Mendelianization: Concentrating Polygenic Signal into a Single Causal Locus

Eric V. Strobl

Abstract

Motivation: Complex disorders such as depression and alcohol use involve numerous genetic variants, and implicated loci continue to grow with sample size. This proliferation hampers interpretability, as the mechanisms by which so many variants jointly contribute to pathophysiology remain unclear. In contrast, classical Mendelian diseases arise from a single causal locus and are easier to interpret.
Results: We introduce *Mendelianization* – an algorithm distinct from Mendelian randomization – that learns weighted combinations of outcomes so that each aggregated phenotype concentrates association at one locus. We prove that this locus is causal under four structural assumptions natural to genetic data. The method handles partial sample overlap, provides calibrated hypothesis tests, maps coefficients to interpretable scales, and quantifies the degree of Mendelianism using summary z-statistics alone. In experiments, Mendelianization enhances statistical power to detect Mendelian symptom profiles even in heterogeneous disorders like major depression, generalized anxiety, and alcohol use disorder.
Availability and Implementation: R implementation: github.com/ericstrobl/Mendelianization.

Key words: causal inference, canonical correlation analysis, GWAS summary statistics, outcome learning

Introduction

Complex diseases arise from diverse genetic and environmental risk factors. As a result, genome-wide association studies (GWAS) have identified hundreds of associated variants, with larger sample sizes only expanding the catalog.[17] However, the accumulation of associations often complicates rather than clarifies disease mechanism. Even after fine-mapping prioritizes putative causal variants within loci,[14] the genetic architecture remains dispersed across many regions and resists reductionist interpretation.[15,9] A major culprit is phenotypic heterogeneity, where complex disorders like depression manifest as different symptom constellations (e.g., low mood, anhedonia, fatigue, insomnia). Conventional GWAS collapse this diversity into a single diagnosis or composite, broadening the phenotype and inflating the number of hits without sharper insight.[5,17]

Existing multi-trait methods do not reduce this proliferation of associated loci. Structure-learning methods (e.g., Genomic-SEM,[6] PDR,[2] SHAHER,[16] FactorGo[18]) learn latent outcomes that preserve trait-variant covariance, typically yielding even more polygenic, diffuse signals rather than locus-specific effects. By contrast, power-seeking methods (e.g., HIPO,[10] metaCCA,[4] fastASSET[11]) combine traits to maximize a detection criterion such as a non-centrality statistic, identifying even more variants rather than clarifying mechanism; their nominal p-values also tend to be anti-conservative because the outcome is learned on the same data used for testing. Meanwhile, causal inference methods either refine signals within loci (multi-trait fine-mapping[1,19]) or estimate exposure-outcome relations (multi-response MR[21] and DAG methods[20]), rather than trying to identify variant-to-outcome causal effects genome-wide.

We thus advance an alternative approach that leverages multiple outcomes to simplify – not expand – the set of associated variants by learning weighted outcome combinations that approximate "Mendelianized" traits whose genetic associations concentrate in a single causal locus (Figure 1). We specifically make the following **contributions**:

1. We introduce *Mendelianization*, a new framework – distinct from Mendelian randomization – that constructs composite outcomes which concentrate associations at a single locus via a substantially modified canonical correlation analysis (CCA).

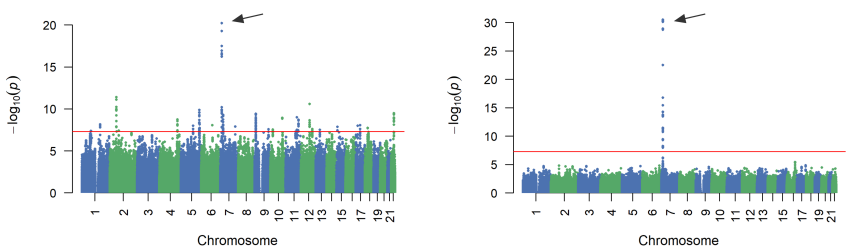


Fig. 1. Main idea. Standard outcomes – such as diagnosis or symptom severity – associate with many loci in large-sample studies of polygenic disease, obscuring underlying mechanisms (left). Mendelianization instead learns an outcome whose association converges on a single locus, yielding a solitary Manhattan plot tower (right). In practice, multiple such outcomes are inferred. The red line marks genome-wide significance (5×10^{-8}).

2. We prove that this locus contains at least one causal variant under just four structural assumptions.
3. We develop a corresponding fast hypothesis test that remains valid after outcome learning.
4. We operationalize Mendelianization at scale by addressing missing data, preserving coefficient interpretability, ensuring accurate inference, and introducing a continuous *Score of Mendelianism* to quantify achieved Mendelian properties.
5. We demonstrate that Mendelianization uncovers interpretable Mendelian-like traits in complex disorders using symptom-level summary z-statistics.[7]

Together, these advances provide a reductionist framework for identifying Mendelian-like symptom profiles dominated by a single causal locus.

Methods

One-Dimensional Canonical Correlation Analysis

One-dimensional CCA will serve as the underlying engine of Mendelianization. We thus consider summary statistics comprised of the partial correlation matrix $r \in \mathbb{R}^{q \times m}$ between q variants V and m outcomes Y. We assume that all covariates like biological sex, age, and ancestry have already been partialed out from V and Y. We focus on population-level properties for ease of presentation and address sample-level issues later. We thus have $r_{jk} = \mathrm{cor}(V_j, Y_k)$. Let r_j correspond to the row of r associated with the variant V_j.

For each V_j, we construct a composite outcome $Y\alpha_j$ by one-dimensional CCA, so that the correlation between V_j and $Y\alpha_j$ is as large as possible:

$$\arg\max_{\alpha \neq 0} \mathrm{cor}(V_j, Y\alpha). \tag{1}$$

We will also show that this procedure implicitly reduces the correlations between loci not containing V_j and $Y\alpha_j$, so that $Y\alpha_j$ resembles a Mendelian trait.

Without loss of generality, we assume that each variant has already been standardized to mean zero and unit variance. We can then simplify the above objective to:

$$\arg\max_{\alpha \neq 0} \frac{r_j \alpha}{\sqrt{\alpha^\top \Sigma \alpha}}, \tag{2}$$

where Σ corresponds to the positive definite correlation matrix of Y. Notice that the numerator maximizes alignment with V_j, while the denominator minimizes the scale of α according to the geometry of Σ.

Correlation and, likewise, the solution to Expression (1) is scale invariant, so we can impose the constraint $\alpha^\top \Sigma \alpha = 1$ and maximize the numerator $r_j \alpha$. We invoke the Cauchy-Schwarz inequality in the Σ-inner product:

$$r_j \alpha = \langle r_j \Sigma^{-1}, \alpha^\top \rangle_\Sigma \leq \|r_j \Sigma^{-1}\|_\Sigma \|\alpha^\top\|_\Sigma = \sqrt{r_j \Sigma^{-1} r_j^\top},$$

where the last equality follows because we imposed the constraint $\|\alpha^\top\|_\Sigma = \sqrt{\alpha^\top \Sigma \alpha} = 1$. The inequality becomes an equality when $\alpha = \frac{\Sigma^{-1} r_j^\top}{\sqrt{r_j \Sigma^{-1} r_j^\top}}$. Recall that the solution to Expression (2) is scale-invariant, so we set $\alpha = A r_j^\top$ with $A = \Sigma^{-1}$ from here on.

Causal Analysis

We establish that the solution of Expression (1) admits a causal interpretation; as the number of outcome variables m increases, the optimization problem achieves *Mendelianization* – the identification of a single *causal* locus – under mild assumptions. The reasoning behind why one-dimensional CCA enables Mendelianism is not obvious, so we construct the argument carefully using mathematical formalism. The main challenge lies in ensuring that the reasoning relies solely on a minimal number of realistic assumptions.

Structure of Latent Confounding

We begin by assuming that latent confounding (separate from the adjusted covariates) lies in a low-dimensional subspace of $\mathbb{R}^m$:

Assumption 1. *(Low-dimensional latent confounding) For sufficiently large m, the confounding effects are contained in a subspace $U^m \subset \mathbb{R}^m$, and there exists an integer constant $r_0 \geq 0$ such that $\dim(U^m) \leq r_0$.*

Remark. Assumption 1 effectively states that V and Y are influenced by only a small number of latent confounders, or – more generally – by any number of redundant latent confounders whose effects lie within a low-dimensional subspace. The assumption is reasonable because, after adjusting for observed covariates, residual bias in genetic data is typically dominated by a small number of processes: batch or site effects, residual population structure, and dynastic effects.[17] These influences act on many outcomes simultaneously, and thus induce an approximately low-dimensional latent confounding subspace.

Assumption 1 also implies that each r_j^m admits the decomposition:

$$r_j^m = r_j^{m\parallel} + r_j^{m\perp},$$

where $r_j^{m\parallel} \in U^m$ and $r_j^{m\perp}$ lies in the orthogonal complement of U^m with respect to the A^m-inner product, i.e., $\langle r_j^{m\parallel}, r_j^{m\perp} \rangle_{A^m} = 0$. We analyze this extended setting in detail below.

Without Latent Confounding

Let $r^m \alpha_j^m$ denote the vector of composites obtained after learning $\alpha_j^m = A^m r_j^{m\top}$. Then for $k = 1, \ldots, q$, we have:

$$\frac{1}{m} r_k^m \alpha_j^m = \frac{1}{m} r_k^m A^m r_j^{m\top} = \frac{1}{m} \langle r_k^m, r_j^m \rangle_{A^m}.$$

If $k = j$, the expression reduces to $\frac{1}{m}\langle r_j^m, r_j^m \rangle_{A^m}$, which is proportional to the numerator of Expression (2) evaluated at our optimizer for any fixed m[1].

In contrast, when $k \neq j$, the term $\frac{1}{m}\langle r_k^m, r_j^m \rangle_{A^m}$ represents an average cross-inner product. We posit that as m increases, the diversity of associations orthogonal to the confounded subspace U^m grows sufficiently to ensure asymptotic orthogonality. In particular, let $\mathcal{S}_j$ denote a finite index set of variants in linkage disequilibrium (LD) with V_j, and $j \in \mathcal{S}_j$. In other words, $\mathcal{S}_j$ corresponds to a locus in LD with V_j. Consider:

[1] Note that rescaling the inner product (i.e., $\Sigma \mapsto \Sigma/m$) does not change the maximizer of Expression (2) because it just multiplies the objective by a constant: $\frac{r_j \alpha}{\sqrt{\alpha^\top \Sigma \alpha / m}} = \sqrt{m} \frac{r_j \alpha}{\sqrt{\alpha^\top \Sigma \alpha}}$. We will thus analyze the average inner products for comparability as m grows, i.e. $\frac{1}{m}\langle \cdot, \cdot \rangle_{A^m}$.

Assumption 2. *(Outcome diversity) We have*

$$\max_{k \notin \mathcal{S}_j} \frac{1}{m} |\langle r_k^{m\perp}, r_j^{m\perp} \rangle_{A^m}| \in o_m(1).$$

Remark. In the absence of confounding, we have $r_j^m = r_j^{m\perp}$ where $r_j^{m\perp}$ represents the causal component of r_j^m. Assumption 2 then implies that, as the number of variables m increases – and hence the granularity of Y grows – the law of large numbers applies so that the average cross-alignment of causal patterns vanishes between V_j and the variants outside $\mathcal{S}_j$. As a result, in the limit $m \to \infty$, the vector $r^m \alpha_j^m$ develops a sharp Manhattan "tower": it has a single dominant peak within $\mathcal{S}_j$, while the remaining entries vanish. This assumption is reasonable because, after incorporating many granular, non-duplicative measures and whitening the outcomes ($A^m = (\Sigma^m)^{-1}$), the causal effects of variants are heterogeneous and weakly correlated across outcomes, so their mixed-sign alignments cancel out as m grows.

