mixWAS: An efficient distributed algorithm for mixed-outcomes genome-wide association studies

Ruowang Li[1]*, Luke Benz[2]*, Rui Duan[2]*, Joshua C. Denny[3], Hakon Hakonarson[4,5,6], Jonathan D. Mosley[7,8], Jordan W. Smoller[9,10], Wei-Qi Wei[8], Marylyn D. Ritchie[11], Jason H. Moore[1], Yong Chen[12]

1. Department of Computational Biomedicine, Cedars-Sinai Medical Center
2. Department of Biostatistics, Harvard T.H. Chan School of Public Health
3. National Human Genome Research Institute, National Institutes of Health
4. Division of Human Genetics, Children's Hospital of Philadelphia
5. Center for Applied Genomics, Children's Hospital of Philadelphia
6. Department of Pediatrics, University of Pennsylvania, Perelman School of Medicine
7. Department of Medicine, Vanderbilt University Medical Center
8. Department of Biomedical Informatics, Vanderbilt University Medical Center
9. Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital
10. Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital
11. Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania
12. Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania

* These authors contributed equally

Corresponding: Ruowang.Li@cshs.org, Jason.Moore@csmc.edu, ychen123@pennmedicine.upenn.edu

Abstract

Genome-wide association studies (GWAS) have been instrumental in identifying genetic associations for various diseases and traits. However, uncovering genetic underpinnings among traits beyond univariate phenotype associations remains a challenge. Multi-phenotype associations (MPA), or genetic pleiotropy, offer important insights into shared genes and pathways among traits, enhancing our understanding of genetic architectures of complex diseases. GWAS of biobank-linked electronic health record (EHR) data are increasingly being utilized to identify MPA among various traits and diseases. However, methodologies that can efficiently take advantage of distributed EHR to detect MPA are still lacking. Here, we introduce mixWAS, a novel algorithm that efficiently and losslessly integrates multiple EHRs via summary statistics, allowing the detection of MPA among mixed phenotypes while accounting for heterogeneities across EHRs. Simulations demonstrate that mixWAS outperforms the widely used MPA detection method, Phenome-wide association study (PheWAS), across diverse scenarios. Applying mixWAS to data from seven EHRs in the US, we identified 4,534 MPA among blood lipids, BMI, and circulatory diseases. Validation in an independent EHR data from UK confirmed 97.7% of the associations. mixWAS fundamentally improves the detection of MPA and is available as a free, open-source software.

Introduction

Genome-wide association studies (GWAS) have systematically identified numerous genetic associations for various diseases and traits[1–3]. However, most GWAS SNPs have small to modest additive phenotypic effects, and merely detecting more genetic associations may not provide direct insights into the shared architectures of diseases and traits[4–6]. It has been observed that

many genetic variants are associated with more than one trait, referred to as multi-phenotype associations (MPA) or genetic pleiotropy[7–9]. Compared to single phenotype GWAS associations, MPA could reveal potential shared genes or pathways among traits and provide key insights into the disease pathogenesis[8,10–13]. Thus, identifying MPA is a crucial next step towards improved understanding of the genetic architectures of complex diseases.

The increasing availability of biobank-linked electronic health record data (EHR), such as the UK Biobank (UKBB) and Electronic Medical Records and Genomics Network (eMERGE), improves our ability to detect MPA, since patients' genetic data are matched with their extensive clinical records that can be extracted into disease phenotypes[14–22]. Utilizing multiple EHRs could improve the power to detect MPA, and at the same time, improve the reproducibility of the detected associations. However, computational methodologies that can fully take advantage of multiple EHR data to identify MPA are lacking. Phenome-Wide Association Studies (PheWAS) is the most widely used method to identify MPA, but its power is hindered by the multiple testing penalties due to the multiplicative number of tests between genetic variants and phenotypes[23,24]. Alternative methods have been developed for more efficient statistical tests but have been generally limited to continuous phenotypes or the analysis of individual datasets[13,25]. In reality, most EHR extracted phenotypes are of mixed continuous and binary outcomes. Extending methods to identify MPA using multiple EHRs is also challenging. Thus far, meta-analysis has been the primary approach to integrate GWAS or PheWAS results from multiple EHR to improve the detection of MPA[13,26–29]. However, meta-analyzed PheWAS has been suggested to be less powerful than the pooled mega analysis when detecting MPA across multiple EHRs[14,29].

To address these issues, we propose a computationally efficient framework, mixWAS, that can losslessly integrate multiple EHRs to detect MPA among mixed (binary and continuous) phenotypes. We implemented four key features in mixWAS to fulfill these goals. First, mixWAS can identify genetic associations among mixed-type phenotypes, including binary (e.g., case-control) and continuous (e.g., lab measurements). Second, mixWAS can efficiently and losslessly integrate multiple EHRs to increase the overall sample size. Third, mixWAS is designed to handle phenotype and confounding covariate heterogeneity that may exist among different EHRs. For example, different age-onset for different diseases or block-wise missing data in some EHRs. Finally, mixWAS only requires data summary statistics from different EHRs, minimizing data transferring and communication costs.

Using simulations, we first demonstrated that the mixWAS algorithm has better power than the commonly used PheWAS approach. Subsequently, the proposed method was utilized to detect MPA among cardiovascular related mixed-type phenotypes using patients from seven EHR data from eMERGE. We then validated our findings in the independent UKBB data. In total, mixWAS identified 4,534 associations in the integrated analysis using all eMERGE EHRs of which 4,428 SNPs (97.7%) were validated by the independent UKBB data. In summary, the proposed mixWAS algorithm can efficiently integrate EHR data from multiple sources to detect MPA among mixed phenotypes, paving the way for discovering new insights into disease mechanisms and potential therapeutic targets.

Result

Overview of the mixWAS algorithm. mixWAS is designed to identify genetic variants that are associated with multiple diseases or traits using multiple distributed datasets. mixWAS is specifically tailored to handle mixed phenotypes (binary and continuous) with heterogeneous data distributions, as often observed in EHR data. The algorithm enables lossless integration of data from multiple sources by utilizing summary statistics of the datasets. The algorithm follows a three-step process: First, intermediate summary statistics of genetic associations are calculated within

each distributed dataset. These summary statistics are then transmitted to a centralized server and analyzed by an analyst, who computes the components (score and variance) of the test statistic. Finally, the test statistics are derived, yielding p-values for subsequent inference (Figure 1).

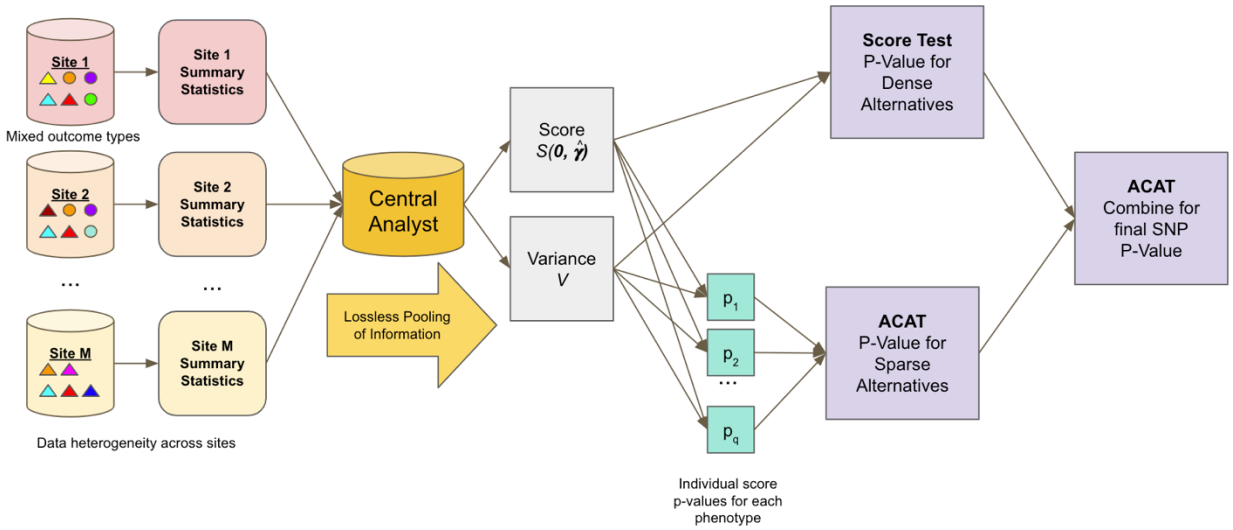


Figure 1: Outline of the mixWAS algorithm. Each site transmits summary level statistics of heterogeneous, mixed-typed phenotypes, specifically the score vector $S_m(\mathbf{0}, \hat{\boldsymbol{\gamma}}_m)$ and variance matrix V_m, to a central analyst. The central analyst pools the site-specific score and variance contributions in a lossless manner for use in a score test, which is powerful against dense phenotypes, and a second test, robust to sparse phenotypes, which combines the p-values of individual score tests for each of q phenotypes using the ACAT method[30]. Finally, these two p-values, one optimized against dense phenotypes and the other optimized against sparse phenotypes, are combined again using the ACAT method.

