Improving causal effect estimation in multi-ancestry multivariable Mendelian randomization with transfer learning

Yihe Yang and Xiaofeng Zhu
Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University

July 10, 2025

Abstract

Multivariable Mendelian randomization (MVMR) is widely used to estimate the causal effects of exposures on disease outcomes. However, its applications have been largely limited to individuals of European ancestry, due to the larger sample sizes available in European genome-wide association studies (GWAS). Although methods that jointly analyze multiple ancestries have been proposed to improve power in MR analyses, most of them have focused on univariable MR, which estimates total rather than direct causal effects, making them less suitable for disentangling the independent contributions of multiple, potentially correlated exposures. Here, we introduce MRBEE-TL, a novel MVMR method that combines transfer learning with bias-corrected estimating equations to improve power in underpowered ancestries and to assess cross-ancestry heterogeneity of disease risk factors. In simulations, MRBEE-TL consistently outperformed MR methods that relied solely on ancestry-specific GWAS data, achieving superior estimation accuracy, statistical power, and Type I error control. In real data analyses, MRBEE-TL not only identified ancestry-consistent and ancestry-specific causal effects missed by conventional approaches, but also improved power in African and East Asian ancestries.

Keywords: Genome-wide association studies, Multivariable Mendelian randomization, Multi-ancestry Mendelian randomization, Transfer learning

1 Introduction

Mendelian randomization (MR) is an epidemiological method that utilizes genetic variants as instrumental variables (IVs) to infer whether an exposure causally influences an outcome (Bowden et al., 2015, 2016; Zhu, 2021; Sanderson et al., 2022). Since the genotypes of individuals are randomly inherited from their parents and generally do not change during their lifetime, genetic variants are considered to be independent of underlying confounders and hence can be used as IVs to eliminate confounding bias. In addition, Mendelian randomization can be performed by using only genome-wide association study (GWAS) summary statistics (Hemani et al., 2018). The growing availability of GWAS summary data through public repositories such as the GWAS Catalog (MacArthur et al., 2017) and dbGaP (Mailman et al., 2007) has further contributed to the rapid development of MR research. Moreover, MR has been generalized to handle multiple exposures (known as multivariable MR (MVMR)) which enables one to estimate the direct causal effect of each exposure on an outcome (Sanderson et al., 2019; Sanderson, 2021). Existing MVMR methods include MVIVW (Bowden et al., 2016), MVMR-lasso (Rees et al., 2019), GRAPPLE (Wang et al., 2021), MVMRcML (Lin et al., 2023), SRIVW (Wu et al., 2024), and MRBEE (Lorincz-Comi et al., 2024), among many others.

Understanding the heterogeneity of disease risk factors across ancestries is important both for improving the generalizability of findings and for uncovering ancestry-specific disease mechanisms. Given that genetic architectures differ across ancestries (Verma et al., 2024), including different allele frequencies and linkage disequilibrium (LD) patterns, MR is typically performed separately within each ancestry, and the resulting causal estimates are compared across ancestries (Wang et al., 2022), rather than conducting a universal MR analysis using the trans-ancestry GWAS data. However, this ancestry-stratified strategy has limitations, such as reducing the sta-

tistical power for non-European ancestries and increasing the risk of selecting weak instruments. In addition, Hemani et al. (2025) recently demonstrated that MR is more likely to yield inconsistent causal estimates across ancestries if not performed appropriately. Such bias can arise when the MR analysis does not adequately address the LD and allele frequency differences across ancestries, valid instrument selection (Hemani et al., 2025), and potential gene-environment and gene-gene interactions (GEIs and GGIs) (Zhu et al., 2024). Thus, it remains challenging to determine whether observed discrepancies in an MR analysis genuinely reflect ancestry-specific disease mechanisms. As a result, there is a growing need for statistical methods that not only account for the aforementioned sources of bias, but also enable joint analysis across diverse ancestries, providing a unified framework to disentangle ancestry-consistent from ancestry-specific disease mechanisms.

Information obtained in European ancestry (EUR) can be borrowed for studies in non-European ancestry, as demonstrated in constructing polygenic risk scores (Ruan et al., 2022) and fine-mapping of causal variants (Yuan et al., 2024). This strategy has recently been introduced into MR as well, aiming to improve the power of causal effect estimation for non-EUR ancestries. For example, Hou et al. (2025) developed trans-ethnic Mendelian randomization (TEMR), one of the first methods designed to enhance causal effect estimation in underrepresented ancestries. However, TEMR is a univariable MR (UVMR) method that relies on the assumption that the estimated causal effects across ancestries follow a multivariate normal distribution, with the correlation matrix estimated from the Z-scores of the selected IVs. As we mentioned above, such an estimated correlation matrix can be sensitive to the selected IVs because of different LD patterns and allele frequencies across ancestral ancestries. When the sample size of the target ancestry is limited, substantial bias may be expected in the estimation of cross-ancestry genetic correlation.

IV selection is another significant challenge in MR and is considered to be one of the main

sources of spurious, inconsistent causal effect estimates across ancestries. To address this issue, Hemani et al. (2025) systematically compared multiple IV selection strategies and found that fixed effects meta-analysis (FEMA) is the most effective strategy for reducing of cross-ancestry heterogeneity, as it selects regionally optimal IVs based on combined evidence across ancestries. In addition, as demonstrated by Zhu et al. (2024), the genetic effect sizes are much more consistent across ancestries when the genetic variants interacting with environmental factors are excluded. Building on this insight, Hemani et al. (2025) proposed an MR-based method to detect and remove IVs likely involved in GEI and GGI, thereby enhancing the robustness and generalizability of multi-ancestry MR studies. These improvements in IV selection have been implemented in the CAMeRa platform (Hemani et al., 2025); however, a key limitation is that CAMeRa remains primarily built upon existing UVMR methods.

While inappropriately conducted MR analysis can lead to spurious heterogeneity in causal effects across ancestries, genuine heterogeneity does exist due to differences in environmental and cultural factors, such as dietary habits, climate, and lifestyle (Peterson et al., 2019). For example, Smith et al. (2024) reported distinct multi-ancestry polygenic mechanisms of type 2 diabetes (T2D), where lipodystrophy-related polygenic risk in East Asian (EAS) is significantly elevated, contributing to increased T2D susceptibility at lower body mass index (BMI) levels compared to other ancestries. These genuine differences in disease etiology across ancestries may manifest as significant differences in causal effect estimates, and the identification of such differences is also of scientific interest in MR studies. Therefore, it is crucial to develop MR methods that not only have sufficient statistical power to detect causal effects, but also can identify heterogeneity in effect sizes across ancestries.

In this paper, we introduce MRBEE-TL, a novel MVMR method that integrates MRBEE (Lorincz-Comi et al., 2024) with transfer learning (Li et al., 2022). In MRBEE-TL, the EUR

ancestry is considered as the source ancestry, and EAS, African (AFR), or other non-EUR ancestries as target ancestries. By leveraging large-scale GWAS data from the source ancestry, MRBEE-TL enhances the statistical power of causal effect estimates in underrepresented groups when effects are consistent across ancestries, while ensuring unbiased estimation when the causal effects are ancestry-specific. In addition, MRBEE-TL adopts FEMA for IV selection while excluding variants likely involved in GGIs or GEIs to avoid bias from interaction effects. In our real data analysis of five stroke subtypes, MRBEE-TL not only observed previously reported ancestry-consistent causal effects (Hemani et al., 2025), but also uncovered new biologically plausible causal exposure-outcome pairs of EAS and AFR through transfer learning. In addition, we find that although the dominant causal exposures vary by stroke subtype, their estimated effect sizes remain consistent across ancestries. These findings suggest that discrepancies in observed causal effects may partly reflect differences in stroke subtype composition across ancestry cohorts, rather than true heterogeneity in causal effect sizes, consistent with previous observations by Hemani et al. (2025).

2 Method

2.1 Multivariable Mendelian randomization model

Let $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{im})^\top$ be an $(m \times 1)$ genotype value vector of m genetic variants, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ be an $(p \times 1)$ vector representing p exposures, and y_i be an outcome. Here, m is the number of IVs, which is usually the number of independent loci with p-values reaching the genome-wide significant level. The MVMR model is below:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta} + \boldsymbol{g}_i^\top \boldsymbol{\gamma} + v_i, \quad \boldsymbol{x}_i = \mathbf{B}^\top \boldsymbol{g}_i + \boldsymbol{u}_i \tag{1}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)^\top$ is an $(m \times p)$ matrix of genetic effects on exposures with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^\top$ being an $(p \times 1)$ vector, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ is an $(p \times 1)$ vector of causal effects of the p exposures on the outcome, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^\top$ is an $(m \times 1)$ vector, known as the horizontal pleiotropy, and $\boldsymbol{u}_i$ and v_i are the noise terms. We assume that the total number of exposures p is fixed and the causal effect $\boldsymbol{\theta}$ is fixed and bounded. The genetic variant g_{ij} is standardized so that $\mathrm{E}(g_{ij}) = 0$ and $\mathrm{var}(g_{ij}) = 1$, and all IVs are in linkage equilibrium (LE), i.e., $\mathrm{cov}(g_{ij}, g_{ik}) = 0$ for $j \neq k$. Substituting for $\boldsymbol{x}_i$ in (1), we obtain the reduced form: $y_i = \boldsymbol{\alpha}^\top \boldsymbol{g}_i + \boldsymbol{\theta}^\top \boldsymbol{u}_i + v_i$, where $\boldsymbol{\alpha}_i = (\alpha_1, \ldots, \alpha_m)^\top$, and the genetic effects of exposures and outcome have the following relationship:

$$\alpha_i = \boldsymbol{\beta}_j^\top \boldsymbol{\theta} + \gamma_i. \tag{2}$$

In practice, neither the outcome genetic effects α_j nor the exposure-specific genetic effects $\boldsymbol{\beta}_j$ are directly observed; instead, they are usually estimated from GWAS. Let $\mathbf{y}^{[0]} = (y_1^{[0]}, \ldots, y_{n_0}^{[0]})^\top$ be the sample vector from outcome GWAS, $\boldsymbol{x}^{[1]} = (x_1^{[1]}, \ldots, x_{n_1}^{[1]})^\top, \ldots, \boldsymbol{x}^{[p]} = (x_1^{[p]}, \ldots, x_{n_p}^{[p]})^\top$ be the sample vectors of the 1st, $\ldots$, pth exposure GWAS cohorts, and $\mathbf{G}^{[0]} = (g_{ij}^{[0]})_{n_0 \times m}, \ldots, \mathbf{G}^{[p]} = (g_{ij}^{[p]})_{n_p \times m}$ be the sample matrices of m genetic variants of the outcome and 1st, $\ldots$, pth exposure GWAS cohorts. The sample size of the sth GWAS cohort is n_s, and the overlapping sample size between the sth and the kth GWAS cohorts is n_{sk}. The estimates of $\hat{\alpha}_j$ and $\{\hat{\beta}_{js}\}$ $(1 \leq s \leq p)$ are thus yielded by $\hat{\alpha}_j = \boldsymbol{g}_j^{[0]\top} \boldsymbol{y}^{[0]}/n_0$ and $\hat{\beta}_{js} = \boldsymbol{g}_j^{[s]\top} \boldsymbol{x}^{[s]}/n_s$, where $\mathrm{var}(\hat{\alpha}_j) \approx \sigma_{yy}/n_0$ and $\mathrm{var}(\hat{\beta}_{js}) \approx \sigma_{x_s x_s}/n_s$. Then the GWAS summary data are formed by $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_m)^\top$, $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jp})^\top$, $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_m)^\top$, the related standard errors (SEs), the p-values, and SNPs information. The summary-statistics-based MR model refers to:

$$\hat{\alpha}_j = \hat{\boldsymbol{\beta}}_j^\top \boldsymbol{\theta} + \gamma_j + \varepsilon_j, \tag{3}$$

where $\hat{\alpha}_j$ and $\hat{\boldsymbol{\beta}}_j$ are estimated from outcome and exposure GWAS for jth IV, and ε_j represents the residual of this regression model. In addition, the GWAS summary data are usually standardized, i.e., $\hat{\alpha}_j$ and $\hat{\beta}_{js}$ by $\hat{\alpha}_j/\text{se}(\hat{\alpha}_j)/\sqrt{n_0}$ and $\hat{\beta}_{js}/\text{se}(\hat{\beta}_{js})/\sqrt{n_s}$, to remove the minor allele frequency effect (Lorincz-Comi et al., 2024).

2.2 Mendelian randomization using transfer learning

We introduce MRBEE-TL, a new MVMR method that combines MRBEE (Lorincz-Comi et al., 2024) with transfer learning techniques (Li et al., 2022). Generally, transfer learning aims to transfer knowledge from related but distinct datasets to improve the learning performance of a target model. In the context of MR, this involves borrowing information from a well-powered source ancestry to enhance causal effect estimation in a less-powered target ancestry. Here, the target and source MVMR models are formulated as:

$$\hat{\alpha}_{\text{target } j} = \hat{\boldsymbol{\beta}}_{\text{target } j}^{\top}\boldsymbol{\theta}_{\text{target}} + \gamma_{\text{target } j} + \varepsilon_{\text{target } j}, \quad 1 \leq j \leq m_{\text{target}} \tag{4}$$

$$\hat{\alpha}_{\text{source } t} = \hat{\boldsymbol{\beta}}_{\text{source } t}^{\top}\boldsymbol{\theta}_{\text{source}} + \gamma_{\text{source } t} + \varepsilon_{\text{source } t}, \quad 1 \leq t \leq m_{\text{source}} \tag{5}$$

where the subscripts "target" and "source" refer to the target and source ancestry, respectively. In particular, MRBEE-TL does not assume that the source and target models share the same set of causal instruments, and hence we use distinct subscripts, j and t, and different total numbers of causal IVs, m_{source} and m_{source}, to characterize the respective MVMR models. For notational simplicity, we omit the subscript "target" in $\hat{\alpha}_{\text{target}t}$, $\hat{\boldsymbol{\beta}}_{\text{target}t}$, $\boldsymbol{\theta}_{\text{target}}$, m_{target} in the rest of the paper.

MRBEE-TL aims to estimate $\boldsymbol{\theta}$ more precisely by borrowing information from a source causal effect estimate $\hat{\boldsymbol{\theta}}_{\text{source}}$, which is typically obtained from a European ancestry. Specifically,

MRBEE-TL considers the following penalized minimization:

$$(\hat{\boldsymbol{\theta}}_{\mathrm{BEE-TL}}, \hat{\boldsymbol{\gamma}}_{\mathrm{BEE-TL}}) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\gamma}} \Bigg\{ \underbrace{\frac{1}{2m}\sum_{j=1}^{m}(\hat{\alpha}_j - \hat{\boldsymbol{\beta}}_j^\top \boldsymbol{\theta} - \gamma_j)^2}_{\text{main loss function}} - \underbrace{\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{W_\beta W_\beta}\boldsymbol{\theta} + \boldsymbol{\sigma}_{W_\beta w_\alpha}^\top \boldsymbol{\theta}}_{\text{correction of estimation error bias}} + \underbrace{\sum_{j=1}^{m} q_\tau(\gamma_i)}_{\text{pleiotropy selection term}} + \underbrace{\sum_{k=1}^{p} \rho_\lambda(\theta_k, 0, \hat{\theta}_{\text{source } k})}_{\text{transfer learning term}} \Bigg\}. \tag{6}$$

In this minimization, there are three penalty terms. The first term is $-\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_{W_\beta W_\beta}\boldsymbol{\theta}/2 + \boldsymbol{\sigma}_{W_\beta w_\alpha}^\top \boldsymbol{\theta}$, which is just the correction term of estimation error bias used in MRBEE (Lorincz-Comi et al., 2024). The second penalty is $\sum_j q_\tau(\gamma_i)$ with a tuning parameter τ, which is employed to identify the horizontal pleiotropy that resemble outliers (Zhao et al., 2020). In practice, lasso is one of the most well-known penalties (Tibshirani, 1996), and MR-Lasso adopts it to identify invalid IVs with large horizontally pleiotropic effects (Rees et al., 2019). Whereas, other variable selection penalties such as the minimax concave penalty (MCP) (Zhang, 2010) can also be used here.

The third penalty $\rho_\lambda(\theta, \vartheta_1, \vartheta_2)$ is the key component that enables MRBEE-TL to distinguish ancestry-consistent and ancestry-specific causal effects. This penalty, known as the bidirectional separation penalty (BDSP) (Tang et al., 2021), encourages each effect to be either zero or similar to its source estimate, whichever makes more sense given the data. Formally, it is defined as:

$$\rho_\lambda(\theta, \vartheta_1, \vartheta_2) = \min\left(\tilde{\rho}_\lambda(\theta - \vartheta_1), \tilde{\rho}_\lambda(\theta - \vartheta_2)\right), \tag{7}$$

where $\tilde{\rho}_\lambda(\theta)$ is a common variable selection penalty such as lasso, and ϑ_1 and ϑ_2 represent two directions of penalization. For the kth target causal effect θ_k, we set $\vartheta_1 = 0$ to shrink small or non-causal effects toward zero, and set $\vartheta_2 = \hat{\theta}_{\text{source } k}$ to pull ancestry-consistent effects toward the source estimate. By using BDSP, MRBEE-TL can achieve the following two goals:

(1) **Identifying ancestry-consistent effects:** When a target causal effect is consistent with the

source causal effect, i.e., $\theta_k \approx \hat{\theta}_{\text{source } k}$, BDSP encourages alignment with $\vartheta_2 = \hat{\theta}_{\text{source } k}$, improving estimation of underpower ancestry via transfer learning.

(2) **Detecting ancestry-specific effects:** When a target causal effect θ_k differs from the source causal effect $\hat{\theta}_{\text{source } k}$, BDSP avoids forcing similarity, allowing data-driven identification of ancestry-specific heterogeneity.

Note that we define exposure as ancestry-consistent in a target ancestry if its true causal effect size is identical to that in the source ancestry. In practice, when the estimated causal effect in the target ancestry is statistically consistent with the source estimate, we say that we have identified an ancestry-consistent effect.

From the Bayesian point of view, the penalty function can be interpreted as the negative log density of a prior distribution on the parameter. Specifically, we consider $p_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \propto -\log f_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$, where $f_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ denotes the prior density on the causal effect vector $\boldsymbol{\theta}_j$. When a standard Lasso penalty is used, the corresponding prior is a Laplace distribution: $f_\lambda(\theta) \propto \exp(-\lambda|\theta|)$, which encourages sparsity by placing most of the prior mass around zero (Park and Casella, 2008). In our bi-directional selection setting, we instead define a prior density proportional to

$$f_\lambda(\theta) \propto \exp\left(-\lambda \cdot \min\left(|\theta|,\ |\theta - \hat{\theta}_{\text{source}}|\right)\right), \tag{8}$$

which is centered at both 0 and the source ancestry estimate. To tune the shrinkage parameter λ, we use a data-driven criterion such as the extended Bayesian information criterion (EBIC) (Chen and Chen, 2008), which balances model fit and complexity while accounting for the large model space, allowing the method to select more informative, weak-to-moderate signals that would otherwise be overshrunk under fixed penalization. This adaptive selection improves statistical power while maintaining interpretability. Similar to the Lasso, other commonly used variable

selection penalties also admit corresponding Bayesian interpretations (Moreno et al., 2015).

It could be difficult to guarantee that the IVs used were biologically causal in practice. To address this limitation, we recommend using FEMA (Hemani et al., 2025) to select a common set of IVs strongly associated with exposures in both the source and target models. Furthermore, we assume that IVs are free from GGI and GEI, and exclude variants exhibiting potential interaction effects. We will discuss how to exclude these IVs that are likely involved in GGI or GEI later.