With Latent Confounding

Peakedness may arise spuriously from latent confounding. For example, one might learn α_j^m that captures only confounding, yielding $\frac{1}{m} r_j^m \alpha_j^m = \frac{1}{m} \langle r_j^{m\parallel}, r_j^{m\parallel} \rangle_{A^m}$. To guard against this scenario, we impose mild assumptions that rule out perfect Mendelianization in the presence of confounding, even as $m \to \infty$.

We assume that the causal and confounded components of r_j^m do not vanish:

Assumption 3. *(Signal floors) For m sufficiently large, if $l \in \mathcal{S}_j$ causes $Y\alpha_j$, then there exists a constant $\beta_c > 0$ such that $\frac{1}{m}\|r_l^{m\perp}\|_{A^m}^2 \geq \beta_c$. Moreover, there exists a constant $\beta_{on} > 0$ such that $\frac{1}{m}\|r_l^{m\parallel}\|_{A^m}^2 \geq \beta_{on}$ for every $l \in \mathcal{S}_j$ that does not cause $Y\alpha_j$.*

Remark. The first clause asserts that if $\mathcal{S}_j$ contains a causal variant V_l, then V_l must exert a non-vanishing causal effect on $Y\alpha_j$, a condition nearly implicit in the causal relationship. The second clause formalizes that all non-causal variants in LD with V_j within the block $\mathcal{S}_j$ remain confounded. This is reasonable because, when a locus's signal is non-causal, a Manhattan tower can only persist as m grows via the low-dimensional confounders of Assumption 1; LD then makes nearby variants inherit essentially the same projection onto the confounding subspace U^m, so the average squared A^m-norm does not vanish across outcomes.

Assumption 4. *(Isolation-spillover) If some $l \in \mathcal{S}_j$ causes $Y\alpha_j$, we have $\max_{k \notin \mathcal{S}_j} \frac{1}{m} |\langle r_k^{m\parallel}, r_j^{m\parallel} \rangle_{A^m}| \to 0$. On the other hand, if there does not exist $l \in \mathcal{S}_j$ that causes $Y\alpha_j$, then $\exists k \notin \mathcal{S}_j$ with $|\cos_{A^m}(r_j^{m\parallel}, r_k^{m\parallel})| \geq \gamma_j > 0$ and $\frac{1}{m}\|r_k^{m\parallel}\|_{A^m}^2 \geq \beta_{off} > 0$ for sufficiently large m.*

Remark. This assumption formalizes a widely observed empirical pattern. In the first clause, a truly causal locus shares little confounding alignment with distant variants because residual confounding remains primarily local (isolated to LD). In the second clause, if a locus' effect arises *purely* from confounding, the same confounder necessarily spills over into at least one additional variant due to low-dimensional confounding, producing measurable cross-alignment.

We are now ready to claim that one-dimensional CCA performs Mendelianization:

Theorem 1. *(Asymptotic Mendelianism) Consider Assumptions 1-4 and suppose Σ^m is positive definite for sufficiently large m. We have $\frac{1}{m} r_j^m \alpha_j^m > 0$ for sufficiently large m and $\max_{k \notin \mathcal{S}_j} |\frac{1}{m} r_k^m \alpha_j^m| \to 0$ (i.e., Mendelianism is asymptotically achieved for $Y\alpha_j^m$ at $\mathcal{S}_j$) if and only if at least one variant $l \in \mathcal{S}_j$ causes $Y\alpha_j^m$.*

Remark. We provide all proofs in the Supplementary Materials. The theorem establishes both sufficiency and necessity. Intuitively, the theorem states that maximizing the correlation between V_j and Y yields causal localization under four structural assumptions natural to genetic data. This contrasts with current practice, which fixes a target (a single α^m that does not maximize correlation) and therefore often yields many significant and potentially causal loci scattered across the genome (Figure 1 left). By comparison, the one-line estimator $\alpha_j^m = A^m r_j^{m\top}$ guarantees that all causal variants for the learned outcome $Y\alpha_j^m$ concentrate within a single LD region (Figure 1 right).

Beyond this localization guarantee, Theorem 1 also changes how we interpret CCA loadings. For each variant V_j, we estimate the canonical coefficients α_j^m and form the composite $Y\alpha_j^m$. **This composite is then held fixed and evaluated *genome-wide*** – not just at V_j. We interpret $Y\alpha_j^m$ as a Mendelian trait when the resulting Manhattan plot shows a single tower confined to an LD region containing V_j.

Finally, notice that we recover classical Mendelianism at $m = 1$ with strengthened assumptions:

Proposition 1. *(Perfect Mendelianism at finite m) Consider Assumptions 5-8 (Supplementary Materials A) with a prespecified m and suppose Σ^m is positive definite. We have $\frac{1}{m} r_j^m \alpha_j^m > 0$ and $\max_{k \notin \mathcal{S}_j} |\frac{1}{m} r_k^m \alpha_j^m| = 0$ (i.e., Mendelianism is perfectly achieved for $\mathcal{S}_j$ at m) if and only if at least one variant $l \in \mathcal{S}_j$ causes $Y\alpha_j^m$.*

Remark. The proof structure is analogous to that of Theorem 1. We emphasize Theorem 1 rather than the proposition because Theorem 1 establishes that Mendelianism can be achieved asymptotically – rather than exactly – using weaker assumptions. This distinction is practically crucial, as perfect single-locus resolution may not be attainable with any fixed m in polygenic disease.

Fast Hypothesis Testing

We now test the null of no dependence between V_j and Y, even though we learn α_j^m from data. We henceforth drop the superscript m to simplify notation. The null and alternative hypotheses correspond to:

$$\begin{aligned} \mathcal{H}_0 &: V_j \perp\!\!\!\perp Y, \\ \mathcal{H}_1 &: r_j \alpha_j > 0. \end{aligned} \tag{3}$$

A naive Pearson test is biased because we learn α_j from the data. However, we can eliminate the bias introduced from learning α_j by forming a statistic in the null geometry.

Under $\mathcal{H}_0$, with i.i.d. samples and finite fourth moments:

$$\sqrt{n}\, \widehat{r}_j := z_j \xrightarrow{d} \mathcal{N}(0, \Sigma), \tag{4}$$

where n denotes sample size[2]. The one-dimensional CCA weight is proportional to $Az_j^\top$, which motivates the on-target quadratic form $Q_j = z_j A z_j^\top$. Write the Cholesky decomposition $\Sigma = LL^\top$ and let $u \sim \mathcal{N}(0, I_m)$ so that $z_j = u^\top L^\top$. Then:

$$Q_j = (Lu)^\top A(Lu) = u^\top (L^\top AL)u = u^\top u,$$

so that $Q_j \sim \chi^2_m$. We then evaluate the right-tail probability of Q_j under $\mathcal{H}_0$ to obtain the p-value.

Note that the above χ^2-distribution result is asymptotic, with finite-sample accuracy governed by the fidelity of Σ. Crucially, we will show that a modified form of Σ can be consistently estimated using millions of variants under realistic LD and sparsity conditions, so that large n and large q asymptotics work decisively in our favor.

Operationalizing Mendelianization at Scale

Executing the above hypothesis test requires access to Σ. However, biobanks rarely include Σ in their summary statistics, and the outcomes needed to estimate Σ are only measured on partially overlapping samples. Scaling Mendelianization also introduces two challenges beyond test calibration: trait redundancy can distort the interpretation of raw canonical coefficients, and the available traits may be insufficient to approach near-perfect Mendelianism. We therefore (a) replace Σ by a quantity estimable from summary z-statistics under partial overlap, (b) show how to estimate it at scale despite LD, (c) rescale coefficients for interpretability, and (d) introduce a metric to quantify the degree of Mendelianism achieved.

From Σ to Γ Under Partial Overlap

If all traits are fully observed and $V \perp\!\!\!\perp Y$, then $\mathbb{E}[\, z_j z_j^\top \,] = \Sigma$ with i.i.d. sampling, so we can consistently estimate Σ using (approximately) independent variants. In practice, samples overlap only partially, so z-statistics and correlations are estimated from unequal sample sizes. Let $M_{a,i} \in \{0,1\}$ indicate whether $Y_{a,i}$ is observed, and define $S_a = \{i : M_{a,i} = 1\}$, $n_a = |S_a|$, and $n_{ab} = |S_a \cap S_b|$. We assume that V contains no missing values, and each Y_a has been centered so that $E(Y_a|M) = 0$ for the realized missingness pattern M. Write the second-moment matrix conditional on two outcomes being observed (i.e., over $S_a \cap S_b$) as $\Sigma^{\text{obs}}_{ab} := \mathbb{E}(Y_{a,i}Y_{b,i}\,|M_a = 1, M_b = 1, M_{-a,-b})$. The following statement holds:

Proposition 2. *Assume i.i.d. samples across individuals, $V \perp\!\!\!\perp Y$ and $V \perp\!\!\!\perp M|Y$. Then $\mathbb{E}[z_{ja}z_{jb} \mid M] = \frac{n_{ab}}{\sqrt{n_a n_b}}\,\Sigma^{\text{obs}}_{ab}$.*

Remark. The assumption $V \perp\!\!\!\perp M \mid Y$ is reasonable in biobank settings since missingness in Y is largely driven by the presence or severity of symptoms themselves (as well as factors such as sex, age, and ancestry, which we already adjusted for). The above proposition also motivates replacing Σ with the matrix Γ in Expression (2), where $\Gamma_{ab} := \frac{n_{ab}}{\sqrt{n_a n_b}}\Sigma^{\text{obs}}_{ab}$.

Replacing Σ with Γ fortunately has a clear interpretation via the outcomes $Y'_a = \sqrt{\frac{n}{n_a}}\, M_a \odot Y_a$. That is, Y'_a sets missing entries of Y_a to zero and rescales by the square root of the observed sample proportion. Then $E(Y'|M) = 0$ for the realized M, hence $E(Y'_a) = 0$. Moreover:

Proposition 3. *Assume i.i.d. samples across individuals, $p_a = \mathbb{P}(M_a = 1) > 0$, and $p_b = \mathbb{P}(M_b = 1) > 0$. We have $\mathbb{E}\Big(\frac{1}{n}\sum_t Y'_{a,t}Y'_{b,t} \mid M\Big) = \Gamma_{ab} \overset{a.s.}{\to} \frac{p_{ab}}{\sqrt{p_a p_b}}\Sigma^{\text{obs}}_{ab}$ and $\mathbb{E}\Big(\frac{1}{n}\sum_t Y'_{a,t}Y'_{b,t}\Big) \to \frac{p_{ab}}{\sqrt{p_a p_b}}\Sigma^{\text{obs}}_{ab}$ as $n \to \infty$.*

Remark. Γ_{ab} is thus the *conditional* covariance of Y' for the realized missingness pattern M. However, it asymptotically approximates the *unconditional* covariance of Y' so that dependence on M vanishes. We can therefore maximize the (asymptotically) unconditional correlation in Expression (2) by replacing Y with Y':

$$\max_{\alpha \neq 0} \frac{\widehat{r}'_j\,\alpha}{\sqrt{\alpha^\top \Gamma \alpha}}, \qquad \text{where } \widehat{r}'_j = \widehat{\text{cov}}(V_j, Y').$$

This reparameterization **does not require missingness-at-random (MAR)**[8] for Σ by targeting Γ.