mixWAS is more powerful than PheWAS in detecting MPA across mixed-type phenotypes. PheWAS is the most commonly used method to detect MPA. In addition, unlike other methods for detecting MPA, PheWAS can be extended to handle mixed phenotypes in multiple datasets through meta-analysis. Thus, mixWAS was directly compared to PheWAS to evaluate its power to detect MPA. To this end, we conducted various simulation studies with diverse settings. Each simulation setting consisted of five independent datasets representing distributed sites. Within each dataset, we simulated MPA associations and confounding covariates, including principal components, age, and sex. Covariate effects for each phenotype were intentionally varied across datasets to reflect heterogeneity across sites. The true MPA associations were devised to two distinct scenarios in simulation studies: 1) Same direction: all phenotypes (binary and continuous) were positively or negatively associated with the genotype in the same direction, and 2) Opposite direction: half of the phenotypes had positive associations with the genotype and half had negative associations. We also incorporated additional factors such as signal sparsity, phenotype correlations, missing data, and different ratios of mixed phenotypes in our simulations. The simulated data was analyzed using mixWAS, PheWAS-Meta, and PheWAS-Mega, described briefly below. mixWAS generated intermediate summary-level statistics in each dataset and performed integration using these statistics to calculate the association p-values. In contrast, separate PheWAS analyses were conducted in each dataset under PheWAS-Meta, where the beta coefficients were combined using meta-analysis. In PheWAS-Mega, the five datasets were pooled to form a combined dataset, followed by a single PheWAS analysis. While not always

plausible in practice, this method was of interest as a point of comparison, since mixWAS uses a lossless decomposition that provides identical results to those that one would obtain by conducting the same score-based procedure if all individual information could be used. Finally, we ascertained the highest possible performance using an oracle model, by applying the score test for dense alternatives in the subset of phenotypes are associated with the SNP. Across all simulation settings, mixWAS consistently outperformed PheWAS (Figures 2 and 3). The gains in power by using mixWAS over PheWAS methods were greatest when the direction of SNP effects went against the direction of the residual correlation between phenotypes. For example, mixWAS outperformed PheWAS the most when SNP effects were positive and residual correlation was negative (Figure 2), reflective of a setting where positive correlation genetic correlation exists among traits in the presence of negative environmental or other correlation. mixWAS also outperformed PheWAS methods when SNP effects had opposite signs with positive residual phenotype correlation (Figure 3). Such power gains result from the fact that unlike PheWAS methods, mixWAS accounts for correlation between phenotypes. The heterogeneity in MPA effects and the residual correlations are frequently encountered in practical settings, underscoring the practical usefulness of mixWAS. Type 1 errors were controlled at the nominal level for all methods. Additional simulations on binary-only phenotypes and designs using shared healthy controls further demonstrated the superior power of mixWAS compared to PheWAS (Supplementary Material).

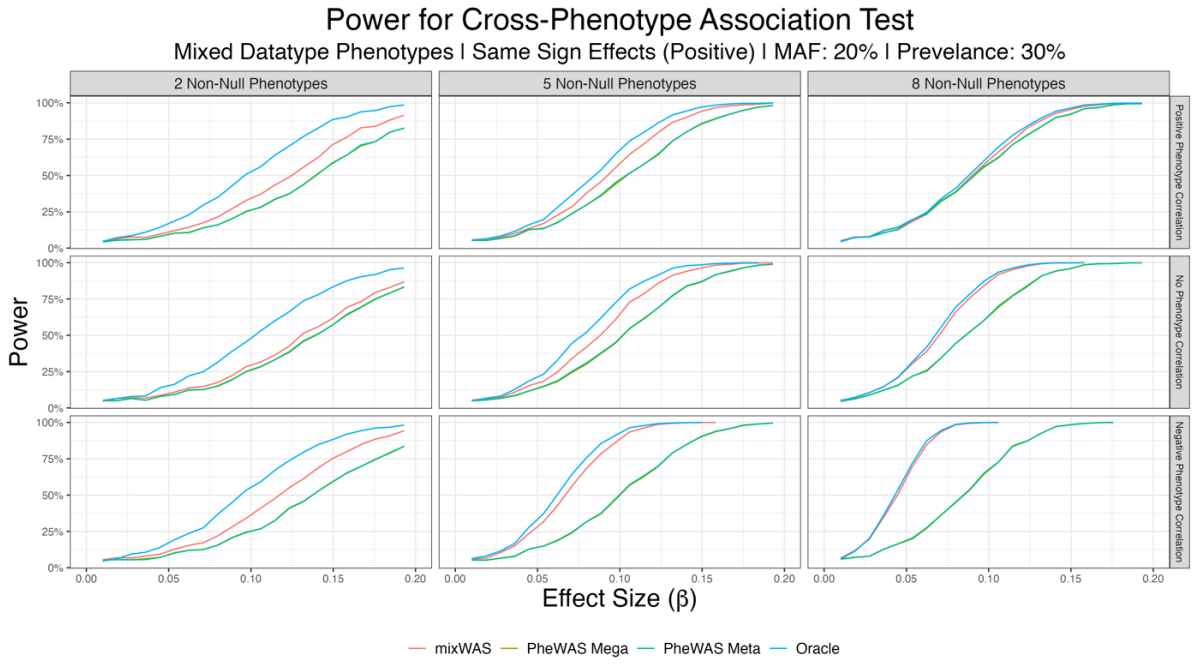


Figure 2: Empirical power curves comparing various cross-phenotype association tests for simulated mixed-type phenotypes. The simulated SNP is positively associated with all phenotypes, while the MPA sparsity (e.g. number of phenotypes with significant association) and correlation between phenotypes vary. Power curves of PheWAS Mega and PheWAS Meta closely overlap in the figure.

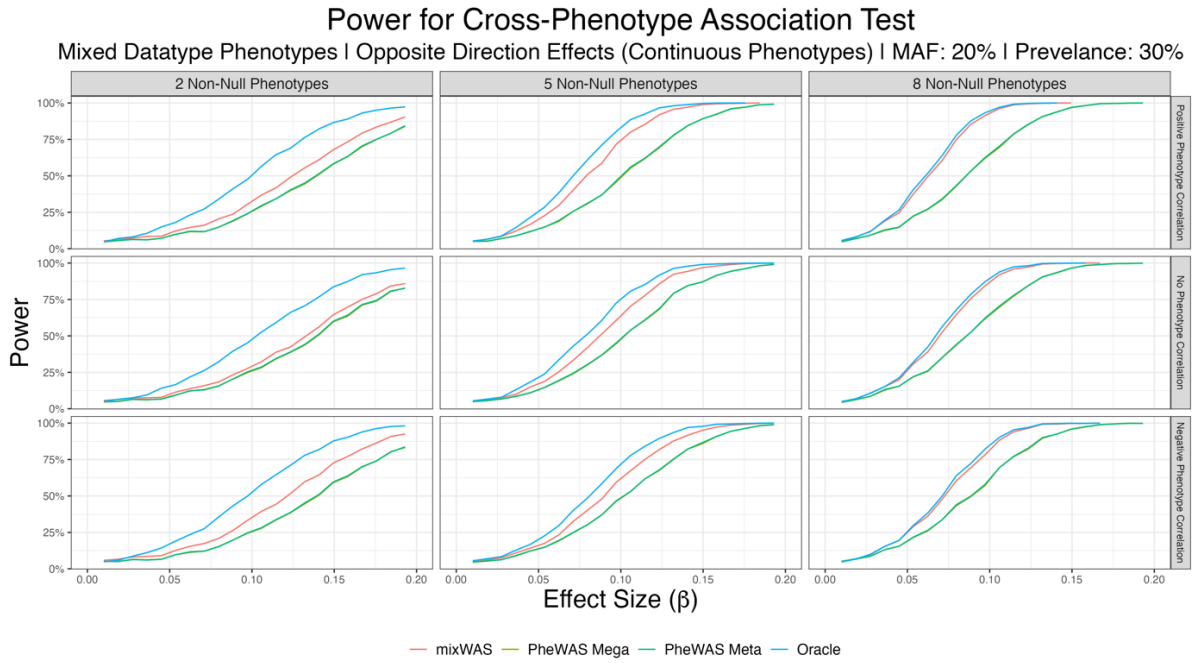


Figure 3: Empirical power curves comparing various cross-phenotype association tests for simulated mixed-type phenotypes. The simulated SNP is positively associated with all binary phenotypes and has both positive and negative associations with continuous phenotypes. MPA sparsity (e.g. number of phenotypes with significant association) and correlation between phenotypes vary. Power curves of PheWAS Mega and PheWAS Meta closely overlap in the figure.

Detecting MPA across blood lipids levels, BMI, and diseases of circulatory system. Previous research has identified MPA between BMI and coronary heart disease [31,32], blood lipids levels[33], and low-density lipoprotein, triglycerides, and cardiovascular diseases[34]. These studies have highlighted the existence of potential shared underlying genetic architecture among these traits/diseases. However, comprehensive investigations of MPA across all traits and diseases have not been carried out. Leveraging multiple EHR datasets and the improved efficiency of our proposed method, we utilized mixWAS to detect MPA among blood lipid levels (high-density lipoprotein (HDL), low-density lipoprotein (LDL), serum total cholesterol, and triglycerides), body mass index (BMI), and circulatory diseases (unspecified essential hypertension, type 2 diabetes (T2D), unspecified hyperlipidemia, benign essential hypertension, atrial fibrillation, congestive heart failure, and coronary atherosclerosis) using eMERGE data from 7 sites. The characteristics of the datasets are presented in Figure 4, illustrating heterogeneities among the EHRs, including variations in the number of patients and patterns of missing data.