2.3 Implementation

MRBEE-TL iteratively updates each parameter in the minimization (6) using a profile-likelihood (Murphy and Van der Vaart, 2000). Specifically, let $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\gamma}^{(t)}$ be the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ at the tth iteration. Following the procedure of Tang et al. (2021), MRBEE-TL will construct a working parameter $\boldsymbol{\vartheta}^{(t+1)} = (\vartheta_1^{(t+1)}, \ldots, \vartheta_p^{(t+1)})^\top$ with the kth entry being

$$\vartheta_k^{(t+1)} = \begin{cases} 0, & \text{if } |\theta_k^{(t)}| < |\theta_k^{(t)} - \hat{\theta}_{\text{source } k}|, \\ \hat{\theta}_{\text{source } k}, & \text{if } |\theta_k^{(t)}| \geq |\theta_k^{(t)} - \hat{\theta}_{\text{source } k}|. \end{cases} \tag{9}$$

In other words, MRBEE-TL clusters each coordinate of $\boldsymbol{\theta}^{(t)}$ around two possible centers: 0 and the corresponding entry in the source estimate $\hat{\boldsymbol{\theta}}_{\text{source}}$. Based on the working center vector $\boldsymbol{\vartheta}^{(t)}$, we define the difference parameter $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\vartheta}^{(t)}$, which captures deviation from either 0 or ancestry-consistent causal effect, thereby allowing the algorithm to adaptively select plausible causal effect during each iteration given the kth exposure. Accordingly, the updated difference

parameter $\boldsymbol{\delta}^{(t+1)}$ is obtained by solving the following optimization problem:

$$\boldsymbol{\delta}^{(t+1)} = \arg\min_{\boldsymbol{\delta}} \left\{ \frac{1}{2m} \sum_{j=1}^{m} (\hat{\alpha}_j - \hat{\boldsymbol{\beta}}_j^\top \boldsymbol{\vartheta}^{(t+1)} - \gamma_j^{(t)} - \hat{\boldsymbol{\beta}}_j^\top \boldsymbol{\delta})^2 - \frac{1}{2} \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_{W_\beta W_\beta} \boldsymbol{\delta} \right.$$
$$\left. + (\boldsymbol{\sigma}_{W_\beta w_\alpha} - \boldsymbol{\Sigma}_{W_\beta W_\beta} \boldsymbol{\vartheta}^{(t+1)})^\top \boldsymbol{\delta} + \sum_{k=1}^{p} \tilde{\rho}_\lambda(\delta_k) \right\}, \tag{10}$$

where $(\boldsymbol{\Sigma}_{W_\beta W_\beta} \boldsymbol{\vartheta}^{(t)})^\top \boldsymbol{\delta}$ is an additional correction term of estimation error bias due to the inclusion of $\boldsymbol{\vartheta}^{(t+1)}$. That is, the introduction of the working center $\boldsymbol{\vartheta}^{(t)}$ transforms the non-common BDSP $\rho_\lambda(\theta, \vartheta_1, \vartheta_2)$ into a common variable selection penalty $\tilde{\rho}_\lambda(\delta)$, which can be efficiently handled by existing algorithms. After obtaining $\boldsymbol{\delta}^{(t+1)}$, $\boldsymbol{\theta}$ is then updated by $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\delta}^{(t+1)} + \boldsymbol{\vartheta}^{(t+1)}$. The update of $\boldsymbol{\gamma}$ follows standard procedures widely used in the literature; see, e.g., cis-MRBEE (Yang et al., 2025). Conditioning on $\boldsymbol{\theta}^{(t+1)}$, the original minimization in (6) reduces to:

$$\boldsymbol{\gamma}^{(t+1)} = \arg\min_{\boldsymbol{\gamma}} \left\{ \frac{1}{2m} \sum_{j=1}^{m} (\hat{\alpha}_j - \hat{\boldsymbol{\beta}}_j^\top \boldsymbol{\theta}^{(t+1)} - \gamma_j)^2 + \sum_{j=1}^{m} q_\tau(\gamma_i) \right\}. \tag{11}$$

For a broad class of penalty functions $q_\tau(\cdot)$, $\boldsymbol{\gamma}^{(t+1)}$ admits a closed-form expression. The detailed procedure for solving the minimization (6) is described in the **Supplementary Materials**.

Although we use lasso as an example to illustrate the principle of BDSP, we actually use SuSiE (Wang et al., 2020; Zou et al., 2022) as the penalty $\tilde{p}(\theta)$ in the minimization (10) to identify non-zero entries of $\boldsymbol{\delta}$. The prior distribution used in SuSiE can be interpreted as the spike-and-slab distribution (Wang et al., 2020), which has been shown to outperform lasso in terms of unbiasedness and model selection consistency (Ročková and George, 2018). In SuSiE, although multiple tuning parameters are available, we only tune the number of single effects L, which we refer to as λ in $\tilde{\rho}_\lambda(\cdot)$, and use the default settings for the remaining parameters. However, because SuSiE typically assumes a small number of single effects L, it is not well suited for selecting horizontal pleiotropy. Therefore, we turn to use MCP (Zhang, 2010) as $q_\tau(\gamma)$

in the minimization (11). In practice, MRBEE-TL adopts the EBIC (Chen and Chen, 2008) to select both the optimal number of single effects L in SuSiE and the tuning parameter τ in MCP.

To estimate the covariance matrix of $\hat{\boldsymbol{\theta}}_{\text{BEE-TL}}$, MRBEE-TL uses a selection stability method (Meinshausen and Bühlmann, 2010), where a $[0.5m]$ number of the independent IVs are resampled without replacement. According to Meinshausen and Bühlmann (2010), the sample size of $[0.5m]$ is chosen as it resembles most closely the bootstrap while allowing computationally efficient implementation, and our previous work (Yang et al., 2024) shows that selection stability works well for estimating the covariance matrix of estimates. Based on the resampled IVs, MRBEE-TL re-estimates the causal effects using the optimal tuning parameters selected by EBIC. The final covariance matrix is computed based on the empirical variation across the bootstrap replicates. In particular, for estimates deemed consistent with the source ancestry, the final covariance matrix incorporates the covariance of the source estimates; that is,

$$\text{cov}(\hat{\boldsymbol{\theta}}^{\mathcal{M}_\theta}_{\text{BEE-TL}}) = \text{cov}(\hat{\boldsymbol{\delta}}^{\mathcal{M}_\theta}_{\text{BEE-TL}}) + \text{cov}(\hat{\boldsymbol{\theta}}^{\mathcal{M}_\theta}_{\text{source}}), \tag{12}$$

where $\mathcal{M}_\theta$ is an index set of exposures whose effect sizes are consistent between source and target population, $\text{cov}(\hat{\boldsymbol{\delta}}^{\mathcal{M}_\theta}_{\text{BEE-TL}})$ is the bootstrap-estimated covariance matrix of the differences, and $\text{cov}(\hat{\boldsymbol{\theta}}^{\mathcal{M}_\theta}_{\text{source}})$ is the covariance matrix of the corresponding source estimate. In other words, the uncertainty of the source estimate is jointly incorporated into MRBEE-TL. In addition, we recommend applying a lenient significance threshold (e.g., P<0.05), below which the causal effect estimate in the source ancestry is forcibly set to zero. This helps avoid misleading interpretations of ancestry-consistent effects that may arise from weak or non-significant signals. See the **Supplementary Materials** for implementation details.

2.4 Simulation

We conducted simulation studies to evaluate the performance of MRBEE-TL. We generated GWAS summary statistics for $p = 10$ exposures, of which three were designed to have nonzero causal effects in the source (EUR) and target (EAS) ancestries. Besides, we considered three configurations for the true causal effect vectors: (i) fully ancestry-consistent effects, where $\boldsymbol{\theta}_{\text{EAS}} = \boldsymbol{\theta}_{\text{EUR}} = (0.5, 0, 0.15, 0, 0.1, 0, 0.1, 0, 0, 0)^\top$; (ii) partially ancestry-consistent effects, where $\boldsymbol{\theta}_{\text{EAS}} = (0.5, 0, 0.25, 0, 0.1, 0, 0, 0, 0, 0)^\top$ and $\boldsymbol{\theta}_{\text{EUR}} = (0.5, 0, 0.15, 0, 0.1, 0, 0.1, 0, 0, 0)^\top$, which differed in the **3rd** and **7th** entries. That is, "fully ancestry-consistent" refers to cases where the effect sizes are identical across the two ancestries, whereas "partially ancestry-consistent" indicates that a subset of exposures have different effect sizes. We fixed the sample sizes at $n_{\text{EUR}} = 400{,}000$ and the number of IVs at $m_{\text{EUR}} = 2{,}000$ and considered $m_{\text{EAS}} = 500$ or 250 when $n_{\text{EAS}} = 100{,}000$ or $50{,}000$, respectively. Likewise, we set the trait heritabilities at $h^2_{\text{EUR}} = 0.10$ and changed $h^2_{\text{EAS}} = 0.025$ or 0.0125 when $n_{\text{EAS}} = 100{,}000$ or $50{,}000$, respectively. The overlap proportions of exposure and outcome GWAS samples are all 95%. We varied the proportion of horizontal pleiotropy to be 0%, 5%, or 10% and fixed the pleiotropy-to-exposure variance ratio at 1:2. Each scenario was repeated for 500 simulation replicates. See details in **Supplementary Materials**.

We compared MRBEE-TL against several existing MVMR methods, including the multivariable versions of MR-Median (Bowden et al., 2016), MR-Lasso (Rees et al., 2019), MRcML-BIC (Lin et al., 2023), and MRBEE (Lorincz-Comi et al., 2024). We did not include MRcML-DP in our comparison due to its high computational cost, and we primarily assessed the performance of MRcML-BIC in terms of unbiasedness rather than Type-I error control. In addition, we did not consider non-robust MVMR methods such as IVW (Burgess et al., 2013) and MR-Egger

(Bowden et al., 2015), as they are susceptible to bias under horizontal pleiotropy. For traditional MVMR methods, we directly input the GWAS summary statistics from the EAS ancestry. In contrast, for MRBEE-TL, we first estimated the causal effects in the EUR ancestry using MRBEE, applied hypothesis testing to zero out non-significant entries, and used the resulting vector as the source estimate $\hat{\boldsymbol{\theta}}_{\text{source}}$ in combination with the EAS summary data. We evaluated method performance across four metrics: (1) root mean square error (RMSE), calculated as $\|\hat{\boldsymbol{\theta}}_{\text{EAS}} - \boldsymbol{\theta}_{\text{EAS}}\|_2$; (2) unbiasedness, visualized through boxplots; (3) coverage frequency of the true non-zero causal effect estimates; and (4) Type-I error rate among the null effect estimates.