Theorem 1 then holds for Y' at the population-level with Σ replaced by Γ. The scaling factors $\sqrt{n/n_a}$ of Y' also do not affect the optimization, since:

$$Y'\alpha = (Y \odot M)D^{1/2}\alpha = (Y \odot M)\gamma,$$

where $D = \text{diag}(n/n_1, \ldots, n/n_m)$ and $\gamma = D^{1/2}\alpha$. Hence:

$$\max_{\alpha \neq 0} \widehat{\text{cor}}(V_j, Y'\alpha) = \max_{\gamma \neq 0} \widehat{\text{cor}}(V_j, (Y \odot M)\gamma).$$

Even better, we never need explicit access to Y' because the sample covariance reduces to:

$$\widehat{r}'_{ja} = \frac{1}{n}\sum_{s=1}^{n} V_{j,s}Y'_{a,s} = \frac{1}{n}\sqrt{\frac{n}{n_a}} \sum_{s:M_{a,s}=1} V_{j,s}Y_{a,s} = \frac{z_{ja}}{\sqrt{n}},$$

which implies:

$$\widehat{\text{cor}}(V_j, Y'\alpha) = \frac{1}{\sqrt{n}}\frac{z_j\alpha}{\sqrt{\alpha^\top\Gamma\alpha}}.$$

except that we have replaced $\frac{1}{n}\sum_t Y'_{a,t}Y'_{b,t}$ in the denominator with its conditional counterpart Γ that has the same large sample limit per Proposition 3. We thus only need access to the z-statistics z_j and an estimate of Γ.

The above argument motivates us to maximize:

$$\arg\max_{\alpha \neq 0} \frac{z_j\alpha}{\sqrt{\alpha^\top\Gamma\alpha}}, \tag{5}$$

which is similar to Expression (2). We henceforth take $\alpha_j = \Omega z_j^\top$ as our solution, where $\Omega = \Gamma^{-1}$.

Estimating Γ with LD

Although Γ is natural to target, its empirical estimation is challenging. In principle, Proposition 2 shows genome-wide z-statistics identify Γ under the global null $V \perp\!\!\!\perp Y$. In practice, two obstacles remain: LD induces cross-variant dependence, and a subset of variants depend on Y. Fortunately, we can circumvent both of these issues:

Theorem 2. *Consider the estimator: $\widehat{S}_{ab} = \frac{1}{q}\sum_{j=1}^{q} z_{ja}z_{jb}$. Let G_q index the variants that are independent of Y, and B_q*

[2] All arguments can be modified by substituting sample sizes with Fisher-effective sample sizes in generalized linear models with $o_p(1)$ remainders.

index the variants that are dependent on Y. Let C_1, C_1', C_2, L and D denote finite constants. Assume:

1. *(Bounded moments for G_q) For $j \in G_q$, we have $\mathbb{E}(z_{ja}z_{jb}|M) = \Gamma_{ab}$ and $\sup_{j\in G_q} \mathrm{Var}(z_{ja}z_{jb}|M) \le C_1$.*
2. *(Bounded moments for B_q) For $j \in B_q$, we have $\mathbb{E}(z_{ja}z_{jb}|M) = \mu_j$ with $\sup_{j\in B_q} |\mu_j - \Gamma_{ab}| \le L$ and $\sup_{j\in B_q} \mathrm{Var}(z_{ja}z_{jb}|M) \le C_1'$.*
3. *(Local LD) There is an undirected dependency graph with edge set E_q and maximum degree D such that $|\mathrm{cov}(z_{ja}z_{jb}, z_{ka}z_{kb}|M)| \le C_2$ when $(j,k) \in E_q$ and $\mathrm{cov}(z_{ja}z_{jb}, z_{ka}z_{kb}|M) = 0$ when $(j,k) \notin E_q$.*
4. *(Vanishing contamination) We have $\frac{|B_q|}{q} \to 0$ as $q \to \infty$.*

Then, $\widehat{S}_{ab} \xrightarrow{p} \Gamma_{ab}$ conditional on M as $q \to \infty$.

Remark. The key condition is the local LD assumption; in practice, investigators bound each variant's LD neighborhood (e.g., 1 Mb) so that the dependency degree is finite. We show in the proof that this allows the variance of $\widehat{S}_{ab}$ to converge to zero so that $\widehat{S}_{ab} \xrightarrow{p} \Gamma_{ab}$ given the realized missingness pattern. The vanishing contamination assumption is also reasonable because only a tiny fraction of variants are usually associated with any outcome after quality control, so the non-null set scales sublinearly.

We conclude that the weak law of large numbers holds despite limited dependence. In practice, we estimate each entry of Γ to negligible error with millions of variants, so we treat Γ as known in subsequent derivations. Consequently, under $\mathcal{H}_0$, $z_j \xrightarrow{d} \mathcal{N}(0, \Gamma)$, and we replace Σ and A with Γ and $\Omega = \Gamma^{-1}$ so that $Q_j = z_j\, \Omega\, z_j^\top \sim \chi^2_m$ in the test of Expression (3). We will show that this test works very well with real data.

Interpretable Coefficients

We can now solve Expression (5) with an estimate of Γ. However, the raw coefficients lack a straightforward interpretation because:

$$\alpha_{ja} = \sqrt{n}\, \widehat{\mathrm{cov}}(V_j, R_a')/\mathrm{var}(R_a'),$$

where R_a' is the residual of Y_a' after regressing it onto Y_{-a}'. In this form, α_{ja} depends on the variance of R_a', which varies across outcomes and complicates comparisons.

To address this issue, we report rescaled coefficients $\widehat{\alpha}_{ja}' := \widehat{\alpha}_{ja}/\sqrt{\Omega_{aa}}$ for interpretability because:

Proposition 4. *We have $\widehat{\alpha}_{ja}' = \sqrt{n}\ \widehat{\mathrm{cor}}(V_j, R_a')$ for every $Y_a \in Y$.*

Remark. It follows that $\widehat{\alpha}_{ja}'$ corresponds to $\widehat{\mathrm{cor}}(V_j, R_a')$ up to a multiplicative constant $\sqrt{n}$. The modified CCA coefficients are also unitless, directly comparable across outcomes for a given variant V_j, and their signs match the genetic direction on the unique component of Y_a' (the residual R_a').

Quantifying the Degree of Mendelianism

We now turn to the final practical challenge: near-perfect Mendelianism is uncommon with small outcome sets (e.g., $\le$ 20). Although Theorem 1 provides a binary criterion, practice demands a continuous measure of how closely $Y\alpha_j$ approaches Mendelian behavior. We therefore fix the coefficients α_j learned at lead variant V_j and, for any variant V_k (including $k = j$), compute:

$$\mathcal{Z}_{k|j} = \frac{z_k\, \alpha_j}{\sqrt{\alpha_j^\top\, \Gamma\, \alpha_j}}.$$

Under $\mathcal{H}_{0,k} : r_k\alpha_j = 0$, we have $\mathcal{Z}_{k|j} \xrightarrow{d} \mathcal{N}(0,1)$ with i.i.d. samples and finite fourth moments, yielding the two-sided p-value $p_{k|j} = 2\Phi(-|\mathcal{Z}_{k|j}|)$[3]. We compute $\mathcal{Z}_{k|j}$ and $p_{k|j}$ for all $V_k \in V$ (i.e., genome-wide).

If we reject $\mathcal{H}_0$ in Expression (3) for V_j, then we summarize concentration at the j-locus with the **Score of Mendelianism (SoM)**:

$$\mathcal{M}_j = \frac{\frac{1}{|\mathcal{S}_j|}\sum_{i\in\mathcal{S}_j} h(\mathcal{Z}_{i|j})}{\frac{1}{|\mathcal{S}_j|}\sum_{i\in\mathcal{S}_j} h(\mathcal{Z}_{i|j}) \;+\; \sum_{k\notin\mathcal{S}_j} h(\mathcal{Z}_{k|j})}, \qquad (6)$$

where $h(\mathcal{Z}) = \big(|\mathcal{Z}| - z_\alpha\big)_+$, and z_α denotes the z-score corresponding to a two-sided p-value (default $\alpha = 5 \times 10^{-8}$). We construct $\mathcal{S}_j$ using proximity-based clustering[12] around V_j (single-linkage with 500kb gap, padded by $\pm$500kb), though alternatives such as LD thresholds relative to V_j are also possible.

The SoM lies in $[0,1]$ and $\mathcal{M}_j = 1$ indicates complete isolation of a single locus. The score decreases toward 0 as polygenic spillover grows, approaching 0 when signal lies entirely outside $\mathcal{S}_j$. We average within $\mathcal{S}_j$ to avoid favoring large loci and do *not* average outside $\mathcal{S}_j$ so that small off-target regions remain visible rather than diluted among millions of variants. Finally, $\mathcal{M}_j$ serves only as a ranking metric among already significant loci; we do not use it for hypothesis testing or interval estimation.

Summary

Mendelianization ultimately corresponds to one-dimensional CCA, modified in five ways: (1) using z-statistics in place of correlation coefficients, (2) overcoming missingness by normalizing with Γ estimated over variants rather than with Σ estimated over outcomes, (3) replacing Wilks' Λ with a χ^2_m test for significance, (4) rescaling canonical coefficients by the square roots of the diagonals of Ω, and (5) applying each significant composite outcome genome-wide by Theorem 1. Pseudocode and a formal time complexity analysis of all components demonstrates that Mendelianization grows cubically with the number of outcome variables but only linearly with the number of variants (Supplementary Materials B). This is computationally practical, as outcomes rarely exceed 100, whereas the number of variants often surpasses one million.

Results

Comparators

We compared Mendelianization against:

- **Meta-analytic Canonical Correlation Analysis (metaCCA)**:[4] performs CCA on summary statistics using Wilks' Λ, which assumes perfectly overlapping and equal sample sizes per trait. metaCCA also estimates Σ from linear regression β coefficients, so it cannot fully combine linear and logistic regression models whose coefficients differ in scale and null variances. Finally, metaCCA lacks causal guarantees, and

[3] The p-values are not always calibrated in practice, because the test assumes that α_j is fixed. We will not use this test for inference but for visualizing and quantifying the degree of Mendelianism only.

its scale-dependent canonical coefficients are not directly comparable across traits.

- **Heritability Informed Power Optimization (HIPO)**:[10] learns multiple sets of outcome weights that increase statistical power by maximizing a standardized non-centrality parameter across – rather than within – loci. We compare against the most powerful, first set of HIPO weights.
- **Fast Association analysis for SubSETs (fastASSET)**:[11] a multi-trait analysis method that first pre-screens traits with suggestive associations and adjusts for this selection. It then searches over all (two-sided) subsets of traits – allowing opposite effect directions – and combines evidence into a single, multiple testing adjusted omnibus p-value per variant.

Mendelianization is the only method that optimizes composite outcomes so that each learned outcome is associated with a single tower in the Manhattan plot that contains at least one causal variant. We finally also compared Mendelianization against **three ablated variants**: (i) using variant-specific outcomes (i.e., $\mathcal{Z}_{i|j} = \mathcal{Z}_{i|i}$) as opposed to fixed outcomes applied genome-wide, (ii) estimating Γ using β coefficients instead of z-statistics, (iii) a combination of both (i) and (ii).

Metrics

Our primary evaluation metric was the **Score of Mendelianism (SoM)**, where higher values indicate a greater ability to isolate a single causal locus. The metric provides a fair basis for comparison because no algorithm, including Mendelianization, optimizes SoM. Given that Mendelianization and metaCCA have different coefficient interpretations, we computed SoM with variant-specific outcomes for metaCCA, and with fixed outcomes applied genome-wide for Mendelianization (the outcome learned at the lead V_j, applied to all variants). This ensures each method is evaluated in line with its intended use.