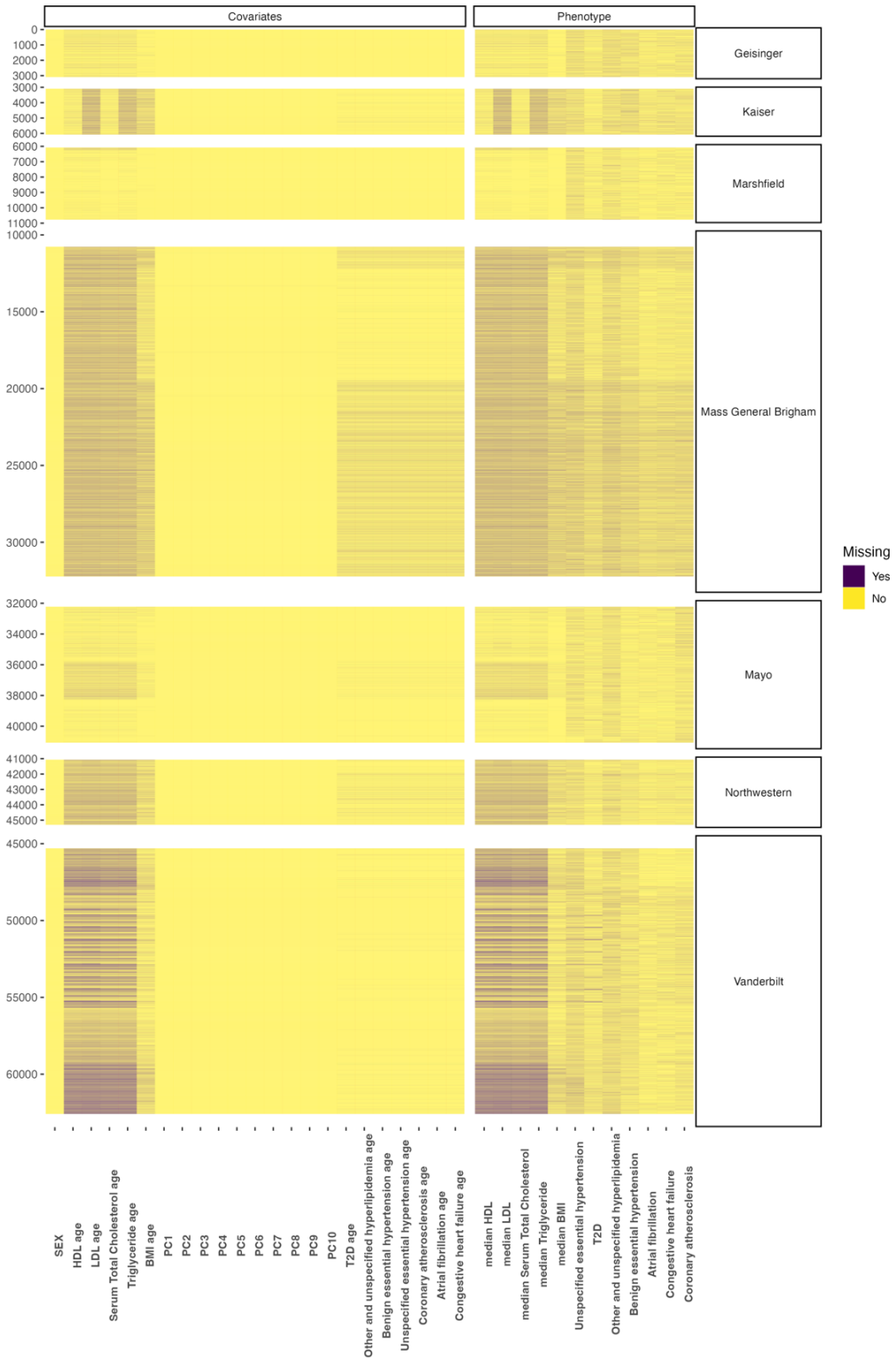
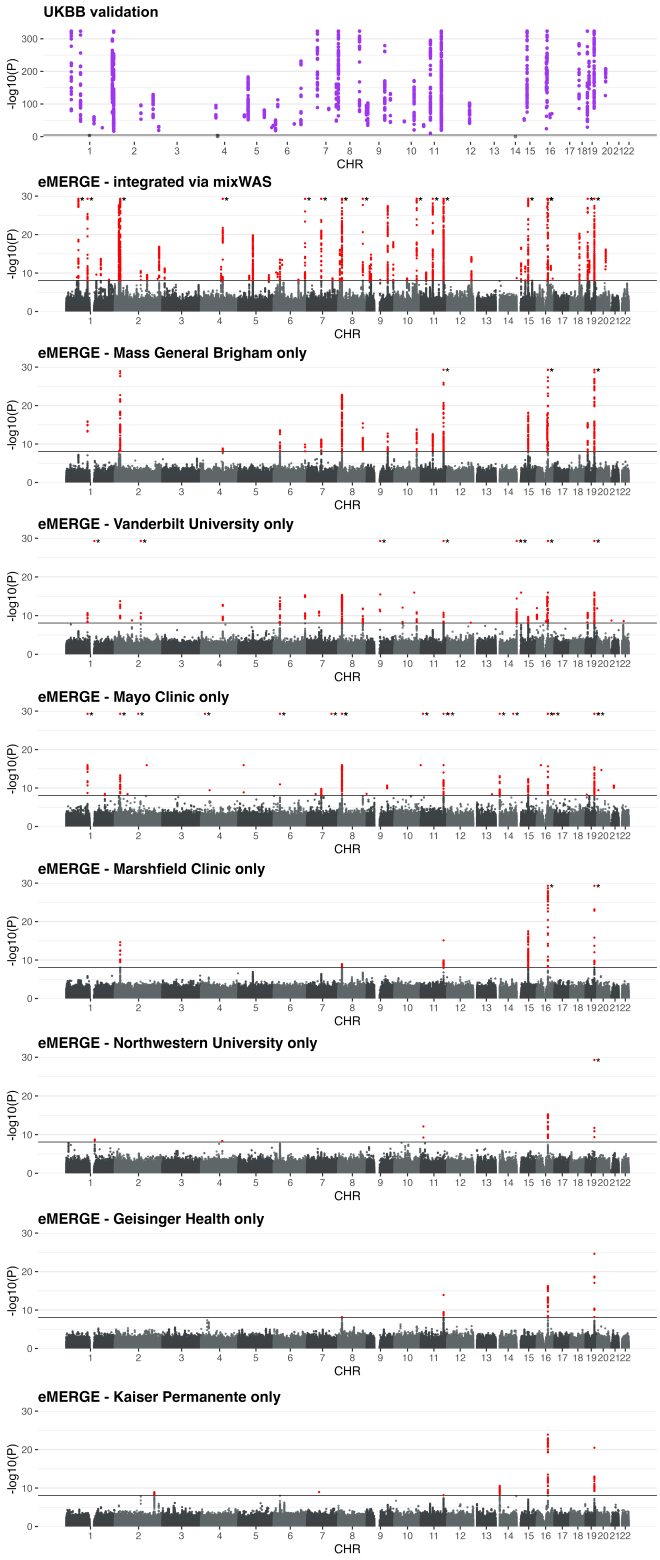


Figure 4. Characteristics of the covariates and phenotypes from the eMERGE dataset. The x-axis displays the phenotypes and covariates, and y-axis displays the number of participants in each dataset. The color indicates whether or not a value (binary or continuous) was present for the individual.



To identify MPA, we conducted three separate analyses: eMERGE single-site analysis, eMERGE integrated analysis via mixWAS, and external validation using the UKBB dataset. In all analyses, patients' sex, ten principal components accounting for population stratification, and trait-associated ages were adjusted in the model. Notably, each trait was measured at a different associated age, leading to distinctive adjustments for each trait. In the first analysis, mixWAS was individually applied to each eMERGE dataset to detect MPA. This allowed the assessment of the power of detecting MPA when each dataset was used independently. The Manhattan plots showed that datasets with smaller sizes (Marshfield, Northwestern, Geisinger, and Kaiser Permanente) exhibited lower power in detecting MPA compared to larger datasets (Mass General Brigham, Vanderbilt, and Mayo). Next, an integrated analysis was performed by mixWAS utilizing all eMERGE datasets. This integrated analysis significantly increased the total sample size and identified a higher number of significant MPA compared to any individual dataset. The Bonferroni-corrected significant associations identified in the eMERGE integrated analysis were validated using the independent UK Biobank data. Out of 4,534 eMERGE associations, 4,428 (97.7%) were successfully replicated in the UKBB dataset, providing further validation for the identified associations (Figure 5).

Figure 5. Manhattan plots of MPA in eMERGE single site analysis, eMERGE integrated analysis, and UKBB validation. The association between each SNP and all phenotypes was tested using mixWAS. The resulting p-value of each SNP was -log10(p) transformed and plotted along the y-axis. SNPs with p-values lower than 5e-30 were replaced with 5e-30 and indicated by asterisk for visualization purpose. The genomic position of the SNP was plotted along the x-axis. The solid horizontal line indicates Bonferroni corrected p-value significance level, respectively to each data. SNPs above the significance threshold are plotted as red points in the eMERGE data and purple points in the UKBB data.

mixWAS identified MPA are associated with multiple traits/diseases. mixWAS-identified MPA could potentially be associated with one or multiple traits. To explore the specific trait/disease driving these MPAs, an exhaustive analysis evaluating all possible combinations of trait and SNP associations (4,534 SNPs x 10 traits) were conducted in the UKBB dataset (Supplemental Table). The significance of the single trait associations was determined using the mixWAS related p-value threshold (Method). The analysis revealed that SNPs showed significant associations with between 0 to 8 traits, with 2 to 4 traits being the most common number (Figure 6a).