2.5 Stroke analysis

We applied MRBEE-TL to perform MVMR analyses across five stroke subtypes: all strokes including both ischemic and hemorrhagic strokes (AS), all ischemic stroke (AIS), large-artery atherosclerotic stroke (LAS), cardioembolic stroke (CES), and small vessel stroke (SVS), where the GWAS summary data are provided by Mishra et al. (2022). The exposures included atrial fibrillation (AF), serum albumin (ALB), BMI, drinks per week (DRNK), HDL cholesterol (HDL-C), LDL cholesterol (LDL-C), systolic blood pressure (SBP), smoking initiation (SMK), triglyceride (TG), and T2D, all of which are recognized as common cardiometabolic risk factors across the outcomes of interest. GWAS summary statistics were derived from Graham et al. (2021); Sinnott-Armstrong et al. (2021); Aragam et al. (2022); Saunders et al. (2022); Miyazawa et al. (2023); Chen et al. (2023); Suzuki et al. (2024); Keaton et al. (2024); Verma et al. (2024). Detailed data sources are listed in Table S1.

We applied SBayesRC (Zheng et al., 2024) to impute missing variants using reference panels containing approximately 7.3 million SNPs for EUR, 6.0 million SNPs for AFR, and 4.9 million SNPs for EAS. The covariance matrices of estimation errors were estimated from genome-wide

SNPs with statistically insignificant associations (Zhu et al., 2015). Next, we followed the idea of FEMA to select IVs (Hemani et al., 2025). Specifically, for each variant, we computed the χ^2-statistic in combining association evidence of EUR, EAS, and AFR, and the corresponding p-values were calculated using a χ^2-distribution with 30 degrees of freedom. In C+T, we used an alternative LD reference panel of 9,680 European individuals randomly selected from the UK Biobank. Although EAS and AFR individuals were not included, this is consistent with the principle of FEMA described by Hemani et al. (2025), which selects variants with strong effects in all ancestries. Furthermore, IVs were clumped using `PLINK` (Purcell et al., 2007) with a 1 Mb window, a threshold of P<5E-8, and a threshold of r^2<0.001. Moreover, including IVs interacted with an environment factor in MR will lead to inconsistent causal effect estimation across ancestries, as suggested by Hemani et al. (2025). Hence, we further excluded the IVs that were in the loci with GEIs (Sung et al., 2018; Zhu et al., 2024). This resulted in 3,082 IVs used in multi-ancestry MVMR analysis.

We consider TEMR as a comparator. We retained the FEMA-based IV selection strategy but applied an exposure-wise joint χ^2 test with 3 degrees of freedom. For clumping and thresholding, we used `PLINK` with a 1 Mb window, a genome-wide significance threshold of P<5E-8, and an LD threshold of r^2<0.001. We also excluded IVs that overlap with known GEI loci in an exposure-wise manner.

3 Results

3.1 Simulation Results

As shown in Figure 1, MRBEE-TL consistently produced lower RMSEs than alternative methods, providing an overall summary of its superior performance. This improvement was at-

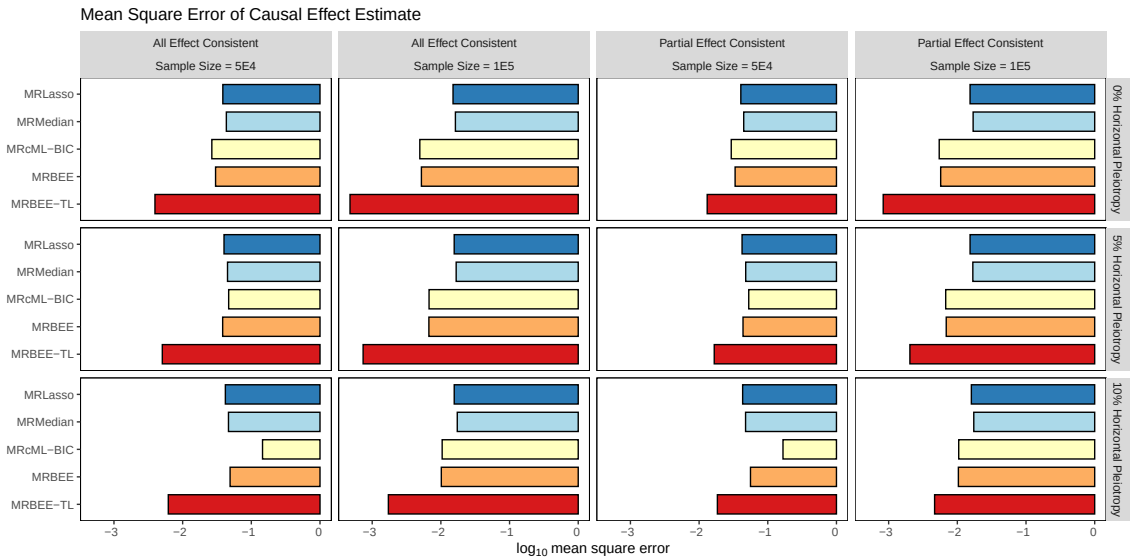

Figure 1: The logarithm of root mean square error (RMSE) of causal effect estimates under full or partial ancestry-consistent effect sizes and different degrees of horizontal pleiotropy. Each panel corresponds to one simulation setting defined by sample size, proportion of horizontal pleiotropy (0%, 5%, or 10%), and whether all or only part of the causal effects are consistent from the source to the target ancestry. Compared methods include MRBEE-TL (proposed), MRBEE, MRcML-BIC, MR-Lasso, and MR-Median. The RMSE is computed as $\|\hat{\boldsymbol{\theta}}_{\text{EAS}} - \boldsymbol{\theta}_{\text{EAS}}\|_2$. Boxplots summarize the results from 500 simulation replicates per setting.

tributable to two key mechanisms:(1) the incorporation of variable selection to shrink the zero effect estimates to exactly zero, and (2) the use of transfer learning to enforce the ancestry-consistent coordinates of $\hat{\boldsymbol{\theta}}_{\text{BEE-TL}}$ to be exactly the coordinates of $\hat{\boldsymbol{\theta}}_{\text{source}}$. In addition, MRBEE and MRcML-BIC performed similarly, reflecting their shared ability to mitigate estimation error bias through bias correction.

Figure 2 presents boxplots of estimated causal effects for four representative components (θ_1, θ_3, and θ_7) under varying degrees of horizontal pleiotropy and whether or not the causal effects are fully consistent. For θ_1 that had a relatively large effect size and were fully ancestry-consistent, MRBEE-TL achieved substantially lower uncertainty compared to alternative methods. This reflected its ability to correctly recognize these effects as ancestry-consistent and thus borrow information from the source estimate, which was based on a much larger sample size. In contrast, for θ_3 and θ_7, shown in Figure 2C and 2E, MRBEE-TL tended to incorrectly classify some