Beyond detecting a Mendelian-like pattern, rigorous applications also require valid, efficient inference. As a result, we evaluated Type I error control using the **signed Kolmogorov–Smirnov (KS) distance** between observed p-values above 0.05 and the uniform distribution on $[0.05, 1]$. Well-calibrated methods should yield nearly uniform p-values in this range, since most variants are uncorrelated with the target. For methods with positive (anti-conservative) KS-distances below 0.01, we then assessed the **statistical power** of the hypothesis tests for p-values below 10^{-4}, since most variants with p-values below this threshold are likely to be associated with their learned outcome. Finally, we measured **run-time**.

Depression and Generalized Anxiety

We evaluated the algorithms using European ancestry summary statistics from the Pan-UK Biobank.[3,7] The outcome variables comprised all 52 non-collinear items in the UK Biobank Mental Health Questionnaire assessing lifetime major depression and generalized anxiety. For all experiments, we restricted analyses to high-quality HapMap3 autosomal, non-MHC, biallelic SNPs with INFO>0.9 and MAF>1% in UK Biobank and, when available, gnomAD genome/exome.

Figure 2 summarizes the results. Mendelianization identified three loci that surpassed the significance threshold of 5×10^{-8}. These loci also attained perfect SoMs of one, whereas most loci identified by the other algorithms had scores close to zero (Figure 2 (a)). Applying each of the three learned composite $Y\alpha_j$ genome-wide and plotting the p-values associated with the z-statistics $\mathcal{Z}_{\cdot|j}$ yielded a single tower per locus, as implied by the SoMs and predicted by Theorem 1 (Figure 2 (b)). The interpretable coefficients further delineated three distinct patient profiles: (i) uncontrollable worry with recurrent depression, (ii) melancholic-like depression characterized by insomnia and worthlessness, and (iii) controllable worry accompanied by somatic complaints (Supplementary Figure 1). Together, these results demonstrate that Mendelianization achieved the greatest spikedness while yielding interpretable coefficient patterns.

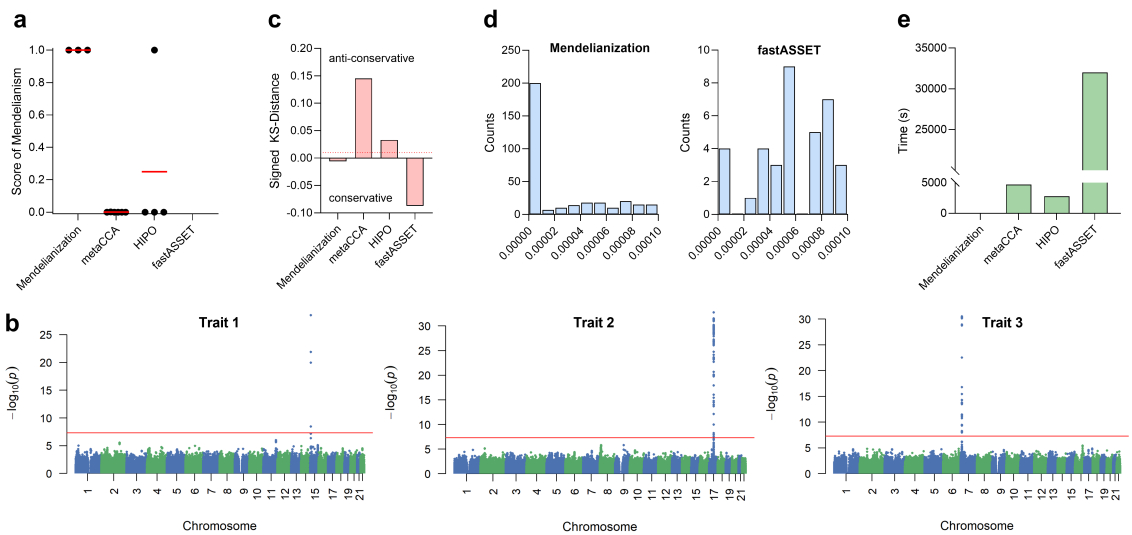


Fig. 2. Main results for depression and generalized anxiety. (a) Each dot denotes the SoM for the lead variant within a Manhattan-plot tower, and the red lines correspond to the mean SoMs. Only Mendelianization achieved perfect SoMs for all lead variants. No variants reached the genome-wide significance threshold under fastASSET. (b) For each of the three dots in the Mendelianization column of panel (a), the learned outcome corresponded to an essentially perfect Mendelian trait because it produced a Manhattan plot with a single tower, consistent with Theorem 1. (c) Only Mendelianization produced well-calibrated p-values; fastASSET was conservative. (d) Among methods with well-calibrated or conservative p-values, Mendelianization achieved the highest statistical power by a wide margin. (e) Mendelianization was approximately four orders of magnitude faster than alternative methods.

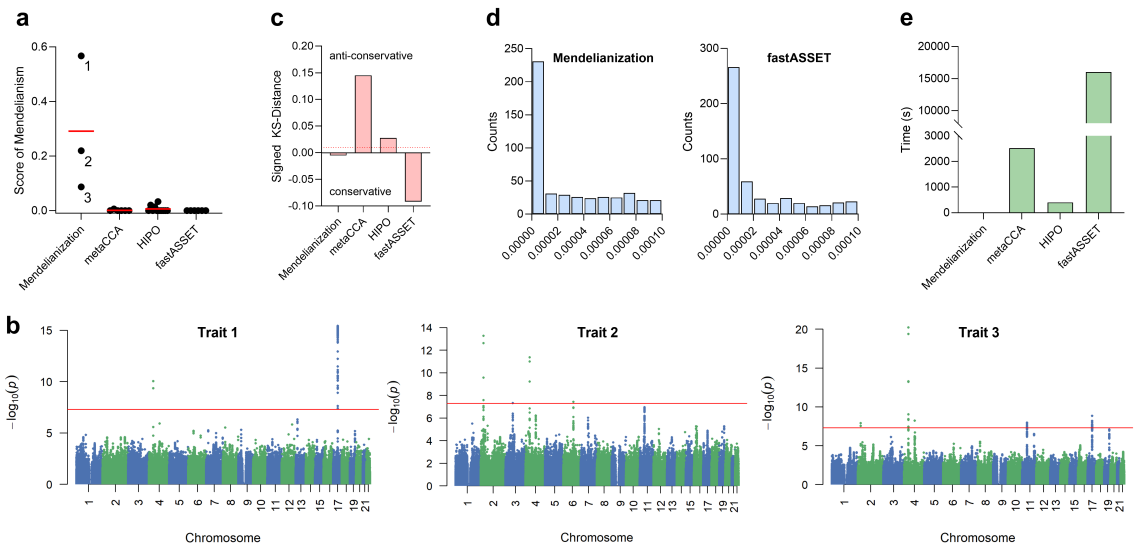


Fig. 3. Main results for alcohol use disorder. (a) Mendelianization achieved the highest SoMs for each lead variant. (b) The learned outcome for each point in the Mendelianization column of panel (a) approximated a Mendelian trait. (c) Only Mendelianization produced well-calibrated p-values. (d) Both Mendelianization and fastASSET attained high statistical power. (e) Mendelianization completed in 0.58 seconds.

We next evaluated secondary performance metrics. Only Mendelianization produced well-calibrated p-values (Figure 2 (c)). FastASSET generated conservative p-values, whereas metaCCA and HIPO yielded anti-conservative estimates, indicating unreliable significance levels. We plot the individual histograms in Supplementary Figure 2. Among methods with calibrated or conservative behavior, Mendelianization achieved the greatest statistical power by far (Figure 2 (d)). The algorithm also required only 2 seconds to complete, while the other methods were slower by at least four orders of magnitude (Figure 2 (e)). Overall, Mendelianization uniquely combined well-calibrated inference, high power, and rapid computation. Finally, ablation results confirmed the necessity of applying canonical coefficients genome-wide and using z-statistics to achieve maximal performance (Supplementary Figure 3).

Alcohol Use Disorder

We increased the difficulty in alcohol use disorder by running the algorithms with only 10 outcome variables corresponding to all individual items of the Alcohol Use Disorders Identification Test (AUDIT).[13] We again used European ancestry summary statistics from the Pan-UK Biobank with the same quality control. According to Theorem 1, Mendelianization is unlikely to recover near-perfect Mendelian traits with only 10 outcomes, but we still expect it to find approximate Mendelian traits.

The results, summarized in Figure 3, support this expectation. Mendelianization did not achieve perfect SoMs, but every locus achieved higher scores than those from any competing method. Visualization of the per-outcome genome-wide Manhattan plots revealed a substantial, though not perfect, degree of Mendelianism (Figure 3 (b)). Each trait was associated with an interpretable set of canonical coefficients (Supplementary Figure 4). As before, Mendelianization was the only method to yield well-calibrated p-values (Figure 3 (c); Supplementary Figure 5), while both Mendelianization and fastASSET achieved high statistical power (Figure 3 (d)). Finally, Mendelianization completed at least three orders of magnitude faster than competing methods (Figure 3 (e)), and ablation results confirmed the necessity of the individual components of the algorithm (Supplementary Figure 6). We conclude that Mendelianization was again the only method that simultaneously achieved high degrees of Mendelianism, well-calibrated inference, high power, and rapid computation.

Discussion

Mendelianization learns composite outcomes whose associations asymptotically concentrate at a single locus. The locus contains at least one causal variant under just four structural assumptions – low-dimensional confounding, outcome diversity, signal floors, and isolation-spillover. The algorithm also handles partial sample-overlap, yields cross-outcome comparable canonical coefficients, and quantifies single-locus concentration via the SoM. In short, Mendelianization performs genome-wide, locus-level causal inference using a substantially modified one-dimensional CCA.

Despite these strengths, Mendelianization has limitations. First, it cannot resolve LD within a locus because it is a one-dimensional method; isolating variant-level causality generally requires fine-mapping – a separate class of techniques fundamentally distinct from Mendelianization. Second, Mendelianization cannot accommodate completely non-overlapping samples between outcome pairs, which poses challenges when distinct populations are assessed with different measures. Third, Mendelianization remains a linear method that may miss nonlinear confounding and causal effects. Future research should therefore explore combining Mendelianization with fine-mapping approaches, extending it to handle fully non-overlapping samples, and generalizing it to nonlinear settings.

In summary, Mendelianization provides a powerful framework for isolating causal loci to the extent permitted by available outcomes, advancing a reductionist approach that helps clarify – rather than complicate – the pathophysiological understanding of complex disease.

Competing interests

No competing interest is declared.

Author contributions statement

E.V.S. conceived, designed, and executed the project in its entirety.