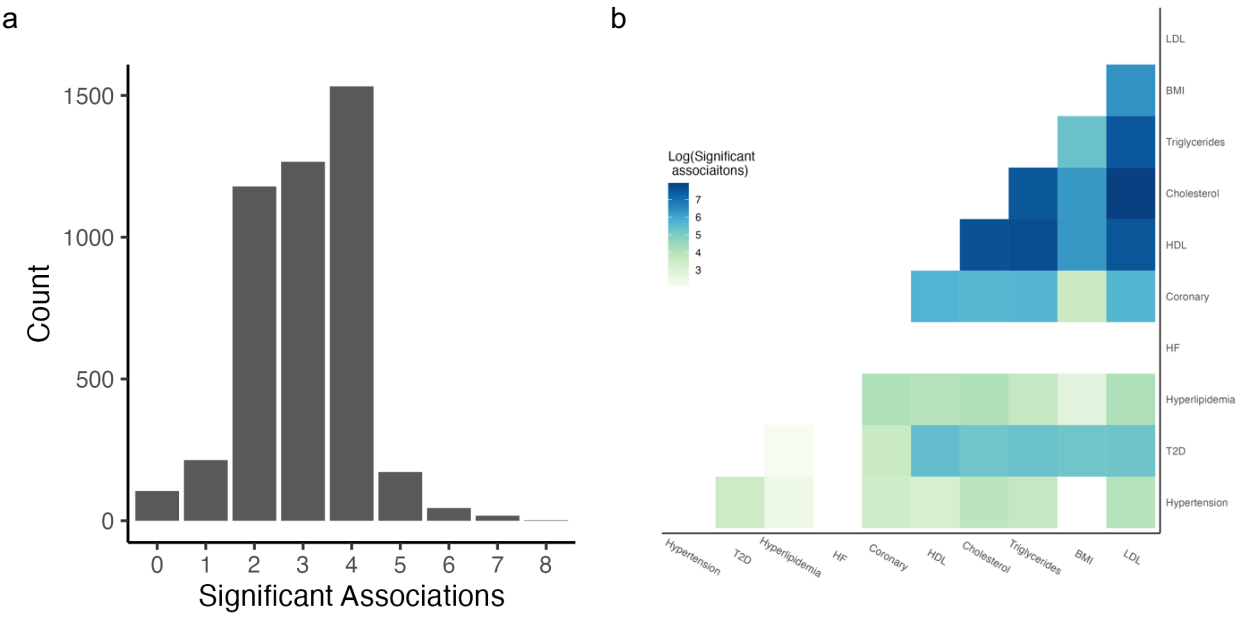


Figure 6. a) The distribution of the number of traits associated with a SNP. Count (y-axis) reflects the number of SNPs with significant associations for varying numbers of phenotypes (x-axis). Each association is determined using regression analysis, while adjusting for the Bonferroni-corrected p-value threshold. **b) Common genetic variants shared among different traits.** For each pair of phenotype combinations, the number of shared same SNP associations that surpassed the significant p-value threshold was calculated and transformed using a logarithmic scale, e.g., for 1000 shared associations, log(1000)= 6.9. The color intensity in the plot reflects the number of shared genetic variants, with darker shades indicating a higher number of shared associations.

Variability in shared genetic variants. The number of genetic associations common to different traits exhibited variability. Blood lipid levels (LDL, HDL, Cholesterol, and Triglycerides) and BMI

displayed the highest number of shared genetic variants among the traits. Additionally, blood lipid levels demonstrated considerable overlap with genetic variants associated with coronary artery disease. For T2D, both blood lipid levels and BMI showed enrichment for shared genetic variants (Figure 6b).

Improved detection powered by mixWAS. Among the initial pool of 4,534 candidate SNPs with detected MPAs across 10 traits, 13,770 significant single trait-SNP associations were identified in the UKBB dataset from. In comparison, directly applying PheWAS to the same SNPs in UKBB would only detect 11,581 significant associations due to the increased number of tests. Consequently, mixWAS detected 18.9% more trait-SNP associations in the UKBB dataset (Figure 7). These additional associations were found for every trait except for heart failure, where neither mixWAS nor PheWAS identified any SNP associated with the disease (see Discussion). The results underscore the improved sensitivity and efficiency of mixWAS in detecting trait-SNP associations compared to traditional PheWAS approaches.

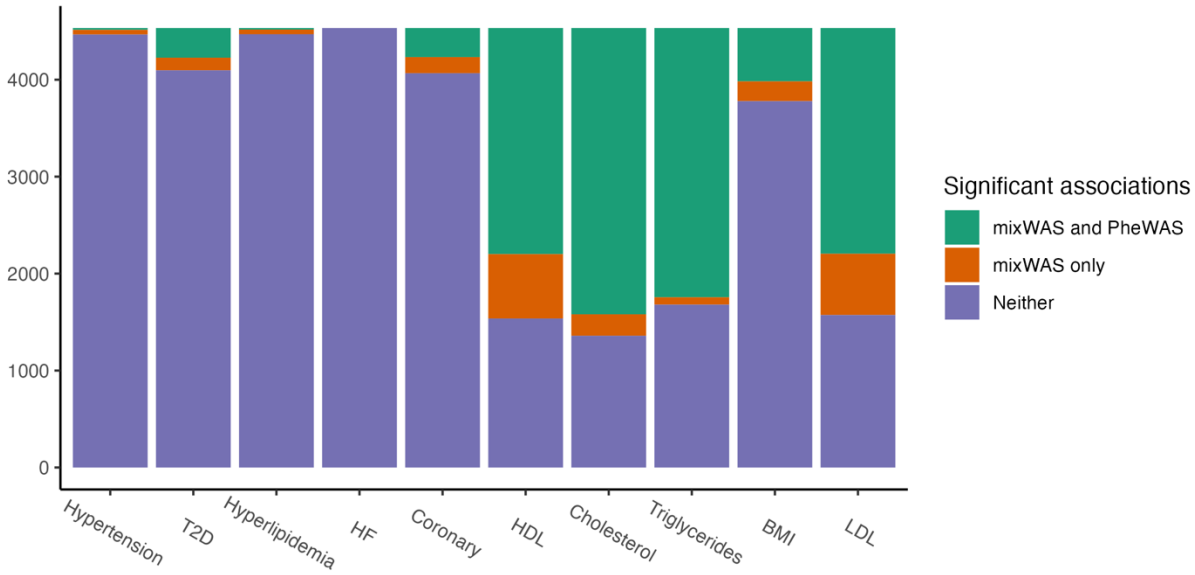


Figure 7. Count of significant trait-SNP associations identified by mixWAS and PheWAS in the UKBB. Method-specific p-value thresholds were applied to mixWAS and PheWAS, corresponding to the number of comparisons performed. The associations detected by PheWAS are a subset of those identified by mixWAS, and are categorized as "mixWAS and PheWAS". The "mixWAS only" category represents genetic associations that were exclusively identified by mixWAS and not by PheWAS. SNPs that were not considered significant by either method are represented by the "Neither" category.

Functional annotation of mixWAS SNPs. The mixWAS SNPs were annotated using the canonical pathways curated in the Human Molecular Signatures Database (MSigDB). These SNPs exhibited enrichments in pathways related to cholesterol metabolism, lipoprotein function, hyperlipidemia, as well as pathways associated with LDL, HDL, and triglycerides (Figure 8). These findings provide additional support for the genetic associations identified through mixWAS (Figure 6 and Figure 7).

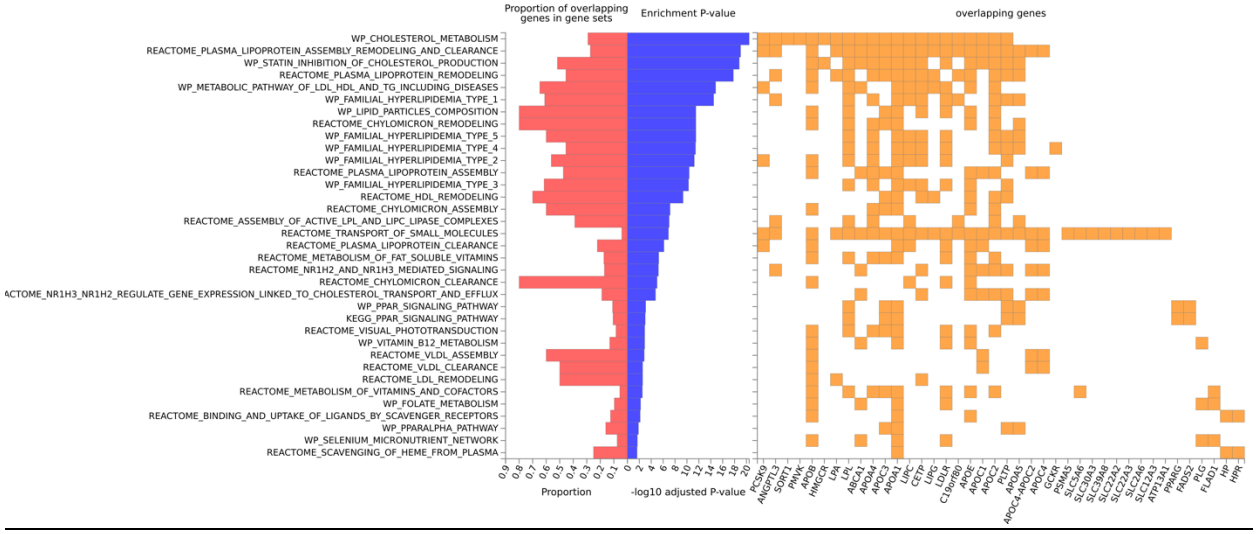


Figure 8. Pathway annotations of mixWAS identified SNPs. The mixWAS identified SNPs were annotated using the MSigDB canonical pathways. The enrichment of a pathway is determined by the number of SNPs associated with the pathway. The pathways are ranked by their enrichment -log10(p-values) from top to bottom (blue bar). SNPs were also mapped to genes within each pathway and their presence in a gene is represented by yellow squares. The proportion of overlapping genes in gene sets is shown as red bars.

