Egger	0.18	0.04	0.03	0.25	0.54	0.35	0.18	0.04	0.03	0.25	0.54	0.35
	TS-RE	-0.03	0.12	0.01	0.17	0.07	0.03	-0.03	0.12	0.01	0.17	0.07	0.03

Table S5: Simulation results for a mixture of IVs from $\mathbb{G}_b$ and $\mathbb{G}_c$. The total number of IVs is $1000 = M_b + M_c$ and the number M_b is varied from 0 to 1000. For the balanced pleiotropy $E(\alpha_c) = 0$, and for the directional pleiotropy $E(\alpha_c) = 0.1$, the InSIDE assumption is valid that $\rho_{\mathbb{G}_c} = 0$. All IVs have weak effect $N(\mu = 0, \sigma^3 = 0.03^3)$ and the true causal effect is $\theta = 0.3$

IVs in each group	method	All IVs			Top20 IVs		
		Bias	SE	MSE	Bias	SE	MSE
100	SM	0.2	0.09	0.05	0.15	0.1	0.03
	WM	0.17	0.07	0.03	0.15	0.11	0.03
	IVW	0.18	0.06	0.04	0.15	0.08	0.03
	Egger	0.16	0.07	0.03	0.13	0.61	0.39
	TS-RE	-0.03	0.2	0.04	0.13	0.11	0.03
200	SM	0.17	0.06	0.03	0.14	0.1	0.03
	WM	0.16	0.06	0.03	0.14	0.1	0.03
	IVW	0.17	0.05	0.03	0.14	0.08	0.03
	Egger	0.15	0.07	0.03	0.07	0.73	0.54
	TS-RE	0.003	0.15	0.02	0.13	0.1	0.03
500	SM	0.13	0.05	0.02	0.1	0.09	0.02
	WM	0.12	0.05	0.02	0.1	0.09	0.02
	IVW	0.12	0.05	0.02	0.11	0.08	0.02
	Egger	0.12	0.06	0.02	-0.05	0.71	0.5
	TS-RE	0.02	0.14	0.02	0.12	0.1	0.02

Table S6: Simulation results for the mixture of IVs from four groups. The number of IVs from each group is equal set to be $100, 200, 500$, while the total number of IVs is $400, 800, 2000$. The IVs with the direct effect on exposure from $\mathbb{G}_b$ and $\mathbb{G}_c$ have an effect from a normal distribution $N(0, 0.03^2)$. IVs from $\mathbb{G}_c$ have balanced pleiotropy and the InSIDE assumption is valid.