Acknowledgments

TBD

References

1. Arvanitis, M., Tayeb, K., Strober, B.J., Battle, A.: Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. The American Journal of Human Genetics **109**(2), 223–239 (2022)
2. Ballard, J.L., O'Connor, L.J.: Shared components of heritability across genetically correlated traits. The American Journal of Human Genetics **109**(6), 989–1006 (2022)
3. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al.: The uk biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)
4. Cichonska, A., Rousu, J., Marttinen, P., Kangas, A.J., Soininen, P., Lehtimäki, T., Raitakari, O.T., Järvelin, M.R., Salomaa, V., Ala-Korpela, M., et al.: metacca: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics **32**(13), 1981–1989 (2016)
5. Fried, E.I., Nesse, R.M.: Depression is not a consistent syndrome: An investigation of unique symptom patterns in the star* d study. Journal of Affective Disorders **172**, 96–102 (2015)
6. Grotzinger, A.D., Rhemtulla, M., de Vlaming, R., Ritchie, S.J., Mallard, T.T., Hill, W.D., Ip, H.F., Marioni, R.E., McIntosh, A.M., Deary, I.J., et al.: Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nature Human Behaviour **3**(5), 513–525 (2019)
7. Karczewski, K.J., Gupta, R., Kanai, M., Lu, W., Tsuo, K., Wang, Y., Walters, R.K., Turley, P., Callier, S., Shah, N.N., et al.: Pan-uk biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. Nature Genetics pp. 1–10 (2025)
8. Little, R.J., Rubin, D.B.: Statistical analysis with missing data. John Wiley & Sons (2019)
9. Mostafavi, H., Spence, J.P., Naqvi, S., Pritchard, J.K.: Systematic differences in discovery of genetic effects on gene expression and complex traits. Nature Genetics **55**(11), 1866–1875 (2023)
10. Qi, G., Chatterjee, N.: Heritability informed power optimization (hipo) leads to enhanced detection of genetic associations across multiple traits. PLoS Genetics **14**(10), e1007549 (2018)
11. Qi, G., Chhetri, S.B., Ray, D., Dutta, D., Battle, A., Bhattacharjee, S., Chatterjee, N.: Genome-wide large-scale multi-trait analysis characterizes global patterns of pleiotropy and unique trait-specific variants. Nature Communications **15**(1), 6985 (2024)
12. Quinlan, A.R., Hall, I.M.: Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)
13. Saunders, J.B., Aasland, O.G., Babor, T.F., De la Fuente, J.R., Grant, M.: Development of the alcohol use disorders identification test (audit): Who collaborative project on early detection of persons with harmful alcohol consumption-ii. Addiction **88**(6), 791–804 (1993)
14. Schaid, D.J., Chen, W., Larson, N.B.: From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics **19**(8), 491–504 (2018)
15. Strobl, E.V., Gamazon, E.R.: Transcriptome-wide root causal inference. PLOS Computational Biology **21**(9), e1013461 (2025)
16. Svishcheva, G.R., Tiys, E.S., Elgaeva, E.E., Feoktistova, S.G., Timmers, P.R., Sharapov, S.Z., Axenovich, T.I., Tsepilov, Y.A.: A novel framework for analysis of the shared genetic background of correlated traits. Genes **13**(10), 1694 (2022)
17. Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., Posthuma, D.: Genome-wide association studies. Nature Reviews Methods Primers **1**(1), 59 (2021)
18. Zhang, Z., Jung, J., Kim, A., Suboc, N., Gazal, S., Mancuso, N.: A scalable approach to characterize pleiotropy across thousands of human diseases and complex traits using gwas summary statistics. The American Journal of Human Genetics **110**(11), 1863–1874 (2023)
19. Zou, Y., Carbonetto, P., Xie, D., Wang, G., Stephens, M.: Fast and flexible joint fine-mapping of multiple traits via the sum of single effects model. bioRxiv pp. 2023–04 (2024)
20. Zuber, V., Cronjé, T., Cai, N., Gill, D., Bottolo, L.: Bayesian causal graphical model for joint mendelian randomization analysis of multiple exposures and outcomes. The American Journal of Human Genetics **112**(5), 1173–1198 (2025)
21. Zuber, V., Lewin, A., Levin, M.G., Haglund, A., Ben-Aicha, S., Emanueli, C., Damrauer, S., Burgess, S., Gill, D., Bottolo, L.: Multi-response mendelian randomization: Identification of shared and distinct exposures for multimorbidity and multiple related disease outcomes. The American Journal of Human Genetics **110**(7), 1177–1199 (2023)

Supplementary Materials

A. Finite m Assumptions

Assumption 5. *(Low-dimensional latent confounding) For a prespecified $m \in \mathbb{N}^+$, the confounding effects are contained in a subspace $U^m \subset \mathbb{R}^m$, and there exists an integer constant $r_0 \geq 0$ such that $\dim(U^m) \leq r_0$.*

Assumption 6. *(Outcome diversity) We have $\max_{k \notin \mathcal{S}_j} \frac{1}{m} |\langle r_k^{m\perp}, r_j^{m\perp} \rangle_{A^m}| = 0$ for a prespecified m.*

Assumption 7. *(Signal floor) For a prespecified m, if $l \in \mathcal{S}_j$ causes $Y\alpha_j$, then there exists a constant $\beta_c > 0$ such that $\frac{1}{m}\|r_l^{m\perp}\|_{A^m}^2 \geq \beta_c$. Moreover, there exists a constant $\beta_{on} > 0$ such that $\frac{1}{m}\|r_l^{m\parallel}\|_{A^m}^2 \geq \beta_{on}$ for every $l \in \mathcal{S}_j$ that does not cause $Y\alpha_j$.*

Assumption 8. *(Isolation-spillover) If some $l \in \mathcal{S}_j$ causes $Y\alpha_j$, we have $\max_{k \notin \mathcal{S}_j} \frac{1}{m} |\langle r_k^{m\parallel}, r_j^{m\parallel} \rangle_{A^m}| = 0$ for a prespecified m. On the other hand, if there does not exist $l \in \mathcal{S}_j$ that causes $Y\alpha_j$, then $\exists k \notin \mathcal{S}_j$ with $|\cos_{A^m}(r_j^{m\parallel}, r_k^{m\parallel})| \geq \gamma_j > 0$ and $\frac{1}{m}\|r_k^{m\parallel}\|_{A^m}^2 \geq \beta_{off} > 0$ for a prespecified m.*

B. Pseudocode & Time Complexity Analysis

We present the pseudocode for Mendelianization in Algorithm 1. The algorithm first computes $\widehat{S}$ in Line 1, which serves as a consistent estimator of Γ under Theorem 2. Using this estimate, it then derives the raw canonical coefficients in Line 2 by solving Expression (5). These coefficients also form the basis of the Q statistics in Expression (3) that enable hypothesis testing; recall that each Q_j follows a χ^2_m distribution under the null, so the algorithm can readily compute p-values in Line 4. Next, Mendelianization computes SoMs for lead variants that surpass a genome-wide significance threshold (e.g., 5×10^{-8}) in Line 5. Finally, the algorithm converts the raw coefficients into their interpretable form α' by applying Proposition 4 in Line 6. Thus, although the theory of Mendelianization is intricate, the resulting algorithm is remarkably straightforward.

Algorithm 1 Mendelianization

Input: multiple outcome z-statistics

Output: interpretable canonical coefficients α'; test statistics Q and p-values; Scores of Mendelianism $\mathcal{M}$

1: $\Gamma \leftarrow$ compute $\widehat{S}$ using z-statistics (Theorem 2)
2: $\alpha \leftarrow \Omega z^\top$
3: $Q \leftarrow \mathrm{diag}(z\alpha)$
4: $p \leftarrow$ right tail probability of each Q_j under χ^2_m
5: $\mathcal{M} \leftarrow$ compute SoMs (Equation (6); Theorem 1)
6: $\alpha' \leftarrow \alpha/\sqrt{\mathrm{diag}(\Omega)}$ (Proposition 4)

The algorithm also has a favorable time complexity. In particular, estimating the correlation matrix Γ described in Line 1 costs $O(qm^2)$ time. In Line 2, inverting Γ requires $O(m^3)$ time, and the matrix product $\Omega z^\top$ required to estimate α costs $O(m^2 q)$. Constructing all Q_j statistics and evaluating their p-values in Lines 3 and 4 adds $O(mq)$, and computing α' in Line 6 requires another $O(mq)$. For the SoM step, suppose there are s significant variants, and each constitutes its own tower to take the worst case. Computing each statistic $\mathcal{Z}_{j|k}$ requires $O(m)$ time for the numerator and $O(m^2)$ for the denominator in Equation (6). Moreover, the numerator must be evaluated genome-wide over q variants, so an upper bound on the total cost of SoM amounts to $O(s(mq+m^2))$. Summing these components, the overall complexity of Mendelianization is $O(m^3 + m^2 q + sqm)$, where the simplification uses $q \gg m$. Thus, the algorithm scales linearly with the number of variants and cubically with the number of outcome variables. In practice, Mendelianization completes within seconds and is at least three orders of magnitude faster than the competing methods.

C. Proofs

Theorem 1. *(Asymptotic Mendelianism) Consider Assumptions 1-4 and suppose Σ^m is positive definite for sufficiently large m. We have $\frac{1}{m} r_j^m \alpha_j^m > 0$ for sufficiently large m and $\max_{k \notin \mathcal{S}_j} |\frac{1}{m} r_k^m \alpha_j^m| \to 0$ (i.e., Mendelianism is asymptotically achieved for $Y\alpha_j^m$ at $\mathcal{S}_j$) if and only if at least one variant $l \in \mathcal{S}_j$ causes $Y\alpha_j^m$.*

Proof We first prove the forward direction by contrapositive. Suppose there does not exist a variant $l \in \mathcal{S}_j$ that causes $Y\alpha_j^m$. As a result, $j \in \mathcal{S}_j$ does not cause $Y\alpha_j^m$. We need to show $\frac{1}{m} r_j^m \alpha_j^m \to 0$ or $\max_{k \notin \mathcal{S}_j} |\frac{1}{m} r_k^m \alpha_j^m| > 0$ (or both) for m sufficiently large. We will show that $\max_{k \notin \mathcal{S}_j} |\frac{1}{m} r_k^m \alpha_j^m| > 0$.

Assumption 1 allows us to write:

$$\frac{1}{m}\langle r_k^m, r_j^m \rangle_{A^m} = \frac{1}{m}\langle r_k^{m\parallel}, r_j^{m\parallel} \rangle_{A^m} + \frac{1}{m}\langle r_k^{m\perp}, r_j^{m\perp} \rangle_{A^m}. \tag{7}$$

Choose $M_1 > 0$. By Assumptions 3 and 4, $\exists M_2 \geq M_1$ such that $\forall m \geq M_2, \exists k \notin \mathcal{S}_j$ where:

$$\left|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m}\right| \geq \|r_j^{m\parallel}\|_{A^m}\|r_k^{m\parallel}\|_{A^m}\gamma_j \geq m\sqrt{\beta_{\text{on}}}\sqrt{\beta_{\text{off}}}\gamma_j.$$

The reverse triangle inequality allows us to write:

$$\begin{aligned}&\left|\langle r_k^m, r_j^m\rangle_{A^m}\right| = \left|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m} + \langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}\right| \\ \geq&\left|\left|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m}\right| - \left|\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}\right|\right| \geq \left|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m}\right| - \left|\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}\right|.\end{aligned}$$

Let $\varepsilon_m = \max_{k\notin\mathcal{S}_j}\frac{1}{m}|\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}|$. Under Assumption 2, $\exists M_3 \geq M_2$ such that $\forall m > M_3$ we have $\varepsilon_m < \sqrt{\beta_{\text{on}}}\sqrt{\beta_{\text{off}}}\gamma_j$. As a result:

$$\frac{1}{m}|r_k^m\alpha_j^m| = \frac{1}{m}\left|\langle r_k^m, r_j^m\rangle_{A^m}\right| \geq \sqrt{\beta_{\text{on}}}\sqrt{\beta_{\text{off}}}\gamma_j - \varepsilon_m > 0,$$

where $\varepsilon_m = \max_{k\notin\mathcal{S}_j}\frac{1}{m}|\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}|$. This completes the proof of sufficiency – i.e., asymptotic Mendelianism is sufficient to establish causality.