Methods

eMERGE data

Genotype data linked with EHR information from the eMERGE network Phase III. This dataset comprised a total of 83,717 genotyped patients across 11 participating sites[35]. For our investigation, we focused on eight adult sites, which included the Marshfield Clinic Research Foundation, Vanderbilt University Medical Center, Kaiser Permanente Washington/University of Washington, Mayo Clinic, Northwestern University, Geisinger, Mt. Sinai, and Mass General Brigham. SNPs were imputed using the Haplotype Reference Consortium 1.1 reference, aligned with genome build 37. This imputation process yielded a set of 39 million genetic variants[36]. Subsequently, we subjected the SNP genotypes to quality filtering and processing, following an established pipeline[37]. The criteria for inclusion required that both the genotype and sample call rates exceeded or equaled 99%, the imputation score exceeded 0.4, Hardy-Weinberg equilibrium p-value used was 0.00001, and the Minor Allele Frequency (MAF) of the SNPs was equal to or exceeded 0.05. To mitigate the potential impact of population structure on our analyses, we restricted our investigation to unrelated individuals of European ancestry. The ancestry was determined using principal components derived from the 1000 Genome Project. In cases where individuals were identified as related, defined by a π-hat value of ≥ 0.25 identity-by-descent, one individual from each related pair was removed. The final dataset comprised 59,136 unrelated individuals, and a curated set of 6,106,952 high-quality SNPs for subsequent analyses.

UK BioBank data

The UK Biobank released comprehensive genetic and phenotypic data, encompassing approximately 500,000 individuals representing diverse regions across the United Kingdom[38]. Genotyping was conducted utilizing two related types of genotype arrays, namely the UK BiLEVE Axiom Array or the UK Biobank Axiom Array, organized into 106 batches and imputed using the merged UK10K and 1000 Genomes phase 3 reference panels[39].

To ensure the quality and reliability of our sample, a series of quality control measures were implemented. First, individuals displaying a SNP missing rate exceeding 5% and exhibiting high levels of heterozygosity were excluded from the study. Second, among related individuals, one individual from each pair was systematically removed to prevent undue influence from familial genetic connections. The threshold for relatedness was set at the level of second-degree relatives, as indicated by an identity-by-descent π-hat value equal to or greater than 0.25. Third, only individuals with White British ancestry were retained in order to match the ancestry of the eMERGE data. Additionally, individuals with discrepancies between their self-reported and genetically-inferred sexes were omitted from the analysis. Finally, genetic variants characterized by imputation info scores lower than 0.3 and MAF less than 0.01 were excluded from consideration.

mixWAS Algorithm

mixWAS is designed to jointly test the association between a single SNP X and multiple mixed-type phenotypes $(Y_1, \ldots, Y_q)$, across M sites, which cannot share individual level patient data across sites due to privacy concerns. We let i index individuals, j index phenotypes, and m index sites. For the i^{th} individual at the m^{th} site we denote the SNP by $X_{im} \in \{0, 1, 2\}$ and the q phenotypes by $\boldsymbol{Y_{im}} = (Y_{i1m}, \ldots, Y_{iqm}) \in R^q$. The outcomes of interest may be of differing data types such as binary, continuous, count, or time to event.

Let $\boldsymbol{Z_{im}} \in R^p$ denote a vector of an individual's covariates, such as age, gender, or ancestry principal component (PCs). As a general framework for phenotype j,

$$Y_{ijm}|X_{im}, \boldsymbol{Z_{im}} \sim f_j(\beta_j X_{im} + \boldsymbol{Z_{im}^T \gamma_{jm}}) \quad (1)$$

where $\beta_j \in R$ denotes the SNP effect, shared across all M sites, and $\boldsymbol{\gamma_{jm}} \in R^p$ denotes the site-, phenotype-specific effect sizes for remaining covariates. These effects are allowed to be site-specific to account for the site-level heterogeneity such as disease prevalence and confounding effects. Finally, f_j denotes the density corresponding to the data type for each phenotype of interest. For example, binary outcomes can be assumed to follow the logistic regression model

$$\log\left(\frac{P(Y_{ijm}=1)}{1-P(Y_{ijm}=1)}|X_{im}, \boldsymbol{Z_{im}}\right) = \alpha_{jm} + \beta_j X_{im} + \boldsymbol{Z_{im}^T \gamma_{jm}} \quad (2)$$

while continuous outcomes can be assumed to follow the linear regression model

$$E(Y_{ijm}|X_{im}, \boldsymbol{Z_{im}}) = \alpha_{jm} + \beta_j X_{im} + \boldsymbol{Z_{im}^T \gamma_{jm}} \quad (3)$$

Following the approach of Li et al[14], q phenotypes are combined using a composite likelihood function, which accounts for complex correlations between mixed-type phenotypes without modeling them directly[40,41]. The log composite likelihood function across all M sites can be expressed as

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{m=1}^{M}\sum_{j=1}^{q} L_{jm}\left(\beta_j, \boldsymbol{\gamma_{jm}}\right) = \sum_{m=1}^{M}\sum_{j=1}^{q}\sum_{i=1}^{n_m} \log(\, f_j(Y_{ijm}, X_{im}, \boldsymbol{Z_{im}} |\, \beta_j, \boldsymbol{\gamma_{jm}}))$$

(4)

In order to determine whether a given SNP has association with any of the phenotypes of interest, we consider testing $H_0: \boldsymbol{\beta} = (\beta_1, \ldots, \beta_q) = \mathbf{0}$ against $H_1: \boldsymbol{\beta} \neq \mathbf{0}$, using an omnibus score-type test, which is computationally efficient and lossless in a federated setting in which sites cannot share individual level data. The score function is defined by

$$S(\mathbf{0}, \hat{\boldsymbol{\gamma}}) = \left(\frac{\partial L(0, \hat{\boldsymbol{\gamma}})}{\partial \beta_1}, \ldots, \frac{\partial L(0, \hat{\boldsymbol{\gamma}})}{\partial \beta_q}\right)$$

(5)

where $\hat{\boldsymbol{\gamma}}$ are the maximum likelihood estimators of $\boldsymbol{\gamma}$ under H_0 with $\boldsymbol{\beta}$ set to 0. Non-SNP coefficients $\boldsymbol{\gamma_{jm}}$ are both site- and phenotype-specific, and they can be estimated independently at each site by $\hat{\boldsymbol{\gamma}}_{\boldsymbol{jm}} = \mathrm{argmax}_{\boldsymbol{\gamma_{jm}}} L_{jm}\left(\mathbf{0}, \boldsymbol{\gamma_{jm}}\right)$. Because of the composite likelihood framework, the score function can be decomposed into a sum of site-specific score vectors as follows

$$S(\mathbf{0}, \hat{\gamma}) = \sum_{m=1}^{M} S_m(\mathbf{0}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{m}}) = \sum_{m=1}^{M}\sum_{j=1}^{q} \nabla_{\beta}\, L_{jm}(\mathbf{0}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{jm}})$$

(6)

Thus, each site only needs to compute and share its own score vector $S_m(\mathbf{0}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{m}})$, and associated $q \times q$ covariance matrix V_m, which contains only summary-level information. V_m can be estimated locally by deriving the influence functions.

mixWAS does not require all M sites to have collected data on all q phenotypes. Additionally, mixWAS does not require all individuals to have data available for all phenotypes collected at the site, provided any missing phenotype data at site level is missing completely at random (MCAR). Specifically, let δ_{ijm} be an indicator denoting whether Y_{ijm} is observed. Then the composite likelihood in equation (4) can be modified as

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{m=1}^{M}\sum_{j=1}^{q} L_{jm}\left(\beta_j, \boldsymbol{\gamma_{jm}}\right) = \sum_{m=1}^{M}\sum_{j=1}^{q}\sum_{i=1}^{n_m} \delta_{ijm} \log(\, f_j(Y_{ijm}, X_{im}, \boldsymbol{Z_{im}} |\, \beta_j, \boldsymbol{\gamma_{jm}}))$$

(7)

Note that when a site has no data on a particular phenotype collected, its contribution to the entry of overall score vector corresponding to that phenotype is just 0. We can obtain an overall test of $H_0: \boldsymbol{\beta} = \mathbf{0}$ by

$$T = S^T V^{-1} S = \left(\sum_{m=1}^{M} S_m(0, \hat{\boldsymbol{\gamma}}_{\boldsymbol{m}})\right)^T \left(\sum_{m=1}^{M} V_m\right)^{-1} \left(\sum_{m=1}^{M} S_m(0, \hat{\boldsymbol{\gamma}}_m)\right)$$

(8)

Asymptotically, $T \sim \chi^2_q$, which we can leverage to obtain a corresponding p-value, p_{score}. This test does well for dense alternatives when many of the q phenotypes are non-null ($\beta_j \neq 0$ for many of the $j \in \{1, \dots, q\}$). However, when q is large and most phenotypes are not significantly associated with the SNP (for example only one or two β_j are non-zero), this score test may not be sufficiently powerful as the signal could be diluted by the majority of null effects.