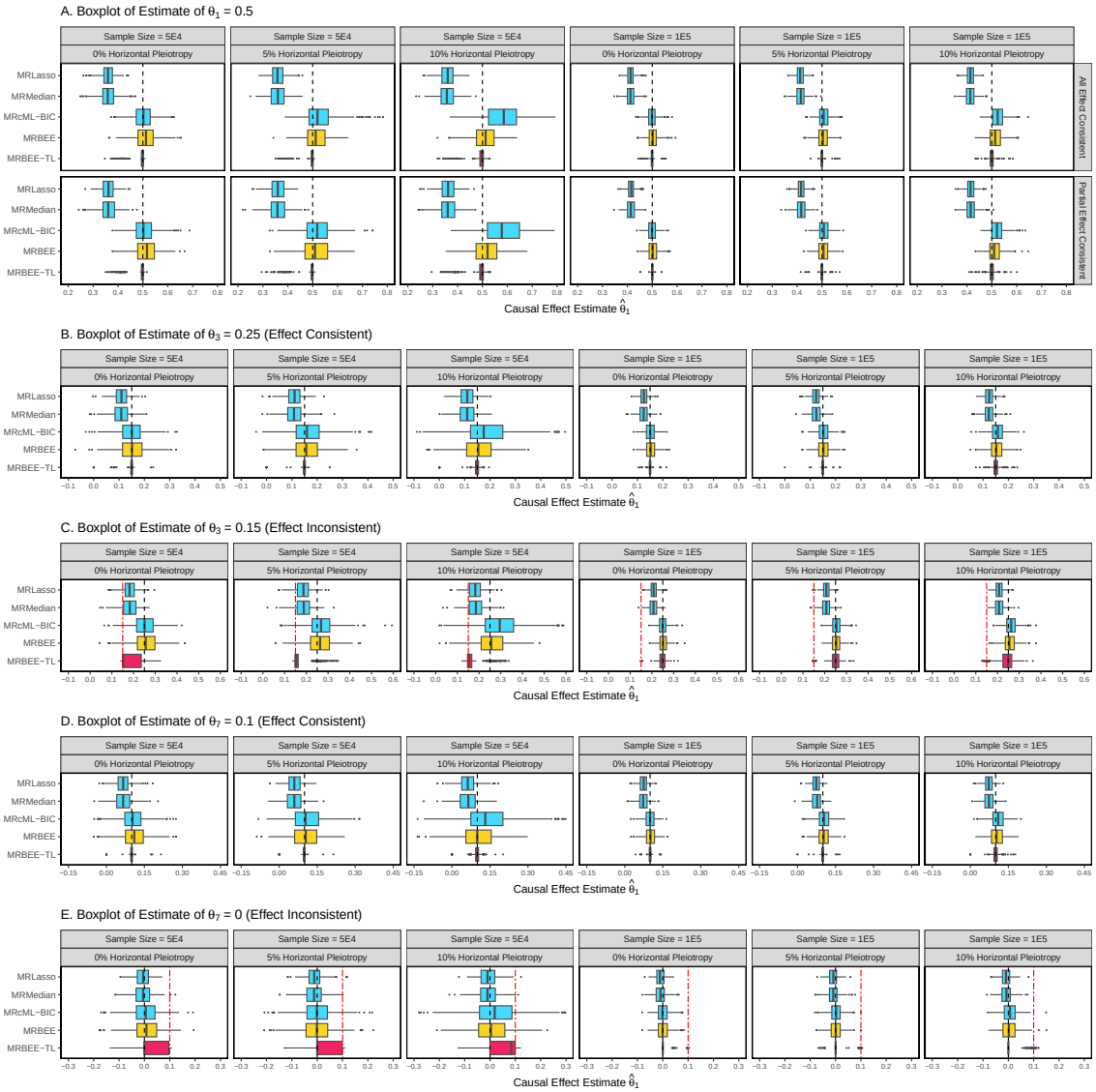

Figure 2: Boxplots of estimated causal effects under full or partial ancestry-consistent effect sizes and different degrees of horizontal pleiotropy. Each panel presents simulation results for one causal component (θ_1, θ_3, or θ_7), under different sample sizes and pleiotropy levels. **A**, **B** and **D** correspond to causal effects that were fully consistent from EUR to EAS ancestries, whereas **C** and **E** represent inconsistent effects. The black solid lines indicate the true causal effect in the EAS ancestry, and the red dashed lines (in **C** and **E**) denote the corresponding effect estimated from the EUR ancestry, shown only for inconsistent cases. MRBEE-TL, MRBEE, MRcML-BIC, MR-Median, and MR-Lasso were compared. All results were based on 500 replicates per setting.

ancestry-specific effects as ancestry-consistent, resulting in over-shrinking toward the EUR estimates and upward bias, when the EAS sample size was small (N = 5E4). This issue became more pronounced with increasing levels of horizontal pleiotropy. Fortunately, this misclassification error dropped sharply when the sample size increased to N = 1E5, suggesting that MRBEE-TL

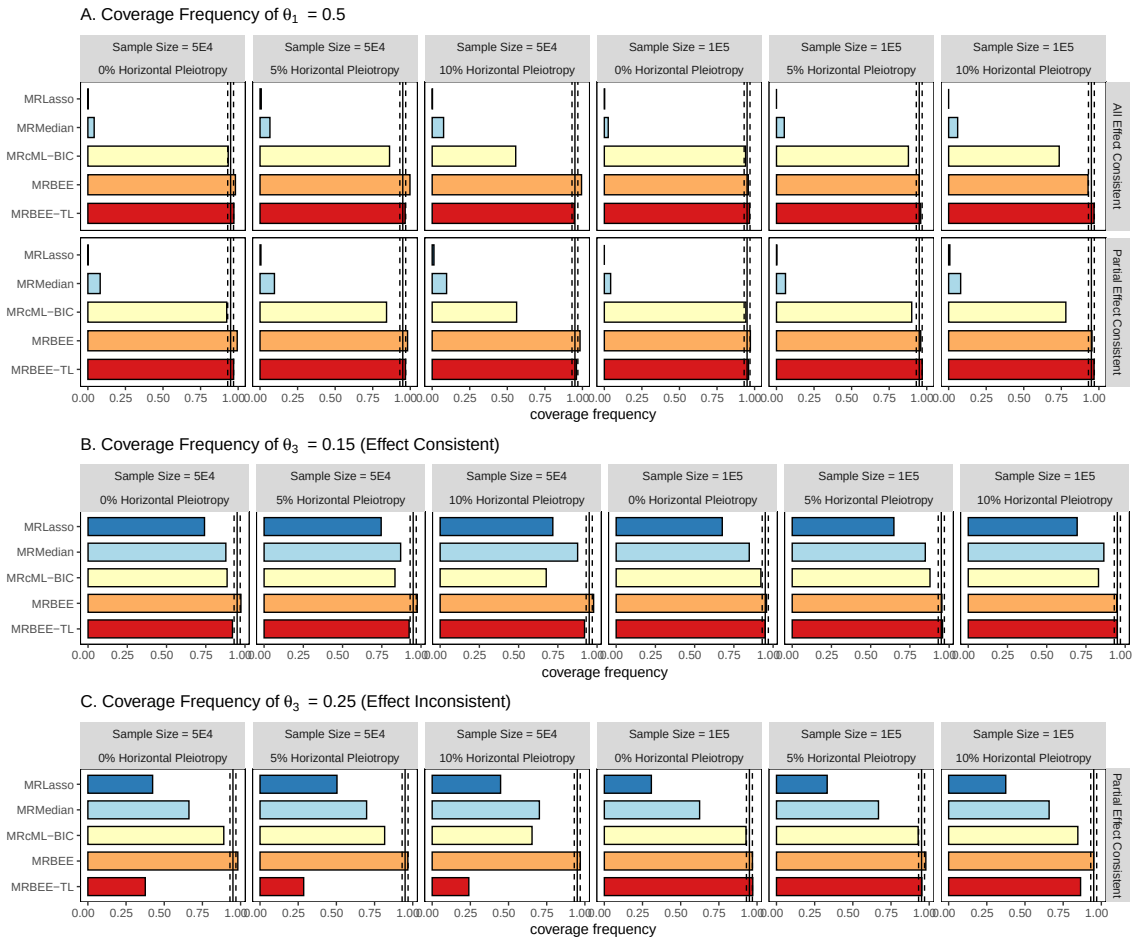

Figure 3: Coverage frequency of 95% confidence intervals for causal effect estimates. Each panel shows the empirical coverage frequency across 500 simulation replicates for selected causal effects (θ_1, θ_3, and θ_7) under varying sample sizes and levels of horizontal pleiotropy. **A** and **B** represent ancestry-consistent effects, while **C** shows a ancestry-specific effect. A well-calibrated method is expected to achieve a coverage frequency close to the nominal level of 0.95. Methods compared include MRBEE-TL, MRBEE, MRcML-BIC, MR-Median, and MR-Lasso. Deviation from the 0.95 line indicates under- or over-coverage due to misclassification of transferability or uncorrected estimation error bias.

was able to more reliably distinguish ancestry-consistent from ancestry-specific effects in larger GWAS studies. On the other hand, MRcML-BIC also exhibited visible bias when the proportion of horizontal pleiotropy was large, consistent with patterns previously observed in Lorincz-Comi et al. (2024). Notably, this bias for MRcML-BIC only emerged under high sample-overlap scenarios, a setting particularly relevant for non-European GWAS, where exposures and outcomes are often measured in the same cohort or biobank. As expected, MR-Lasso and MR-Median suffered from severe estimation error bias across all scenarios.

Figure 3 reports the coverage frequency of 95% confidence intervals for representative causal effects across simulation settings. Coverage frequency refers to the proportion of replicates in which the true effect falls within the 95% confidence interval, and a well-calibrated method should yield coverage near 0.95. MRBEE-TL generally maintained good coverage across most scenarios. However, as the proportion of horizontal pleiotropy increased, coverage declined slightly, most notably in Figure 3C under 10% pleiotropy (last column). When the sample size was small (N = 5E4), MRBEE-TL sometimes misclassified ancestry-specific effects as ancestry-consistent. This led to low coverage in the first three columns of Panel C, where the frequency dropped to around 0.25. MRBEE consistently showed over-coverage across all scenarios, consistent with previous observations (Wu et al., 2024). As expected, MR-Lasso and MR-Median had substantially lower coverage throughout, likely due to uncorrected estimation error bias.

Figure 4 evaluates the statistical properties of the compared methods in terms of Type-I error rates and power. Figure 4A and 4C report the empirical rejection frequencies under the null hypotheses $H_0 : \theta_2 = 0$ and $H_0 : \theta_7 = 0$, respectively, while Figure 4B shows the power for detecting a weak but nonzero causal effect ($\theta_7 = 0.1$) and is consistent between EUR and EAS. Overall, MRBEE-TL controlled Type-I errors effectively across most scenarios. Notably, since MRBEE-TL used variable selection to enforce sparsity, it could achieve Type-I error rates well below the nominal 0.05 level in many settings. However, in Figure 4C, where θ_7 was ancestry-specific, MRBEE-TL showed inflated Type-I error rates when the sample size was small (N = 5E4), particularly under higher levels of horizontal pleiotropy. This reflected the occasional misclassification of ancestry-specific effects as ancestry-consistent in sample-size-limited settings. MRBEE again showed overconservative behavior, with Type-I error rates below 0.05 in many cases, consistent with earlier observations (Wu et al., 2024). MR-Median and MR-Lasso exhibited inflated Type-I errors in nearly all settings, which was likely due to uncorrected estimation