To prove the backward direction, assume that there exists at least one variant $l \in \mathcal{S}_j$ that causes $Y\alpha_j^m$. Then, we can invoke Assumption 1 and write Equation (7) for any k. We have two situations:

1. If $k \notin \mathcal{S}_j$, then $\max_{k\notin\mathcal{S}_j}\frac{1}{m}|\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}| \to 0$ by Assumption 2. Moreover, $\frac{1}{m}|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m}| \to 0$ by the first clause of Assumption 4. Hence, $\max_{k\notin\mathcal{S}_j}\frac{1}{m}|\langle r_k, r_j\rangle_{A^m}| \to 0$.
2. If $k \in \mathcal{S}_j$ and $k = j$, then we have two situations. If j causes $Y\alpha_j^m$, then the first clause of Assumption 3 allows us to write:

$$\frac{1}{m}\langle r_j^m, r_j^m\rangle_{A^m} = \frac{1}{m}|\langle r_j^m, r_j^m\rangle_{A^m}| = \frac{1}{m}\|r_j^{m\perp}\|_{A^m}^2 + \frac{1}{m}\|r_j^{m\parallel}\|_{A^m}^2 \geq \frac{1}{m}\|r_j^{m\perp}\|_{A^m}^2 \geq \beta_{\text{c}} > 0,$$

for sufficiently large m. On the other hand, if j does not cause $Y\alpha_j^m$, then the second clause of Assumption 3 allows us to write:

$$\frac{1}{m}\langle r_j^m, r_j^m\rangle_{A^m} = \frac{1}{m}\|r_j^{m\perp}\|_{A^m}^2 + \frac{1}{m}\|r_j^{m\parallel}\|_{A^m}^2 \geq \frac{1}{m}\|r_j^{m\parallel}\|_{A^m}^2 \geq \beta_{\text{on}} > 0,$$

for sufficiently large m.

We conclude that $\frac{1}{m}r_j^m\alpha_j^m > 0$ for sufficiently large m and $\max_{k\notin\mathcal{S}_j}|\frac{1}{m}r_k^m\alpha_j^m| \to 0$, which completes the proof of necessity – i.e., asymptotic Mendelianism is necessary to establish causality. □

Proposition 1. *(Perfect Mendelianism at finite m) Consider Assumptions 5-8 with a prespecified m and suppose Σ^m is positive definite. We have $\frac{1}{m}r_j^m\alpha_j^m > 0$ and $\max_{k\notin\mathcal{S}_j}|\frac{1}{m}r_k^m\alpha_j^m| = 0$ (i.e., Mendelianism is perfectly achieved for $\mathcal{S}_j$ at m) if and only if at least one variant $l \in \mathcal{S}_j$ causes $Y\alpha_j^m$.*

Proof The proof almost exactly parallels that of Theorem 1, but we include it here for completeness. We first prove the forward direction by contrapositive. Suppose there does not exist a variant $l \in \mathcal{S}_j$ that causes $Y\alpha_j^m$. As a result, $j \in \mathcal{S}_j$ does not cause $Y\alpha_j^m$. We need to show $\frac{1}{m}r_j^m\alpha_j^m = 0$ or $\max_{k\notin\mathcal{S}_j}|\frac{1}{m}r_k^m\alpha_j^m| > 0$ (or both) for prespecified m. We will show that $\max_{k\notin\mathcal{S}_j}|\frac{1}{m}r_k^m\alpha_j^m| > 0$.

Assumption 5 allows us to write:

$$\frac{1}{m}\langle r_k^m, r_j^m\rangle_{A^m} = \frac{1}{m}\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m} + \frac{1}{m}\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}. \tag{8}$$

By Assumptions 7 and 8, there exists $k \notin \mathcal{S}_j$ where:

$$\left|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m}\right| \geq \|r_j^{m\parallel}\|_{A^m}\|r_k^{m\parallel}\|_{A^m}\gamma_j \geq m\sqrt{\beta_{\text{on}}}\sqrt{\beta_{\text{off}}}\gamma_j.$$

The reverse triangle inequality allows us to write:

$$\begin{aligned}&\left|\langle r_k^m, r_j^m\rangle_{A^m}\right| = \left|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m} + \langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}\right| \\ \geq&\left|\left|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m}\right| - \left|\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}\right|\right| \geq \left|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m}\right| - \left|\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}\right|.\end{aligned}$$

Thus, under Assumption 6, we have:

$$\frac{1}{m}|r_k^m\alpha_j^m| = \frac{1}{m}\left|\langle r_k^m, r_j^m\rangle_{A^m}\right| \geq \sqrt{\beta_{\text{on}}}\sqrt{\beta_{\text{off}}}\gamma_j > 0.$$

This completes the proof of sufficiency – i.e., perfect Mendelianism is sufficient to establish causality.

To prove the backward direction, assume that there exists at least one variant $l \in \mathcal{S}_j$ that causes $Y\alpha_j^m$. Then, we can invoke Assumption 5 and write Equation (8) for any k. We have two situations:

1.If $k \notin \mathcal{S}_j$, then $\max_{k \notin \mathcal{S}_j} \frac{1}{m} |\langle r_k^{m\perp}, r_j^{m\perp}\rangle_{A^m}| = 0$ by Assumption 6. Moreover, $\frac{1}{m}|\langle r_k^{m\parallel}, r_j^{m\parallel}\rangle_{A^m}| = 0$ by the first clause of Assumption 8. Hence, $\max_{k \notin \mathcal{S}_j} \frac{1}{m}|\langle r_k, r_j\rangle_{A^m}| = 0$.
2.If $k \in \mathcal{S}_j$ and $k = j$, then we have two situations. If j causes $Y\alpha_j^m$, then the first clause of Assumption 7 allows us to write:

$$\frac{1}{m}\langle r_j^m, r_j^m\rangle_{A^m} = \frac{1}{m}|\langle r_j^m, r_j^m\rangle_{A^m}| = \frac{1}{m}\|r_j^{m\perp}\|_{A^m}^2 + \frac{1}{m}\|r_j^{m\parallel}\|_{A^m}^2 \geq \frac{1}{m}\|r_j^{m\perp}\|_{A^m}^2 \geq \beta_{\mathrm{c}} > 0.$$

On the other hand, if j does not cause $Y\alpha_j^m$, then the second clause of Assumption 7 allows us to write:

$$\frac{1}{m}\langle r_j^m, r_j^m\rangle_{A^m} = \frac{1}{m}\|r_j^{m\perp}\|_{A^m}^2 + \frac{1}{m}\|r_j^{m\parallel}\|_{A^m}^2 \geq \frac{1}{m}\|r_j^{m\parallel}\|_{A^m}^2 \geq \beta_{\mathrm{on}} > 0.$$

We conclude that $\frac{1}{m} r_j^m \alpha_j^m > 0$ and $\max_{k \notin \mathcal{S}_j} |\frac{1}{m} r_k^m \alpha_j^m| = 0$, which completes the proof of necessity – i.e., perfect Mendelianism is necessary to establish causality. □

Proposition 2. *Assume i.i.d. samples across individuals, $V \perp\!\!\!\perp Y$ and $V \perp\!\!\!\perp M|Y$. Then $\mathbb{E}[z_{ja} z_{jb} \mid M] = \frac{n_{ab}}{\sqrt{n_a n_b}} \Sigma_{ab}^{\mathrm{obs}}$.*

Proof Note that z-statistics are computed on observed samples, so we can write:

$$\begin{aligned}\mathbb{E}[z_{ja} z_{jb} \mid M] &= \mathbb{E}\left(\sqrt{n_a}\widehat{r}_{ja}\sqrt{n_b}\widehat{r}_{jb} \mid M\right) = \mathbb{E}\left(\sqrt{n_a}\widehat{r}_{ja}\sqrt{n_b}\widehat{r}_{jb} \mid M_a = 1, M_b = 1, M_{-a,-b}\right) \\ &= \mathbb{E}\left(\frac{\sqrt{n_a}}{n_a}\sum_{i \in S_a} V_{ji} Y_{ai} \frac{\sqrt{n_b}}{n_b}\sum_{k \in S_b} V_{jk} Y_{bk} \mid M_a = 1, M_b = 1, M_{-a,-b}\right).\end{aligned} \tag{9}$$

Cross-terms cancel by i.i.d. samples. Moreover, $V \perp\!\!\!\perp Y$ and $V \perp\!\!\!\perp M|Y$ imply $V \perp\!\!\!\perp Y|M$ by the contraction and weak union semi-graphoid axioms. As a result, we have:

$$\mathbb{E}\left(\frac{\sqrt{n_a}}{n_a}\sum_{i \in S_a} V_{ji} Y_{ai} \frac{\sqrt{n_b}}{n_b}\sum_{k \in S_b} V_{jk} Y_{bk} \mid \widetilde{M}\right) = \frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\mathbb{E}\left(V_j^2 \mid \widetilde{M}\right)\mathbb{E}\left(Y_a Y_b \mid \widetilde{M}\right),$$

where $\widetilde{M} = \{M_a = 1, M_b = 1, M_{-a,-b}\}$. Recall that genotypes were standardized so that $\mathbb{E}(V_j) = 0$ and $\mathrm{Var}(V_j) = 1$. Thus:

$$\mathbb{E}\left(V_j^2 \mid \widetilde{M}\right) = \mathbb{E}_Y\left(\mathbb{E}\left(V_j^2 \mid Y, \widetilde{M}\right) \mid \widetilde{M}\right) = \mathbb{E}_Y\left(\mathbb{E}\left(V_j^2 \mid Y\right) \mid \widetilde{M}\right) = \mathbb{E}_Y(1 \mid \widetilde{M}) = 1.$$

We applied $V \perp\!\!\!\perp M|Y$ in the second equality, and $V \perp\!\!\!\perp Y$ and standardization in the third. Hence:

$$\frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\mathbb{E}\left(V_j^2 \mid \widetilde{M}\right)\mathbb{E}\left(Y_a Y_b \mid \widetilde{M}\right) = \frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\mathbb{E}\left(Y_a Y_b \mid \widetilde{M}\right) = \frac{n_{ab}}{\sqrt{n_a n_b}}\Sigma_{ab}^{\mathrm{obs}}.$$

We have established the equality $\mathbb{E}[z_{ja} z_{jb} \mid \widetilde{M}] = \mathbb{E}[z_{ja} z_{jb} \mid M] = \frac{n_{ab}}{\sqrt{n_a n_b}}\Sigma_{ab}^{\mathrm{obs}}$. □

Proposition 3. *Assume i.i.d. samples across individuals, $p_a = \mathbb{P}(M_a = 1) > 0$, and $p_b = \mathbb{P}(M_b = 1) > 0$. We have $\mathbb{E}\left(\frac{1}{n}\sum_t Y'_{a,t} Y'_{b,t} \mid M\right) = \Gamma_{ab} \overset{a.s.}{\to} \frac{p_{ab}}{\sqrt{p_a p_b}}\Sigma_{ab}^{\mathrm{obs}}$ and $\mathbb{E}\left(\frac{1}{n}\sum_t Y'_{a,t} Y'_{b,t}\right) \to \frac{p_{ab}}{\sqrt{p_a p_b}}\Sigma_{ab}^{\mathrm{obs}}$ as $n \to \infty$.*

Proof For the first statement, we may write:

$$\mathbb{E}\left(\frac{1}{n}\sum_t Y'_{a,t} Y'_{b,t} \mid M\right) = \frac{1}{\sqrt{n_a}\sqrt{n_b}}\mathbb{E}\left(\sum_t M_{at} Y_{at} M_{bt} Y_{bt} \mid M\right) = \frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\mathbb{E}\left(Y_a Y_b | M_a = 1, M_b = 1\right) = \frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\Sigma_{ab}^{\mathrm{obs}} = \Gamma_{ab}.$$

Note that:

$$\frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}} = \frac{n\widehat{p}_{ab}}{\sqrt{n\widehat{p}_a}\sqrt{n\widehat{p}_b}} = \frac{\widehat{p}_{ab}}{\sqrt{\widehat{p}_a}\sqrt{\widehat{p}_b}} \overset{a.s.}{\to} \frac{p_{ab}}{\sqrt{p_a}\sqrt{p_b}}.$$

where $p_{ab} = \mathbb{P}(S_a \cap S_b)$, and almost sure convergence holds by the strong law of large numbers and the continuous mapping theorem. Hence, $\Gamma_{ab} \overset{a.s.}{\to} \frac{p_{ab}}{\sqrt{p_a}\sqrt{p_b}}\Sigma_{ab}^{\mathrm{obs}}$.