Thus, we also consider a second test that is powerful under sparse alternatives, when the majority of the SNP effects are zero. Specifically, let

$$\boldsymbol{z} = V^{-1/2} S = \left(\sum_{i=1}^{m} V_m\right)^{-1/2} \left(\sum_{i=1}^{m} S(\mathbf{0}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{m}})\right)$$

(9)

$\boldsymbol{z} = (z_1, \dots, z_q) \sim^{\text{iid}} N(0,1)$ and thus p-values $p_1, \dots, p_q$ for the corresponding hypotheses $H_0: \beta_j = 0$ can be obtained $p_j = 2\Phi(-|z_j|)$ where $\Phi(\cdot)$ is the CDF of the standard Normal distribution. In order to combine the p-values, we use the aggregated Cauchy association test (ACAT)[42].

$$t_{ACAT} = \frac{1}{q}\sum_{j=1}^{q} \tan\left(\pi\left[\frac{1}{2} - p_j\right]\right)$$
$$p_{ACAT} = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}(t_{ACAT})$$

(10)

Originally developed as a fast, computationally efficient p-value combination method for rare variant analyses, ACAT was shown particularly powerful in the presence of only a small number of causal variants in a variant set. Via simulations, we will show that such a test boosts power in the case of sparse alternatives.

ACAT has also been shown to be useful as a method for combining p-values from tests powerful in differing scenarios to create an omnibus test[42]. It is particularly appealing as a way to combine our two component p-values, as it does not require one to estimate or account for potentially very complex correlation between component p-value. Thus, in order to create a test robust to both dense and sparse alternatives, we use ACAT to combine p_{score}, which is powerful against dense alternatives, and p_{ACAT}, which boosts power against sparse alternatives, as shown in Equation (11).

$$t_{SNP} = \frac{1}{2}\left(\tan\left(\pi\left[\frac{1}{2} - p_{score}\right]\right) + \tan\left(\pi\left[\frac{1}{2} - p_{ACAT}\right]\right)\right)$$

$$p_{SNP} = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}(t_{SNP})$$

(11)

Figure 1 outlines the full mixWAS algorithm, while Algorithm S1 in the supplementary material provides pseudo-code. mixWAS is lossless in the sense that there is no approximation error brought by the federated decomposition and we obtain identical results as the pooled analysis. It is also communication-efficient because only one round of communication is required across sites. Also, it is highly computationally efficient since the reduced model is shared across all genetic variants and the summary-level statistics have closed-form expressions.

PheWAS

PheWAS methods were utilized as a set of baselines against which to compare mixWAS. For continuous phenotypes, a linear regression was fit for each phenotype, while logistic regressions were fit for each binary phenotype. We first consider site specific estimates $\hat{\beta}_{jm}$ for phenotype j at site m, as sites can not pool their individual data, and combine estimates across sites using inverse-variance weighting to obtain $(\hat{\beta}_1, \ldots, \hat{\beta}_q)$. In contrast to this PheWAS-Meta-analysis estimator, we also consider a PheWAS-Mega-analysis estimator where all individual data are pooled together prior to analysis, and then linear/logistic regression is applied to each phenotype depending on the data type. While this may be unrealistic in practice due to privacy restrictions that may prevent sites from sharing individual level information, it serves as a useful benchmark against which to compare, as there is no approximation error brought by the federated decomposition used in mixWAS' distributed score tests, which obtain identical results as a pooled analysis.

In these two methods, which we refer to as PheWAS-Meta and PheWAS-Mega, respectively, the overall p-value for the SNP is obtained by taking the minimum of the q Bonferonni adjusted p-values.

Simulating MPA across multiple EHRs

Various multi-phenotype association models were simulated to compare mixWAS with existing PheWAS baselines. As mixWAS is designed to integrate summary-level data across sites when individual-level data is unable to be pooled, we generate data at 5 sites. For comparison against mixWAS, we compute three PheWAS methods, as described in the PheWAS section above.

We first outline a general process for generating data in simulations. We begin by drawing six covariates for each subject at each of $M = 5$ sites, denoted by $\boldsymbol{Z_{im}} = 4$ principal components (PCs), age, and gender. Corresponding coefficients for these covariates, $\boldsymbol{\gamma_{jm}}$ that are both site- and phenotype-specific are generated for each phenotype and site. Distributional choices for generating $\boldsymbol{Z_{im}}$ and $\boldsymbol{\gamma_{jm}}$ are shown in Table 1.

Variable	$\boldsymbol{Z_{im}}$ Generation	$\boldsymbol{\gamma_{jm}}$ Generation
Principal Components (4)	$N(0,1)$	$\text{Uniform}(-0.5, 0.5)$
Age (Centered)	$\text{N}(0, 15^2)$	$\text{Uniform}(-0.05, 0.05)$
Gender	$\text{Bernoulli}(0.5)$	$\text{Uniform}(-0.1, 0.1)$

Table 1: Data generation mechanism for individual covariates ($\boldsymbol{Z_{im}}$) and corresponding coefficients ($\boldsymbol{\gamma_{jm}}$). Note that $\boldsymbol{\gamma_{jm}}$ coefficient vectors are generated for each phenotype independently.

SNPs X are drawn from Binomial$(2, \text{MAF})$, where the minor allele frequency (MAF) is a parameter of the specific simulations. SNPs are centered to have mean 0 by subtracting $E[X] = 2 \times \text{MAF}$ so that the size of any SNP effect does not change the prevalence of any binary phenotype. q_c continuous phenotypes are generated from the following linear model

$$Y_{ijm} = \beta_j X_{im} + \boldsymbol{Z_{im}^T \gamma_{jm}} + \epsilon_{ijm} \tag{12}$$

Random noise $\boldsymbol{\epsilon_{im}} = \left(\epsilon_{i1m,} \dots, \epsilon_{iq_cm}\right) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_c})$ is added to each subjects' phenotype vector $\boldsymbol{Y_{im}}$ to induce covariance $\boldsymbol{\Sigma_c}$ across phenotypes. $q_b = q - q_c$ binary phenotypes are generated by the following logistic regression model

$$\log\left(\frac{P\left(Y_{ijm}=1\right)}{1-P\left(Y_{ijm}=1\right)} \,|X_{im}, \boldsymbol{Z_{im}}\right) = \alpha + \beta_j X_{im} + \boldsymbol{Z_{im}^T \gamma_{jm}} \tag{13}$$

where $\alpha = \log\left(\frac{\text{Prevalence}}{1-\text{Prevalence}}\right)$ and the prevalence = $E\left(Y_{ijm}\right)$, which is a simulation parameter, is set to be the same for each binary phenotype. Each subjects q_b binary phenotypes are drawn using the rmvbin() in R's bindata library[43], which creates correlated multivariate binary random variables by thresholding a normal distribution. Marginal probabilities are obtained by applying the inverse logit function to equation (13), and covariance matrix $\boldsymbol{\Sigma_b}$ is used to induce a correlation for the resulting binary phenotypes. Let $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_c} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma_b} \end{pmatrix}$ denote the overall covariance matrix for all phenotypes. In all simulations, 10% of phenotypes are set to be missing completely at random, and additionally we do not require that each site has all data available for all q phenotypes.

We consider $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_q\right)$ of varying direction and sparsity. We define the sparsity of $\boldsymbol{\beta}$ to be the number of components that are non-zero. All simulations set $q = 8$, and we consider sparsities ranging from $\beta = (\beta, \beta, 0,0,0,0,0,0)$ (2 non-null phenotypes) to $\boldsymbol{\beta} = (\beta, \beta, \beta, \beta, \beta, \beta, \beta, \beta)$ (8 non-null phentypes). While we always set the magnitude $\left|\beta_j\right| = \beta$ to be the same for each non-null phenotype, we consider cases where all non-zero components are positive (e.g. $\boldsymbol{\beta} = (\beta, \beta, \beta, \beta, 0,0,0,0)$), all non-zero components are negative (e.g. $\boldsymbol{\beta} = (-\beta, -\beta, -\beta, -\beta, 0,0,0,0)$), and a case where the direction of the effects are opposite, with half non-zero components being positive and half being negative (e.g. $\boldsymbol{\beta} = (\beta, -\beta, \beta, -\beta, 0,0,0,0)$). For each magnitude of effect, β, 2000 simulated datasets were generated, and power for each method was calculated as the percentage of times a method identified significant multi-phenotype associations out of 2000 repetitions.

To demonstrate the utility of mixWAS, we first consider a simulation setting in which q = 8 phenotypes are available at M = 5 sites, with n_m = 1000 subjects at each site. Of the 8 available phenotypes, q_c = 4 are continuous $(\boldsymbol{Y_1}, \dots, \boldsymbol{Y_4})$ and q_b = 4 are binary $(\boldsymbol{Y_5}, \dots \boldsymbol{Y_8})$. A common variant setting was chosen for binary phenotypes with MAF = 20% and the prevalence of each phenotype = 30%. Effect sizes $|\beta|$ considered ranged from [0.01, 0.35].