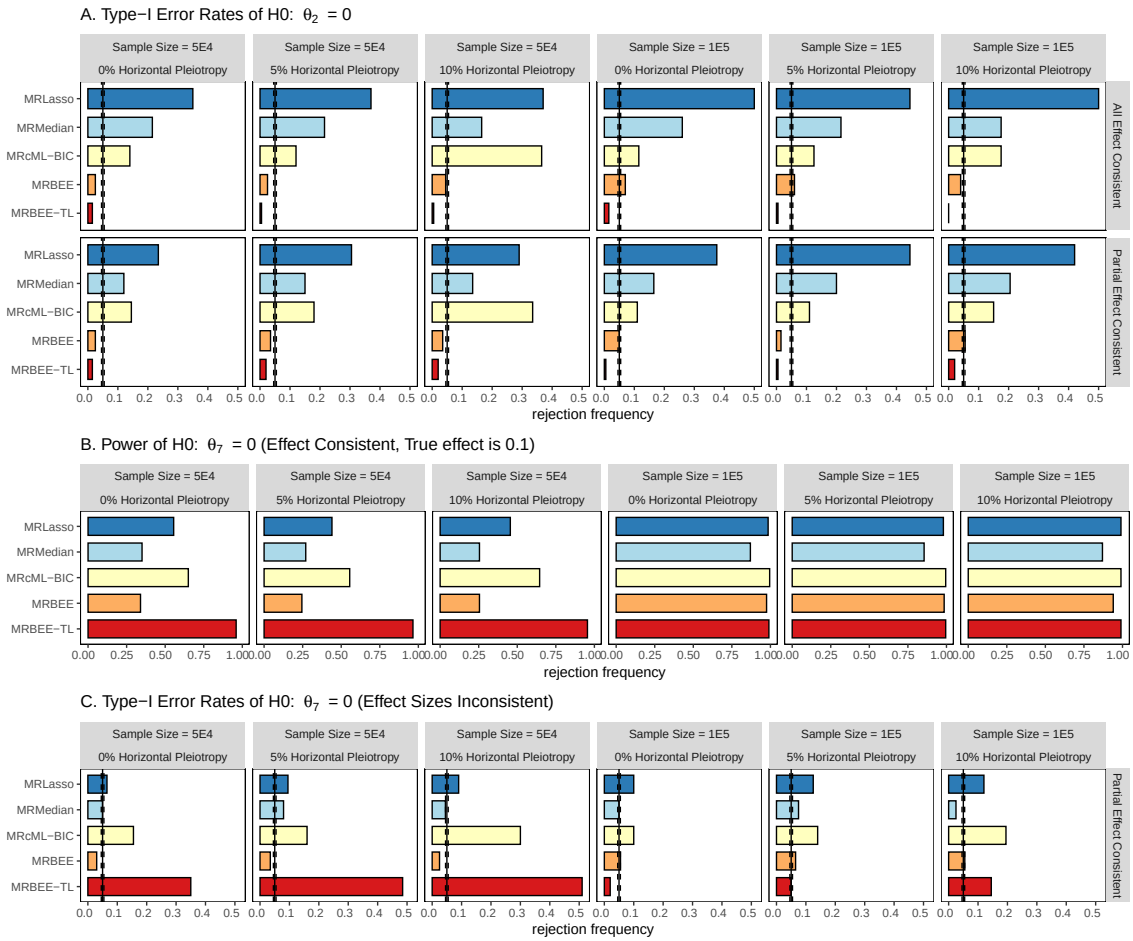

Figure 4: Type-I error rates and power under different simulation settings. Each panel reports the empirical rejection frequency across 500 replicates under varying sample sizes and levels of horizontal pleiotropy. **A** shows Type-I error rates for testing $H_0 : \theta_2 = 0$; **B** presents the power for detecting a weak but ancestry-consistent causal effect $\theta_7 = 0.1$; and **C** evaluates Type-I error rates for testing $H_0 : \theta_7 = 0$, where the effect is population-specific. The horizontal reference line corresponds to the nominal significance level of 0.05. Compared methods include MRBEE-TL, MRBEE, MRcML-BIC, MR-Median, and MR-Lasso. MRBEE-TL employed variable selection and nonparametric bootstrap inference, allowing for effective control of Type-I errors in most scenarios while also achieving superior power when effects were ancestry-consistent.

error bias, which will raise upward bias when the sample-overlap proportion is large (Lorincz-Comi et al., 2024). Figure 4B further illustrates that when the true causal effect was small but ancestry-consistent, MRBEE-TL achieved substantially higher power than the other methods. This gain was especially visible when N = 5E4, as MRBEE-TL leveraged information from the EUR ancestry to enhance signal detection. MRBEE and MRcML-BIC also achieved moderate power, while MR-Median and MR-Lasso underperformed across all scenarios.

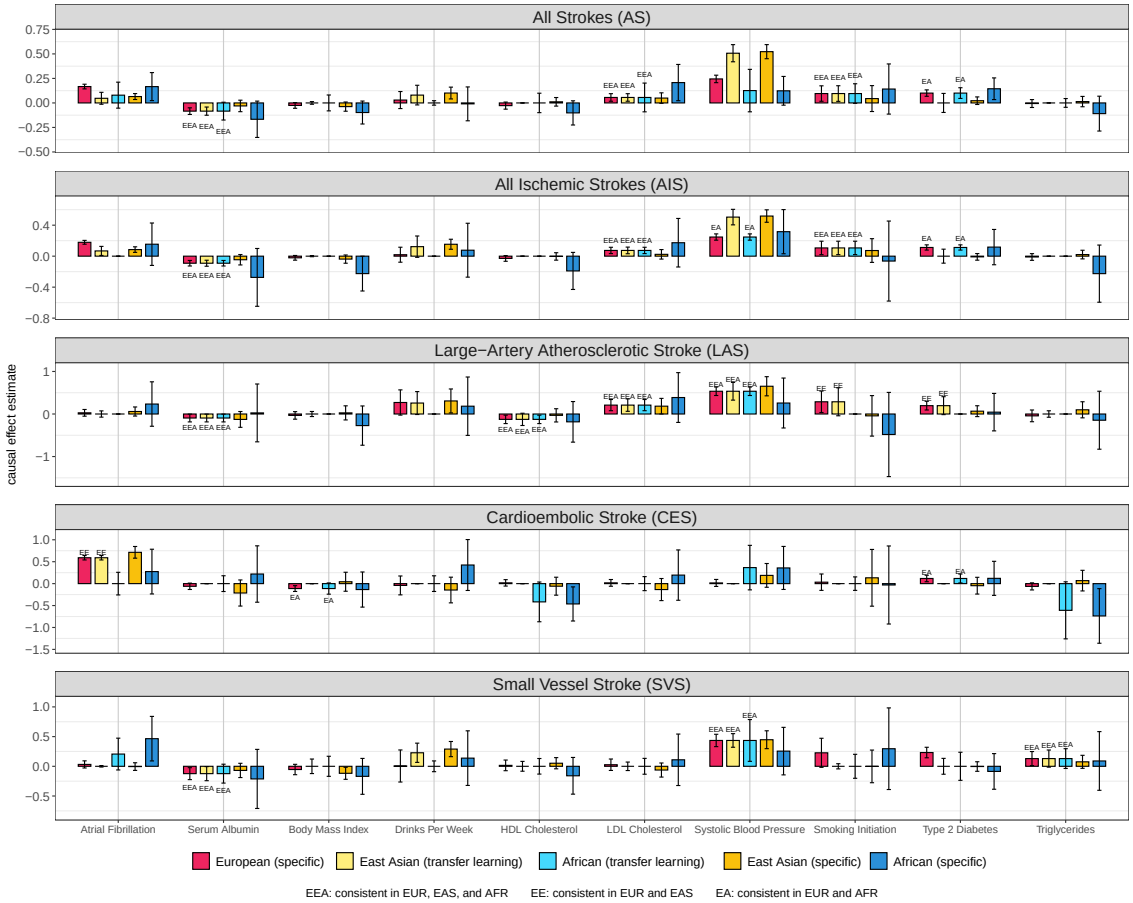

Figure 5: Causal effect estimates of selected exposures on five stroke outcomes across European, East Asian, and African ancestries. Each exposure-outcome pair is evaluated using five settings: European (specific), East Asian (transfer learning), African (transfer learning), East Asian (specific), and African (specific) (by using the original MRBEE). Bar heights represent point estimates of causal effects, with error bars indicating 95% confidence intervals. Annotations "EA", "EE", and "EEA" above bars denote consistent causal effect estimates between European and African, European and East Asian, and all three ancestries, respectively, identified by transfer learning.

3.2 Stroke analysis results

Figure 5 displays the causal effect estimates in East Asian and African Ancestries from MRBEE-TL across five stroke subtypes, in comparison to the estimates from MRBEE using ancestry-specific GWAS data. The European causal effect estimates (source) are also present for comparison. Overall, MRBEE-TL resulted in more consistent causal estimates between ancestries than population-specific analysis. For example, MRBEE-TL suggested that ALB was protective and SBP, LDL-C, and SMK are risk factors for AS and AIS in East Asian and African ancestries,

which are consistent with European ancestry. In contrast, SMK was not significant in EAS and AFR in ancestry-specific analysis. The causal risk factors varied considerably across the three ischemic stroke subtypes. For LAS, LDL-C and SBP emerged as the primary risk contributors, while ALB and HDL-C were protective factors. Regarding CES, AF was identified as the primary risk factor. T2D was a risk factor in European and East Asian ancestries. The causal factors of CES are less consistent across the three ancestries. For SVS, SBP and TG were consistent risk factors and ALB was protective in the three ancestries. These findings indicate the different causal risk components for stroke subtypes, but the magnitude of causal effects is consistent in the three ancestries for each subtype.

We observed that multiple causal signals were detected in AFR by MRBEE-TL but were missed by AFR-specific analysis, which can be attributed to the small sample sizes in AFR GWAS. Similarly, in EAS, several causal exposure-outcome pairs, including LDL-C-AS, LDLC-AIS, SMK-AS, SMK-AIS, HDL-C-LAS, and T2D-LAS, were identified only through MRBEE-TL. These results underscore the value of transfer learning in enhancing statistical power for underrepresented ancestries in GWAS and demonstrate its effectiveness as a scalable extension to traditional MVMR approaches in multi-ancestry analyses. Nevertheless, for CES, BMI was identified as a spurious protective factor in both EUR and AFR ancestries. We currently speculate that this may be due to collinearity among exposures rather than a “novel” exposure with an ancestry-consistent causal effect.