We apply the law of total expectation for the second statement:

$$\mathbb{E}\left(\frac{1}{n}\sum_t Y'_{a,t} Y'_{b,t}\right) = \mathbb{E}_M\left(\mathbb{E}\left(\frac{1}{n}\sum_t Y'_{a,t} Y'_{b,t} \mid M\right)\right) = \mathbb{E}\left(\frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\Sigma_{ab}^{\mathrm{obs}}\right) = \mathbb{E}\left(\frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\right)\Sigma_{ab}^{\mathrm{obs}}.$$

Moreover, we have $0 \leq \frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}} \leq 1$ by cosine similarity, so the sequence is uniformly bounded. We combine this with the almost sure convergence above to invoke the dominated convergence theorem:

$$\mathbb{E}\left(\frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\right) \to \frac{p_{ab}}{\sqrt{p_a}\sqrt{p_b}}.$$

It follows that:

$$\mathbb{E}\left(\frac{n_{ab}}{\sqrt{n_a}\sqrt{n_b}}\right)\Sigma_{ab}^{\mathrm{obs}} \to \frac{p_{ab}}{\sqrt{p_a}\sqrt{p_b}}\Sigma_{ab}^{\mathrm{obs}}.$$

Hence, $\mathbb{E}\left(\frac{1}{n}\sum_t Y'_{a,t} Y'_{b,t}\right) \to \frac{p_{ab}}{\sqrt{p_a}\sqrt{p_b}}\Sigma_{ab}^{\mathrm{obs}}$. □

Theorem 2. *Consider the estimator:* $\widehat{S}_{ab} = \frac{1}{q}\sum_{j=1}^{q} z_{ja}z_{jb}$. *Let* G_q *index the variants that are independent of* Y, *and* B_q *index the variants that are dependent on* Y. *Let* C_1, C_1', C_2, L *and* D *denote finite constants. Assume:*

1. *(Bounded moments for* G_q*) For* $j \in G_q$, *we have* $\mathbb{E}(z_{ja}z_{jb}|M) = \Gamma_{ab}$ *and* $\sup_{j\in G_q} \mathrm{Var}(z_{ja}z_{jb}|M) \leq C_1$.
2. *(Bounded moments for* B_q*) For* $j \in B_q$, *we have* $\mathbb{E}(z_{ja}z_{jb}|M) = \mu_j$ *with* $\sup_{j\in B_q} |\mu_j - \Gamma_{ab}| \leq L$ *and* $\sup_{j\in B_q} \mathrm{Var}(z_{ja}z_{jb}|M) \leq C_1'$.
3. *(Local LD) There is an undirected dependency graph with edge set* E_q *and maximum degree* D *such that* $|\mathrm{cov}(z_{ja}z_{jb}, z_{ka}z_{kb}|M)| \leq C_2$ *when* $(j,k) \in E_q$ *and* $\mathrm{cov}(z_{ja}z_{jb}, z_{ka}z_{kb}|M) = 0$ *when* $(j,k) \notin E_q$.
4. *(Vanishing contamination) We have* $\frac{|B_q|}{q} \to 0$ *as* $q \to \infty$.

Then, $\widehat{S}_{ab} \xrightarrow{p} \Gamma_{ab}$ *conditional on* M *as* $q \to \infty$.

Proof Let $X_j = z_{ja}z_{jb}$. Consider the numerator of $\widehat{S}_{ab}$ and write:

$$\sum_{j=1}^{q}(X_j - \Gamma_{ab}) = \underbrace{\sum_{j\in G_q}(X_j - \Gamma_{ab})}_{N_G} + \underbrace{\sum_{j\in B_q}(X_j - \Gamma_{ab})}_{N_B}.$$

We have $\mathbb{E}(N_G|M) = 0$ by Assumption 1. We can also write:

$$|\mathbb{E}(N_B|M)| = |\sum_{j\in B_q}(\mu_j - \Gamma_{ab})| \leq L|B_q|.$$

The inequality follows by Assumption 2. Thus $|\mathbb{E}(N_B|M)|/q \leq L|B_q|/q \to 0$ by Assumption 4. We proved that $\mathbb{E}(\widehat{S}_{ab}|M) \to \Gamma_{ab}$.

With Assumption 3, we can write:

$$\begin{aligned}\mathrm{Var}(\widehat{S}_{ab}|M) &= \sum_j \frac{1}{q^2}\mathrm{Var}(X_j|M) + \sum_{(j,k)\in E_q}\frac{1}{q^2}\mathrm{cov}(X_j, X_k|M)\\ &\leq \frac{\max(C_1, C_1')q}{q^2} + \frac{C_2 D q}{q^2}.\end{aligned}$$

The first and second ratios converge to zero. Hence, $\mathrm{Var}(\widehat{S}_{ab}|M) \to 0$.

Let $m = \mathbb{E}(\widehat{S}_{ab}|M)$. We invoke the triangle inequality to write:

$$\mathbb{P}_{|M}(|\widehat{S}_{ab} - \Gamma_{ab}| > \varepsilon) = \mathbb{P}_{|M}(|(\widehat{S}_{ab} - m) + (m - \Gamma_{ab})| > \varepsilon) \leq \mathbb{P}_{|M}(|\widehat{S}_{ab} - m| > \varepsilon - |m - \Gamma_{ab}|),$$

where we have moved the conditioning notation to the subscripts to avoid confusion with absolute value notation. We previously showed that $|m - \Gamma_{ab}| \to 0$. Now apply Chebyshev's inequality when $|m - \Gamma_{ab}| < \varepsilon$:

$$\mathbb{P}_{|M}(|\widehat{S}_{ab} - \Gamma_{ab}| > \varepsilon) \leq \mathbb{P}_{|M}(|\widehat{S}_{ab} - m| > \varepsilon - |m - \Gamma_{ab}|) \leq \frac{\mathrm{Var}(\widehat{S}_{ab}|M)}{(\varepsilon - |m - \Gamma_{ab}|)^2}.$$

We also already showed that $\mathrm{Var}(\widehat{S}_{ab}|M) \to 0$, so $\widehat{S}_{ab} \xrightarrow{p} \Gamma_{ab}$ as $q \to \infty$. □

Proposition 4. *We have* $\widehat{\alpha}'_{ja} = \sqrt{n}\,\widehat{\mathrm{cor}}(V_j, R'_a)$ *for every* $Y_a \in Y$.

Proof Choose any $Y_a \in Y$. Consider regressing Y'_a on Y'_{-a}:

$$\beta_a = (\Gamma_{-a,-a})^{-1}\Gamma_{-a,a} = -\frac{\Omega_{-a,a}}{\Omega_{aa}}.$$

We thus have:

$$R'_a = Y'_a - Y'_{-a}\beta_a = Y'w_a,$$

where $w_a = \frac{\Omega e_a}{\Omega_{aa}}$ and e_a denotes the ath canonical basis vector. We also have $\mathrm{Var}(R'_a) = 1/\Omega_{aa}$. Now consider:

$$\widehat{\mathrm{cov}}(V_j, R'_a) = \widehat{\mathrm{cov}}(V_j, Y')w_a = \widehat{r}'_j\frac{\Omega e_a}{\Omega_{aa}} = \frac{(\widehat{r}'_j\Omega)_a}{\Omega_{aa}}.$$

Since $\mathrm{Var}(V_j) = 1$ and $\mathrm{Var}(R'_a) = 1/\Omega_{aa}$, we have:

$$\widehat{\mathrm{cor}}(V_j, R'_a) = \frac{(\widehat{r}'_j\Omega)_a}{\sqrt{\Omega_{aa}}}.$$

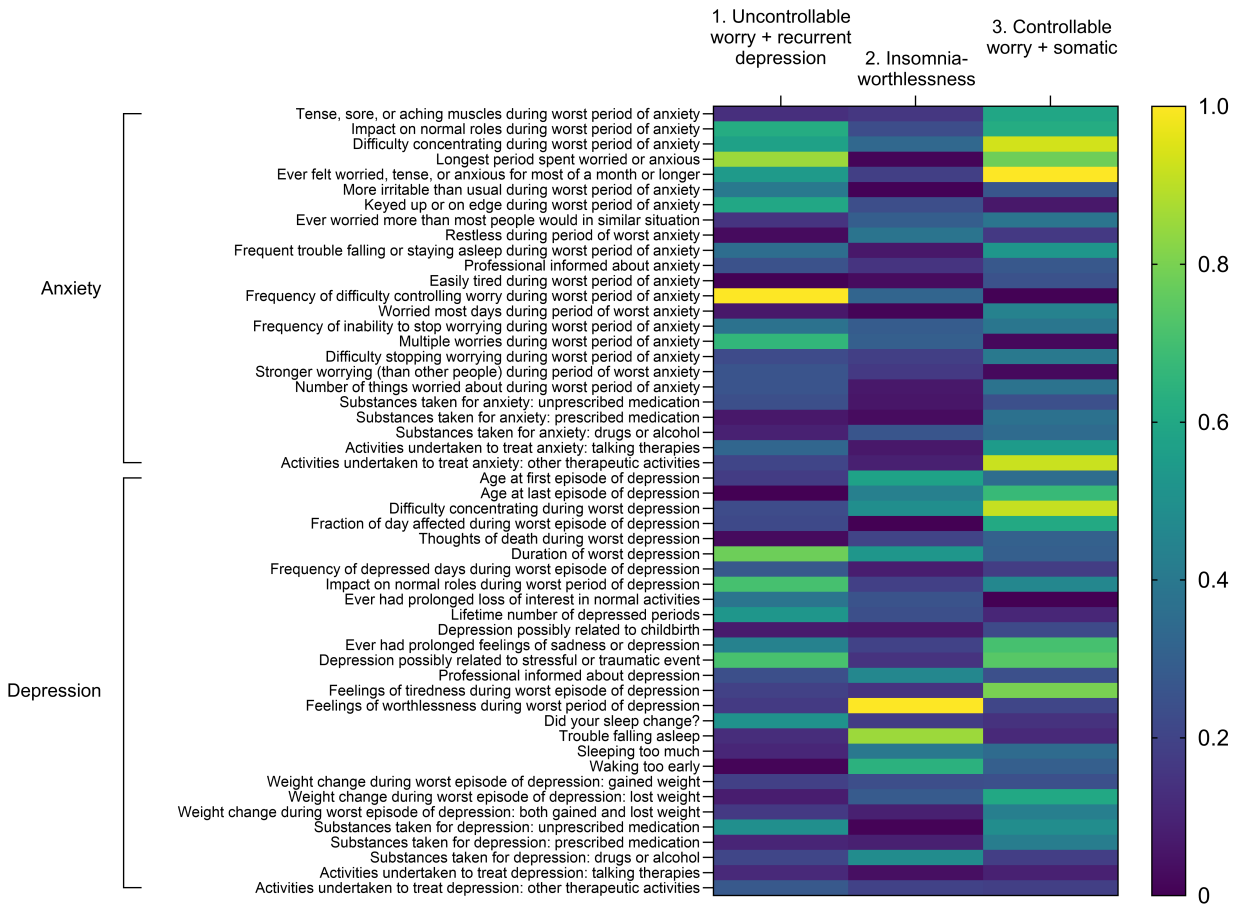
We previously showed that $\widehat{r}'_j = \widehat{\mathrm{cov}}(V_j, Y') = \frac{z_j}{\sqrt{n}}$. It thus follows that:

$$\widehat{\mathrm{cor}}(V_j, R'_a) = \frac{(z_j\Omega)_a}{\sqrt{\Omega_{aa}}\sqrt{n}} = \frac{(\Omega z_j^\top)_a}{\sqrt{\Omega_{aa}}\sqrt{n}} = \frac{\alpha_{ja}}{\sqrt{\Omega_{aa}}\sqrt{n}} = \frac{\alpha'_{ja}}{\sqrt{n}}.$$