For this setting, binary phenotypes are always considered to be positive, indicating a SNP increases the likelihood of each phenotype (disease). We consider cases where 2 (1 binary + 1 continuous), 5 (3 binary + 2 continuous), and all 8 β_j are non-zero. While binary phenotypes are always positive, we consider cases where the continuous phenotypes either all positive (Figure 2) or are in opposite directions (e.g. 1 negative, 1 negative/1 positive, and 2 negative/2 positive in the 3 levels of sparsity, respectively) [Figure 3].

We consider 3 types of residual correlation between phenotypes: positive, no correlation, and negative. Correlations between phenotypes are shown in Figure 9 When all effects are positive, residual correlation is added to both continuous and binary phenotypes. In the case of opposite direction continuous phenotypes, we only include residual correlation for continuous phenotypes. As this setting only considers positive binary phenotype, we wanted to understand the interaction between oppositive direction effects and different directions of correlations without power varying due to the direction of residual correlation of the positive binary effects, which can heavily influence power (Figure 2). Note that $\boldsymbol{\Sigma_c}$, the covariance matrix for the continuous phenotypes is simply the upper block diagonal of the correlation matrix shown in Figure 10 scaled by $\sigma^2 = 2.3^2$.

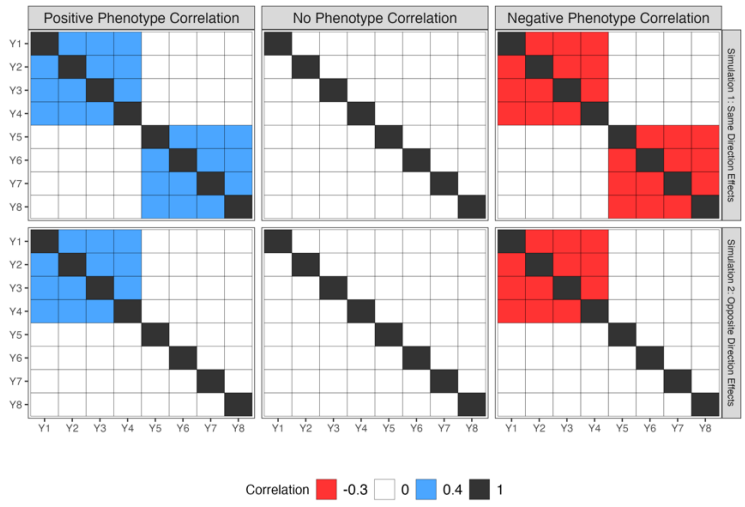


Figure 9: Correlation matrix for mixed data type simulation. $(\boldsymbol{Y_1}, \dots, \boldsymbol{Y_4})$ are continuous phenotypes while $(\boldsymbol{Y_5}, \dots \boldsymbol{Y_8})$ are binary phenotypes. To get the full covariance matrix for the continuous phenotypes, $\boldsymbol{\Sigma_c}$, the upper block diagonal is scaled by $\sigma^2 = 2.3^2$. For binary phenotypes, the correlation matrix is supplied to R function rmvbin() in the bindata package.

mixWAS was compared against the PheWAS-Mega/meta-analysis methods described in the PheWAS section, as well as an oracle score test, which is a score test using only de-correlated z −scores (Equation (9)) for non-null phenotypes. This test is considered an oracle test because in which phenotypes have non-null associations with the SNP is unknown in practice. As such, this reference gives a helpful upper bound on the power of score-based hypothesis tests under this setting.

Power curves from the simulation are shown in Figures 2 and 3. Most notably, mixWAS has higher power compared to standard PheWAS methods in all scenarios, especially in cases where only 2 of the 8 SNP effects are non-null, and when the SNP effects oppose the direction of residual correlation between phenotypes.

Utilizing mixWAS to detect MPA using eMERGE and UK Biobank

From the eMERGE study, patients from multiple adult electronic health records (EHRs), including Marshfield Clinic, Vanderbilt University, Kaiser Permanente/University of Washington, Mayo Clinic, Northwestern University, Geisinger, and Mass General Brigham, were included in the research. Binary disease statuses for individuals were determined based on specific ICD-9 codes, including unspecified essential hypertension (ICD-9 401.9), type 2 diabetes (ICD-9 250.00), hyperlipidemia (ICD-9 272.4), benign essential hypertension (ICD-9 401.1), atrial fibrillation (ICD-9 427.31), congestive heart failure (ICD-9 428.0), and coronary atherosclerosis (ICD-9 414.00). Additionally, median laboratory measures, including LDL, HDL, serum total cholesterol, triglycerides, and BMI for each patient, were calculated and utilized as continuous outcomes.

mixWAS was applied to each SNP to detect MPA among the mixed binary and continuous outcomes. This analysis included adjustments for age, sex, and the top 10 principal components to account for population stratification. Given the different ages of disease onset for each condition, distinct disease-associated ages were incorporated into the mixWAS model. The disease-associated age was computed as the median age for each continuous laboratory measure. For binary diseases, the median age of the ICD-9 code assignments for a disease was employed as the age for cases, while for controls, the age was determined as the patients' age at their last EHR record.

The MPA identified by eMERGE were independently validated using data from the UK Biobank (UKBB). Since the UKBB primarily utilizes ICD-10 codes for clinical diagnosis, a mapping process was carried out to convert the ICD-9 codes used in eMERGE data to their corresponding ICD-10 codes. The converted ICD-10 codes were unspecified essential hypertension and benign essential hypertension (ICD-10 I10), type 2 diabetes (ICD-10 E119), hyperlipidemia (ICD-10 E784 and ICD-10 E785), atrial fibrillation (ICD-10 I489), congestive heart failure (ICD-10 I509), and coronary atherosclerosis (ICD-10 I251). The continuous laboratory measures were extracted from the following fields, including LDL (field 30780), HDL (field 30760), total cholesterol (field 30690), triglycerides (field 30870), and BMI (field 12001). Notably, unspecified essential hypertension (ICD-9 401.9) and benign essential hypertension (ICD-9 401.1) from eMERGE were consolidated into a single condition, essential (primary) hypertension (ICD-10 I10), in the UKBB dataset.

In the eMERGE discovery analysis, the significance threshold for SNPs' p-value was set as 8.19×10^{-9} (0.05/6,106,952), corresponding to the Bonferroni adjusted p-value threshold. Subsequently, the 4,534 significant SNPs identified were re-evaluated in the UKBB dataset using the mixWAS algorithm, with a significance threshold set at $0.05/4{,}534 = 1.103 \times 10^{-5}$. Furthermore, these 4,534 significant SNPs underwent PheWAS analysis in the UKBB to identify specific SNP-phenotype associations driving the MPAs. The Bonferroni-adjusted p-value threshold for this analysis was set at $(0.05/4{,}534)/10 = 1.103 \times 10^{\wedge-6}$. In contrast, the standard PheWAS Bonferroni-corrected p-value threshold is $7.90 \times 10^{\wedge-10}$, which accounts for analyzing all SNPs in the UKBB dataset.

Functional annotation of the mixWAS-identified SNPs was carried out using the FUMA software[44,45], and these SNPs were annotated using canonical pathways from the Human Molecular Signatures Database (MSigDB)[46].

Discussion

Recently, numerous initiatives have made EHR-linked genetic data available for genomics research. The wealth of extensive and well-characterized patients' phenotype data extracted from EHRs presents an unprecedented opportunity to investigate genetic variants associated with multiple diseases or traits, which can potentially shed light on shared underlying genetic architectures among diverse phenotypes. The availability of EHR-linked genetic data from various institutions, health systems, and population studies further enables us to exchange and integrate data from different sources, thereby enhancing the power and reproducibility of our research findings.

However, despite these opportunities, significant data sharing constraints and methodological challenges have hindered the full utilization of multiple EHR-linked genetic datasets for studying MPA. To overcome these challenges, we developed the mixWAS method, which efficiently and losslessly integrates summary statistics from multiple data sources to effectively identify genetic variants associated with multiple binary or continuous phenotypes, enabling a more comprehensive exploration of the shared genetic basis underlying various phenotypic traits.

We used simulation studies to demonstrate that mixWAS outperforms standard statistical approaches used in most PheWAS across a range of realistic settings that incorporate heterogeneity across sites, ranging direction, magnitude, and sparsity of phenotype effects, missing data, healthy volunteer biobanks, and common/rare genetic variants. By accounting for correlation between phenotypes in a manner that does not require individual level information, mixWAS gained the most power in settings where residual correlation existed between phenotypes, and SNP effects went against the correlation of these effects. Given that MPAs can often be difficult to detect, in large part due weak associations and multiple testing penalties in standard PheWAS methods, and given its improved power in most settings, mixWAS is a superior method for studying the shared genetic basis underlying multiple phenotypic traits in complex multi-EHR settings.

Towards this end, we employed the mixWAS method to study MPA across blood lipid levels, BMI, and diseases of the circulatory system using seven EHR sites from the eMERGE project, and we validated our findings using data from the UKBB. Figure 3 illustrates heterogeneities in data characteristics across different eMERGE study sites. Notably, Vanderbilt and Mass General Brigham had the largest relative sample sizes compared to other sites, but both datasets had significant missing blood lipid measurements. The presence of differential missing data patterns is expected when integrating data from multiple real EHR sources, given the varying clinical protocols and patient populations among hospitals. Nevertheless, the mixWAS method can effectively account for the differential missing data across hospitals.