MRBEE-TL was able to identify ancestry-specific risk factors. For example, DRNK was detected as a causal factor for AIS, LAS, and SVS specifically in EAS. In addition, the estimated causal effects of SBP on AS and AIS were much larger in EAS than in EUR or AFR. In contrast, the effect sizes of SBP on LAS and SVS were consistent across ancestries. We interpret this apparent discrepancy as potentially arising from differences in the subtype composition within

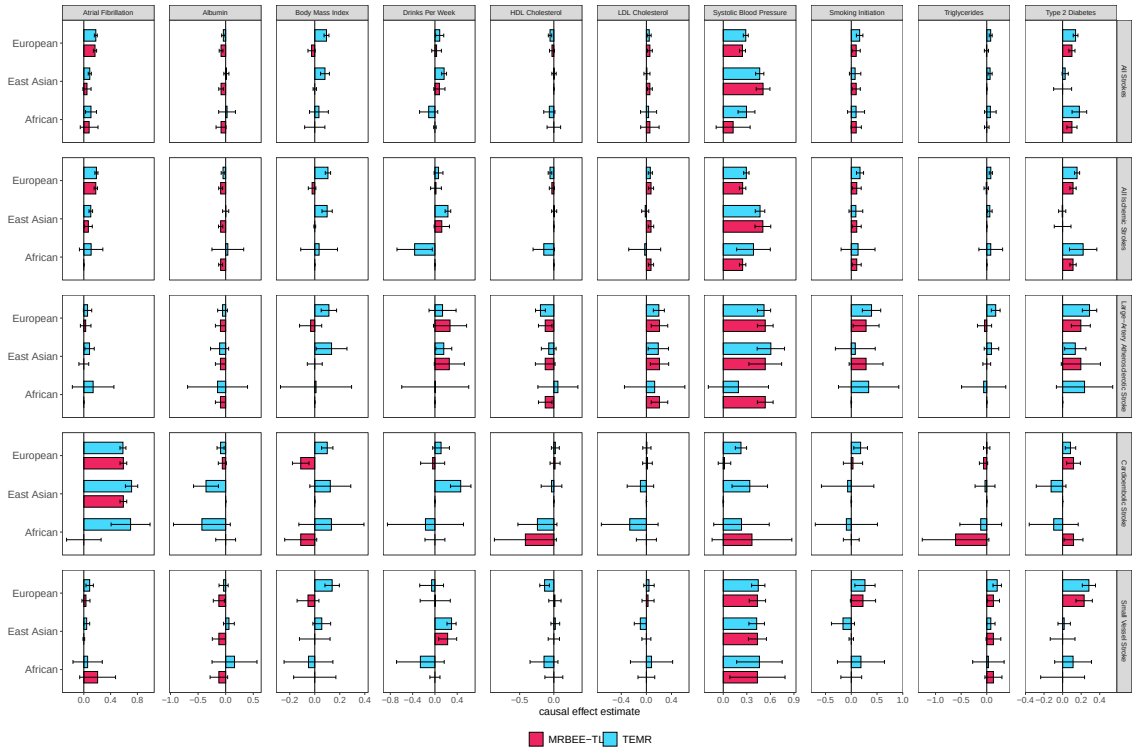

Figure 6: Comparison of causal effect estimates derived from MRBEE-TL and TEMR across three ancestries (for European, the causal effect was derived from MRBEE). Each panel represents one exposure, and within each panel, causal effects on five stroke outcomes are shown. Bars represent point estimates, with error bars indicating 95% confidence intervals. Within each exposure-outcome pair, three ancestry-specific estimates are presented from left to right: European, East Asian, and African American. Color indicates the method: MRBEE-TL (transfer learning) or TEMR (univariable baseline).

AS and AIS categories across ancestries, rather than from true heterogeneity in the underlying causal effects, which has also been discussed by Hemani et al. (2025).

Figure 6 compares the causal effect estimates obtained from MRBEE-TL and TEMR, where TEMR was performed for each exposure separately. In European ancestry, TEMR identified BMI and TG as significant risk factors and HDL-C as a protective factor for multiple stroke outcomes, while most of these risk-outcome pairs were not significant by MRBEE-TL. The main reason for this discrepancy can be attributed to that MRBEE-TL models all risk factors simultaneously, suggesting that most of the risk factors identified by TERM do not directly contribute to stroke outcomes. Consequently, TEMR frequently identified BMI and TG as risk factors and HDL-C as protective factors in East Asian and African ancestries. For T2D, TEMR yielded relatively

larger effect size estimates than MRBEE-TL, although their 95% CIs always overlapped. In AFR, TEMR identified 7 significant causal exposure-outcome pairs, compared to 12 identified by MRBEE-TL, suggesting that MRBEE-TL is more efficient and powerful than TEMR in utilizing source ancestry information.

4 Discussion

Current multi-ancestry MVMR analyses face two major challenges. The first is how to address the limited statistical power caused by the relatively small sample sizes in non-EUR populations. The second is how to distinguish ancestry-consistent causal effects from ancestry-specific ones, which may provide important insights into ancestry-specific disease mechanisms and help inform population-tailored public health policies. To resolve these challenges, we propose MRBEE-TL, an extension of MRBEE that incorporates transfer learning to simultaneously identify ancestry-consistent, ancestry-specific, and non-causal effects. From a mathematical perspective, MRBEE-TL can outperform traditional MVMR approaches as long as a subset of the causal effects is shared between the source and target ancestries (Li et al., 2022), an assumption that is broadly supported by growing empirical evidence (Hemani et al., 2025). Furthermore, MRBEE-TL integrates bias correction and pleiotropy detection, improving robustness in the presence of sample overlap bias and horizontal pleiotropy bias frequently encountered in empirical studies. To reduce bias introduced by inappropriate IV selection, we employ FEMA for IV selection and exclude variants that exhibit GEIs and GGIs.

Our real data analysis suggests that causal factors and their corresponding effect sizes of strokes are generally consistent across ancestral populations, which is consistent with the recent findings by Hemani et al. (2025). In particular, their UVMR analysis identified LDL-C

as causal only for LAS, but MRBEE-TL indicated its consistent causal effects on AS, AIS and LAS in all three ancestries. The MRBEE-TL result likely suggests that LDL-C shares common causal pathways for AS, AIS, and LAS across ancestries. In addition, we identified multiple ancestry-consistent exposure-outcome pairs across the three ancestries, including ALB–AS, ALB–AIS, ALB–LAS, ALB–SVS, SMK–AS, SMK–AIS, HDL-C–LAS, SBP–LAS, SBP–SVS, and TG–SVS. Some effects, such as T2D–AS, T2D–AIS, and T2D–CES, were consistent across EUR and AFR, while others, such as T2D–LAS, SMK–LAS, and CES–AF, were consistent across EUR and EAS. Nevertheless, many of these effects might be undetectable in AFR-specific MR analyses due to the limited sample sizes, highlighting that MRBEE-TL enhances statistical power for underrepresented ancestries.

Beyond identifying ancestry-consistent causal relationships, MRBEE-TL also uncovered significant ancestry-specific heterogeneity. First, as also pointed out by Hemani et al. (2025), some of the observed cross-ancestry discrepancies in causal effect estimates, particularly for composite outcomes like all stroke or all ischemic stroke, may reflect differences in subtype composition rather than true heterogeneity in causal mechanisms. For instance, the estimated effects of SBP on AS and AIS were consistently smaller in EUR compared to EAS. In contrast, SBP exhibited broadly consistent effects across ancestries for LAS and SVS. Indeed, the total number of EUR cases for these specific subtypes was only 24,114, substantially less than the 73,652 AS cases and 62,100 AIS cases, which may partly explain their different contributions to overall stroke risk across ancestries. In addition, our analysis identified DRNK as a potential ancestry-specific risk factor for stroke in EAS, a finding that, to our knowledge, has rarely been discussed in the literature. Further investigation of ancestry differences in the genetic architecture of alcohol consumption and its downstream disease mechanisms may offer important insight and represent a promising direction for future research.

In the comparison between TEMR and MRBEE-TL, we observed a fundamental distinction between the two methods: MRBEE-TL is an MVMR method that estimates the direct causal effect, whereas TEMR is a UVMR method that estimates the total causal effect, which is the sum of the direct causal effect and the medication effects through other exposures. While both quantities are interpretable within a causal inference framework, they reflect different aspects of the underlying biology (Sanderson et al., 2019). In our opinion, an MVMR method could be suitable than a UVRM method if the goal is disentangling the independent contributions of multiple, potentially correlated exposures. In addition, in this comparison, MRBEE-TL identified more causal exposure–outcome pairs for AFR (12 vs. 7), possibly because its use of transfer learning allows for more effective borrowing of information from high-powered source populations.

MRBEE-TL also has several limitations. First, MRBEE-TL treats the source causal effect estimates as fixed inputs, without jointly updating them alongside the target estimates. As a result, for exposures with truly ancestry-consistent effects, MRBEE-TL does not leverage information from the target ancestry, leading to a potential loss in statistical power. A unified transfer learning framework that jointly estimates causal effects across multiple ancestries could address this limitation. Second, although MRBEE-TL might have selected several additional exposure-outcome pairs via a SuSiE-based variable selection procedure, such as the HDL-C–AS and SMK–AS pair for AFR, their 95% confidence intervals were wide and included zero, suggesting a lack of statistical significance. Further investigation with larger sample sizes is necessary to make a clear causal conclusion. In addition, for certain exposure-outcome pairs, such as SBP–SVS in AFR, the confidence intervals were noticeably wider than in EUR, despite showing consistent effect sizes. Both phenomena likely stem from the same source: in some replicates of stability selection, the exposure was not consistently classified as ancestry-consistent. Furthermore, as Hemani et al. (2025) emphasized, IV selection and other improper quality controls in multi-ancestry MR

are subject to several inherent challenges. MRBEE-TL, like other MR methods, relies on the validity of IV selection and cannot inherently correct for biases arising from suboptimal IV selection. Therefore, the ancestry-specific heterogeneity identified by MRBEE-TL should also be interpreted with caution, as it may partially reflect biases introduced during IV selection rather than true biological differences. Therefore, we recommend using FEMA for IV selection and removing variants with known GEIs when applying MRBEE-TL.

Data and code availability

The GWAS summary data used in this study can be downloaded from the "Data available" section of the related literature. The code for the analyses presented in this paper is available in the Supplemental Materials, complete with step-by-step instructions. MRBEEX package is available at `https://github.com/harryyiheyang/MRBEEX`.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT-4o in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgments

This work was supported by grants HG011052 and HG011052-03S1 (to X.Z.) from the National Human Genome Research Institute (NHGRI), USA.