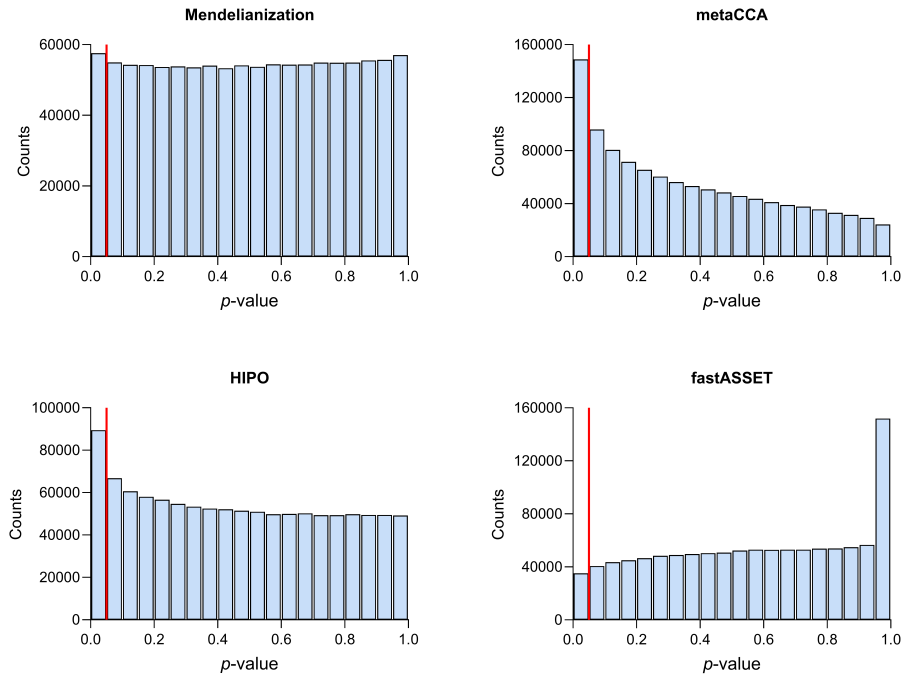
We conclude that $\alpha'_{ja} = \sqrt{n}\widehat{\mathrm{cor}}(V_j, R'_a)$ for every $Y_a \in Y$ because we chose Y_a arbitrarily. □

D. Additional Experimental Results

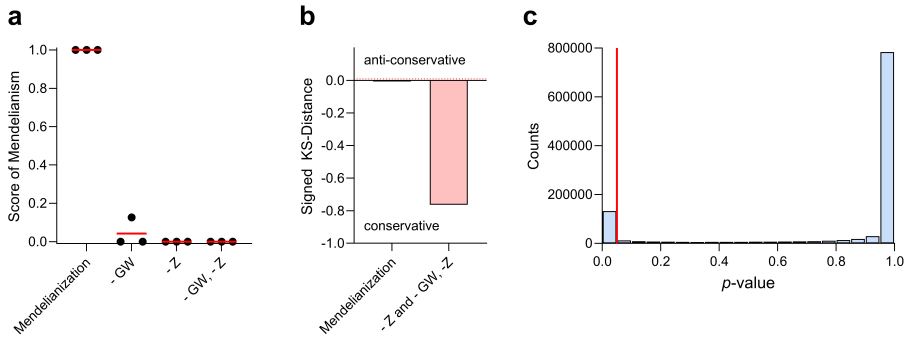
Depression and Generalized Anxiety



Supplementary Fig. 1. Heatmap of $|\alpha'|$ values for the lead variants in depression and anxiety. Each column displays the magnitudes of the interpretable coefficients for a lead variant, normalized to the range $[0, 1]$. The coefficient patterns resolve three phenotypic profiles: (i) uncontrollable worry with recurrent depression; (ii) melancholic-like depression characterized by insomnia, worthlessness, and cognitive slowing; and (iii) largely controllable worry accompanied by prominent somatic complaints. These profiles illustrate distinct modes of experiencing depression and anxiety, underscoring the disorders' heterogeneity.

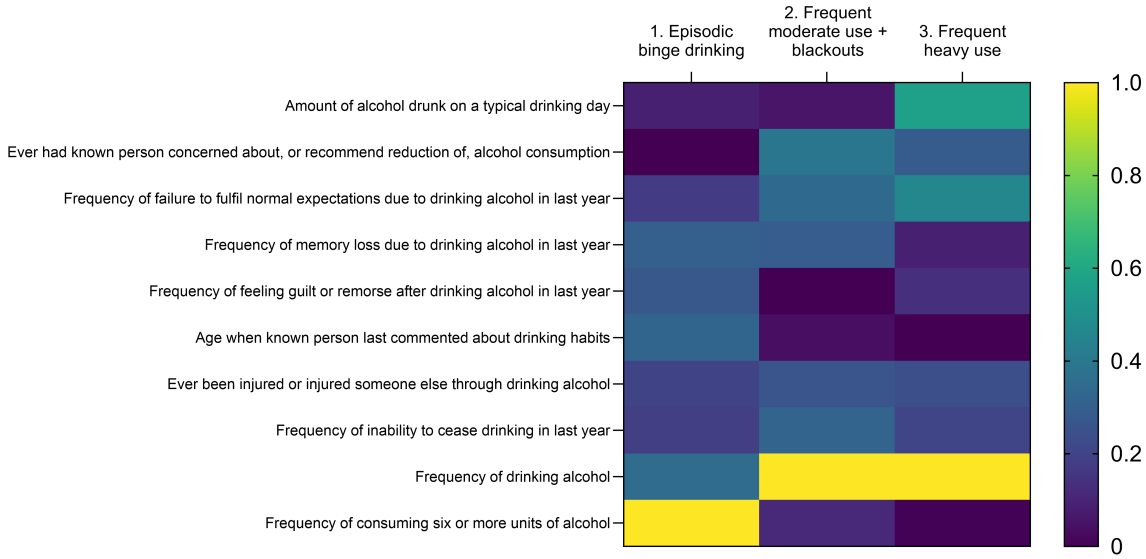


Supplementary Fig. 2. Histograms of p-values across all variants. Mendelianization was the only method to produce a near-uniform distribution for variants associated with p-values above 0.05 (red line). By contrast, metaCCA and HIPO yielded anti-conservative distributions with shrunken p-values, whereas fastASSET was overly conservative, producing p-values near one for many variants.

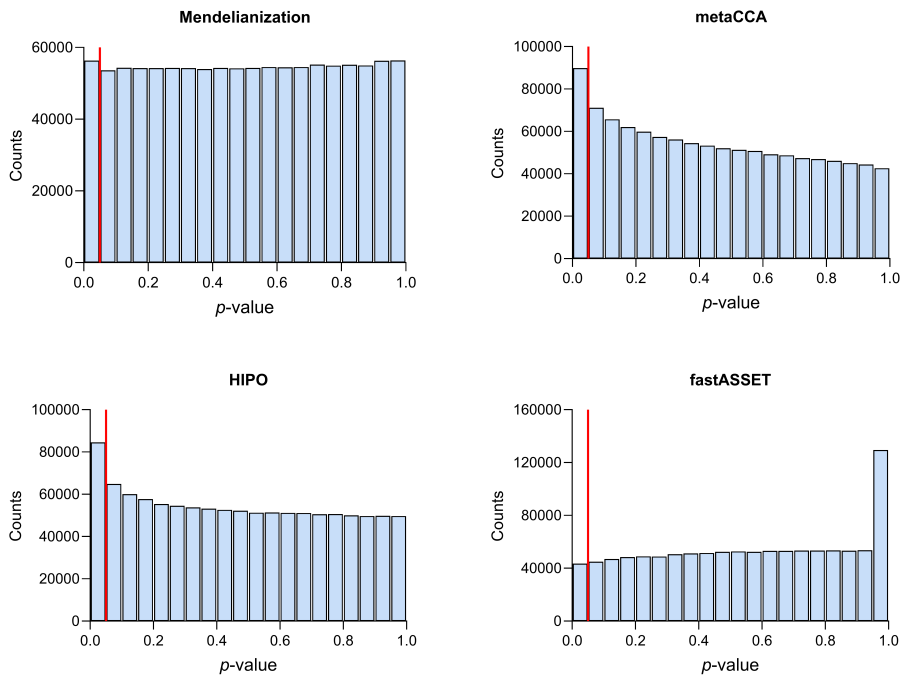


Supplementary Fig. 3. Ablation results. -GW denotes not applying each outcome genome-wide, and -Z denotes not using z-statistics to estimate Γ. (a) Misapplying the learned outcomes or using β coefficients in place of z-statistics (or both) decreased the SoM score to near zero. (b) The histogram obtained by β coefficients was very conservative. (c) Inspection of the histogram derived from the β coefficient estimates revealed highly miscalibrated p-values following a U-shape curve.

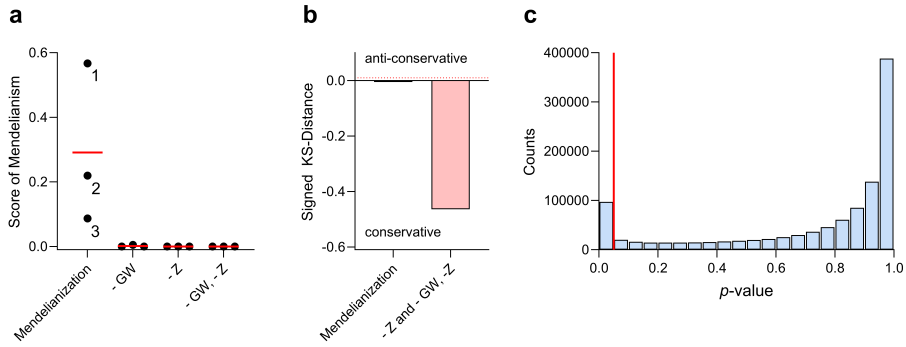
Alcohol Use Disorder



Supplementary Fig. 4. Heatmap of $|\alpha'|$ values for the lead variants in alcohol use disorder. The coefficient patterns correspond to three phenotypic profiles: (i) episodic binge drinking, (ii) frequent moderate use of alcohol with blackouts, and (iii) frequent heavy use without blackouts.



Supplementary Fig. 5. Histograms of p-values across all variants. Mendelianization was again the only method to produce a near-uniform distribution on $[0.05, 1]$.



Supplementary Fig. 6. Ablation results. (a) Mendelianization again achieved the highest SoMs, and (b, c) estimating Γ from β coefficients yielded highly miscalibrated p-values.