References

Aragam, K. G., T. Jiang, A. Goel, S. Kanoni, B. N. Wolford, D. S. Atri, E. M. Weeks, M. Wang, G. Hindy, W. Zhou, et al. (2022). Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet. 54*(12), 1803–1815.

Bowden, J., G. Davey Smith, and S. Burgess (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol. 44*(2), 512–525.

Bowden, J., G. Davey Smith, P. C. Haycock, and S. Burgess (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol. 40*(4), 304–314.

Burgess, S., A. Butterworth, and S. G. Thompson (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol. 37*(7), 658–665.

Chen, C.-Y., T.-T. Chen, Y.-C. A. Feng, M. Yu, S.-C. Lin, R. J. Longchamps, S.-H. Wang, Y.-H. Hsu, H.-I. Yang, P.-H. Kuo, et al. (2023). Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell Genom. 3*(12).

Chen, J. and Z. Chen (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Graham, S. E., S. L. Clarke, K.-H. H. Wu, S. Kanoni, G. J. Zajac, S. Ramdas, I. Surakka, I. Ntalla, S. Vedantam, T. W. Winkler, et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. *Nature 600*(7890), 675–679.

Hemani, G., Y. Cho, A. Chong, T. Palmer, A. Mason, J. Ferguson, D. Evans, and G. Davey Smith (2025). Jointly modelling multiple ancestral populations using gwas summary data improves causal inference. *Research Square*.

Hemani, G., J. Zheng, B. Elsworth, K. H. Wade, V. Haberland, D. Baird, C. Laurin, S. Burgess, J. Bowden, R. Langdon, et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife 7*, e34408.

Hou, L., S. Wu, Z. Yuan, F. Xue, and H. Li (2025). TEMR: Trans-ethnic Mendelian randomization method using large-scale gwas summary datasets. *Am. J. Hum. Genet. 112*(1), 28–43.

Keaton, J. M., Z. Kamali, T. Xie, A. Vaez, A. Williams, S. B. Goleva, et al. (2024). Genome-wide analysis in over 1 million individuals of european ancestry yields improved polygenic risk scores for blood pressure traits. *Nat. Genet. 56*(5), 778–791.

Li, S., T. T. Cai, and H. Li (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *J. R. Stat. Soc. Ser. B Stat. Methodol. 84*(1), 149–173.

Lin, Z., H. Xue, and W. Pan (2023). Robust multivariable Mendelian randomization based on constrained maximum likelihood. *Am. J. Hum. Genet. 110*(4), 592–605.

Lorincz-Comi, N., Y. Yang, G. Li, and X. Zhu (2024). MRBEE: A bias-corrected multivariable Mendelian randomization method. *Hum. Genet. Genom. Adv. 5*(3).

MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res. 45*(D1), D896–D901.

Mailman, M. D., M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet. 39*(10), 1181–1186.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol. 72*(4), 417–473.

Mishra, A., R. Malik, T. Hachiya, T. Jürgenson, S. Namba, D. C. Posner, F. K. Kamanu, M. Koido, Q. Le Grand, M. Shi, et al. (2022). Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature 611*(7934), 115–123.

Miyazawa, K., K. Ito, M. Ito, Z. Zou, M. Kubota, S. Nomura, H. Matsunaga, S. Koyama, H. Ieki, M. Akiyama, et al. (2023). Cross-ancestry genome-wide analysis of atrial fibrillation unveils disease biology and enables cardioembolic risk prediction. *Nat. Genet. 55*(2), 187–197.

Moreno, E., J. Girón, and G. Casella (2015). Posterior model consistency in variable selection as the model dimension grows. *Stat. Sci. 30*(2), 228–241.

Murphy, S. A. and A. W. Van der Vaart (2000). On profile likelihood. *J. Am. Stat. Assoc. 95*(450), 449–465.

Park, T. and G. Casella (2008). The bayesian lasso. *J. Am. Stat. Assoc. 103*(482), 681–686.

Peterson, R. E., K. Kuchenbaecker, R. K. Walters, C.-Y. Chen, A. B. Popejoy, S. Periyasamy, M. Lam, C. Iyegbe, R. J. Strawbridge, L. Brick, et al. (2019). Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell 179*(3), 589–603.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet. 81*(3), 559–575.

Rees, J. M., A. M. Wood, F. Dudbridge, and S. Burgess (2019). Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PloS One 14*(9), e0222362.

Ročková, V. and E. I. George (2018). The spike-and-slab lasso. *J. Am. Stat. Assoc. 113*(521), 431–444.

Ruan, Y., Y.-F. Lin, Y.-C. A. Feng, C.-Y. Chen, M. Lam, Z. Guo, L. He, A. Sawa, A. R. Martin, et al. (2022). Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet. 54*(5), 573–580.

Sanderson, E. (2021). Multivariable Mendelian randomization and mediation. *Cold Spring Harb. Perspect. Med. 11*(2), a038984.

Sanderson, E., G. Davey Smith, F. Windmeijer, and J. Bowden (2019). An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int. J. Epidemiol. 48*(3), 713–727.

Sanderson, E., M. M. Glymour, M. V. Holmes, H. Kang, J. Morrison, M. R. Munafò, et al. (2022). Mendelian randomization. *Nat. Rev. Methods Primers 2*(1), 6.

Saunders, G. R., X. Wang, F. Chen, S.-K. Jang, M. Liu, C. Wang, S. Gao, Y. Jiang, C. Khunsriraksakul, J. M. Otto, et al. (2022). Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature 612*(7941), 720–724.

Sinnott-Armstrong, N., Y. Tanigawa, D. Amar, N. Mars, C. Benner, M. Aguirre, G. R. Venkataraman, M. Wainberg, H. M. Ollila, T. Kiiskinen, et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet. 53*(2), 185–194.

Smith, K., A. J. Deutsch, C. McGrail, H. Kim, S. Hsu, A. Huerta-Chagoya, R. Mandla, P. H. Schroeder, K. E. Westerman, L. Szczerbinski, et al. (2024). Multi-ancestry polygenic mechanisms of type 2 diabetes. *Nat. Med. 30*(4), 1065–1074.

Sung, Y. J., T. W. Winkler, L. de Las Fuentes, A. R. Bentley, M. R. Brown, A. T. Kraja, K. Schwander, I. Ntalla, X. Guo, N. Franceschini, et al. (2018). A large-scale multi-ancestry genome-wide study accounting for smoking behavior identifies multiple significant loci for blood pressure. *Am. J. Hum. Genet. 102*(3), 375–400.

Suzuki, K., K. Hatzikotoulas, L. Southam, H. J. Taylor, X. Yin, K. M. Lorenz, R. Mandla, A. Huerta-Chagoya, G. E. Melloni, S. Kanoni, et al. (2024). Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature 627*(8003), 347–357.

Tang, X., F. Xue, and A. Qu (2021). Individualized multidirectional variable selection. *J. Am. Stat. Assoc. 116*(535), 1280–1296.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol. 58*(1), 267–288.

Verma, A., J. E. Huffman, A. Rodriguez, M. Conery, M. Liu, Y.-L. Ho, Y. Kim, D. A. Heise,

L. Guare, V. A. Panickan, et al. (2024). Diversity and scale: Genetic architecture of 2068 traits in the va Million Veteran Program. *Science 385*(6706), eadj1182.

Wang, G., A. Sarkar, P. Carbonetto, and M. Stephens (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol. 82*(5), 1273–1300.

Wang, J., Q. Zhao, J. Bowden, G. Hemani, G. Davey Smith, D. S. Small, and N. R. Zhang (2021). Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS Genet. 17*(6), e1009575.

Wang, K., X. Shi, Z. Zhu, X. Hao, L. Chen, S. Cheng, et al. (2022). Mendelian randomization analysis of 37 clinical factors and coronary artery disease in East Asian and European populations. *Genome Med. 14*(1), 63.

Wu, Y., H. Kang, and T. Ye (2024). A more credible approach to multivariable Mendelian randomization. *arXiv preprint 2402.00307*.

Yang, Y., N. Lorincz-Comi, M. Li, and X. Zhu (2025). A multivariable *cis*-Mendelian randomization method robust to weak instrument bias and horizontal pleiotropy bias. *Brief. Bioinform. 26*(3), bbaf250.

Yang, Y., N. Lorincz-Comi, and X. Zhu (2024). Estimation of a genetic Gaussian network using GWAS summary data. *Biometrics 80*(4), ujae148.

Yuan, K., R. J. Longchamps, A. F. Pardiñas, M. Yu, T.-T. Chen, S.-C. Lin, Y. Chen, M. Lam, R. Liu, Y. Xia, et al. (2024). Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. *Nat. Genet. 56*(9), 1841–1850.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*(1), 894–942.

Zhao, Q., J. Wang, G. Hemani, J. Bowden, and D. Small (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Stat. 48*(3), 1742–1769.

Zheng, Z., S. Liu, J. Sidorenko, Y. Wang, T. Lin, L. Yengo, P. Turley, A. Ani, R. Wang, I. M. Nolte, et al. (2024). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nat. Genet. 56*(5), 767–777.

Zhu, X. (2021). Mendelian randomization and pleiotropy analysis. *Quant. Biol. 9*(2), 122–132.

Zhu, X., T. Feng, B. O. Tayo, J. Liang, J. H. Young, N. Franceschini, J. A. Smith, L. R. Yanek, Y. V. Sun, T. L. Edwards, et al. (2015). Meta-analysis of correlated traits via summary statistics from gwass with an application in hypertension. *Am. J. Hum. Genet. 96*(1), 21–36.

Zhu, X., Y. Yang, N. Lorincz-Comi, G. Li, A. R. Bentley, P. S. de Vries, M. Brown, A. C. Morrison, C. N. Rotimi, W. J. Gauderman, et al. (2024). An approach to identify gene-environment interactions and reveal new biological insight in complex traits. *Nat. Commun. 15*(1), 3385.

Zou, Y., P. Carbonetto, G. Wang, and M. Stephens (2022). Fine-mapping from summary data with the “Sum of Single Effects” model. *PLoS Genet. 18*(7), e1010299.