Applying mixWAS separately to each eMERGE site or across all sites yielded significantly different numbers of significant genetic associations. Comparing results between individual sites revealed a strong correlation between sample size and the number of detected genetic associations. Notably, the locations of the significant associations remained consistent between different datasets, suggesting the detection of the same MPAs across different EHRs, with only variations in the number of associations. The integrated eMERGE analysis identified the highest number of significant associations compared to any individual site (Figure 4). Importantly, the integrated analysis identified additional genetic associations that are not present in any single-site data, underscoring the benefits of this integrated approach.

The 4,534 mixWAS-identified MPA in eMERGE were further validated in the UKBB data. Using the p-value thresholds corresponding to the number of MPA, 4,428 MPA reached the significance threshold in UKBB (Figure 5). Given the distinct study populations and data generation processes between the two datasets (US and UK), we believe the 4,428 genetic variants represent robust MPAs for the studied diseases and traits. A common challenge in interpreting MPAs lies in distinguishing SNPs that are associated with only one phenotype from those associated with multiple phenotypes. However, a joint test, such as mixWAS, can effectively detect both types of associations equivalently. To further investigate the specific trait-SNP associations driving the MPAs, we performed additional single phenotype and SNP associations for all MPA SNPs identified in eMERGE. MPA SNPs were found to be significantly associated with 0 to 8 traits, with 2 to 4 traits being the most common, and the majority of MPA SNPs were associated with more than 2 phenotypes (Figure 6a).

Among the traits, lipid levels (including LDL, HDL, Cholesterol, and Triglycerides) shared the largest number of associated genetic variants, followed by BMI (Figure 6b). Additionally, coronary artery disease and T2D showed common MPAs with protein lipid levels and BMI. For heart failure, no significant associations were detected; however, some of the genetic associations were just below the significance threshold. The evaluated genetic variants are specifically those that showed MPAs across diseases, and the lack of identified associations may be due to limited number of heart failure cases, or shared genetic effects between heart failure and other diseases, or it may indicate inadequate power to detect smaller associations, or phenotype heterogeneity, or a combination of multiple factors. We additionally performed a separate GWAS analysis on heart failure alone in eMERGE and no SNP associations were found to be significant. This supports our hypothesis that the data was underpowered for studying heart failure. However, further studies are needed to confirm these results, as this study represents the first identification of MPAs in these diseases and traits.

Moreover, we observed improved power in detecting specific trait-SNP associations from the 4,534 mixWAS-detected MPAs in UKBB. Compared to investigating all trait-SNP associations, or a PheWAS analysis, using mixWAS MPAs resulted in an 18.9% increase in the number of detected associations (Figure 7). This increased number of associations can provide additional insights into the shared underlying genetics among different diseases and traits. Functional analysis of the mixWAS-detected MPA confirmed that the MPA are enriched for pathways related to cholesterol metabolism, lipoprotein function, hyperlipidemia, as well as pathways associated with LDL, HDL, and triglycerides (Figure 8). Together, these findings support that the mixWAS has improved power to detect more MPA that are functionally relevant to the studied diseases/traits.

In our evaluation of mixWAS, we tested it against a fixed number of phenotypes and found that it outperformed typical PheWAS statistical approaches. Importantly, this approach could be scaled to an arbitrarily large number of phenotypes, including different approaches taken to defining the phenome. mixWAS also extends its utility beyond genetic datasets and can be applied to any datasets that contain binary or continuous outcomes. Nevertheless, we also recognize several limitations of the study. First, while mixWAS can accommodate differential missing data patterns in each dataset, this relies on the assumption that the data is missing at random. Second, the presence of a substantial number of null phenotypes—those unrelated to genetic variants—can diminish the power of mixWAS. Lastly, mixWAS requires the sharing of more extensive summary-level statistics compared to traditional meta-analysis methods, resulting in higher data communication costs. However, these costs remain orders of magnitude lower than the transmission of entire datasets.

Code availability

The mixWAS algorithm and the code associated with this study have been deposited at:

https://github.com/lbenz730/mixWAS

Acknowledgment

NIH R01 LM010098, AG066833, GM148494, LM014344, LM012607, LM013519, AI130460, AG073435, RF1AG077820, R56AG069880, R56AG074604, U01TR003709, R21AI167418 and R21EY034179. MDR was funded by R01HG010067 and R01HL169458.

eMERGE Network (Phase III). This phase of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG8657 (Group Health Cooperative/University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine).

UK Biobank. All data for this cohort pertained to project 32133 – "Integration of multi-organ imaging phenotypes, clinical phenotypes, and genomic data".

References

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362–7 (2009).
2. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).
3. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *American Journal of Human Genetics* vol. 110 179–194 Preprint at https://doi.org/10.1016/j.ajhg.2022.12.011 (2023).
4. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009).
5. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
6. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications* vol. 11 1–3 Preprint at https://doi.org/10.1038/s41467-020-19653-5 (2020).
7. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, (2014).
8. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51**, 1339–1348 (2019).

9. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* vol. 169 1177–1186 Preprint at https://doi.org/10.1016/j.cell.2017.05.038 (2017).
10. Gratten, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: Implications for genomic medicine. *Genome Medicine* vol. 8 78 Preprint at https://doi.org/10.1186/s13073-016-0332-x (2016).
11. Chun, S. *et al.* Leveraging pleiotropy to discover and interpret GWAS results for sleep-associated traits. *PLoS Genet* **18**, e1010557 (2022).
12. Zhang, X. *et al.* Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and nervous system disorders. *Nat Commun* **13**, (2022).
13. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics* vol. 14 483–495 Preprint at https://doi.org/10.1038/nrg3461 (2013).
14. Li, R. *et al.* Lossless integration of multiple electronic health records for identifying pleiotropy using summary statistics. *Nat Commun* **12**, 1–10 (2021).
15. Li, R. *et al.* A regression framework to uncover pleiotropy in large-scale electronic health record data. *Journal of the American Medical Informatics Association* **26**, 1083–1090 (2019).
16. Tyler, A. L., Crawford, D. C. & Pendergrass, S. A. The detection and characterization of pleiotropy: discovery, progress, and promise. *Brief Bioinform* **17**, 13–22 (2016).
17. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
18. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
19. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2–S8 (2017).
20. Fritsche, L. G. *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am J Hum Genet* **102**, 1048–1061 (2018).
21. Kho, A. N. *et al.* Electronic medical records for genetic research: Results of the eMERGE consortium. *Sci Transl Med* **3**, (2011).
22. McCarty, C. A. *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* **4**, 13 (2011).
23. Verma, A. *et al.* PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *The American Journal of Human Genetics* **102**, 592–608 (2018).
24. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
25. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biology* vol. 7 Preprint at https://doi.org/10.1098/rsob.170125 (2017).
26. Johnson, E. C. *et al.* A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry* **7**, 1032–1045 (2020).

27. Verma, A. *et al.* A Phenome-Wide Association Study of genes associated with COVID-19 severity reveals shared genetics with complex diseases in the Million Veteran Program. *PLoS Genet* **18**, e1010113 (2022).
28. Diogo, D. *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nat Commun* **9**, (2018).
29. Deflaux, N. *et al.* Demonstrating paths for unlocking the value of cloud genomics through cross cohort analysis. *Nat Commun* **14**, (2023).
30. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410–421 (2019).
31. Rankinen, T., Sarzynski, M. A., Ghosh, S. & Bouchard, C. Are there genetic paths common to obesity, cardiovascular disease outcomes, and cardiovascular risk factors? *Circ Res* **116**, 909–922 (2015).
32. Xu, L., Borges, M. C., Hemani, G. & Lawlor, D. A. The role of glycaemic and lipid risk factors in mediating the effect of BMI on coronary heart disease: a two-step, two-sample Mendelian randomisation study. *Diabetologia* **60**, 2210 (2017).
33. Edwards, K. L., Mahaney, M. C., Motulsky, A. G. & Austin, M. A. Pleiotropic genetic effects on LDL size, plasma triglyceride, and HDL cholesterol in families. *Arterioscler Thromb Vasc Biol* **19**, 2456–2464 (1999).
34. Thomas, D. G., Wei, Y. & Tall, A. R. Lipid and metabolic syndrome traits in coronary artery disease: A Mendelian randomization study. *J Lipid Res* **62**, 100044 (2021).
35. Stanaway, I. B. *et al.* The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet Epidemiol* **43**, 63–81 (2019).
36. Stanaway, I. B. *et al.* The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet Epidemiol* **43**, 63–81 (2019).
37. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Front Genet* **5**, 370 (2014).
38. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
39. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**, (2015).
40. Reid, N., Varin, C. & Firth, D. *An Overview of Composite Likelihood Methods*. *Statistica Sinica* vol. 21 https://www.researchgate.net/publication/228634405 (2011).
41. Lindsay, B. G. Composite likelihood methods. *Comtemporary Mathematics* **80**, 221–239 (1988).
42. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410–421 (2019).
43. Generation of Artificial Binary Data [R package bindata version 0.9-20]. (2021).
44. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat Commun* **10**, 1–13 (2019).
45. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1–11 (2017).

46. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
47. Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am J Hum Genet* **90**, 821–835 (2012).
48. Firth, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80**, 27 (1993).

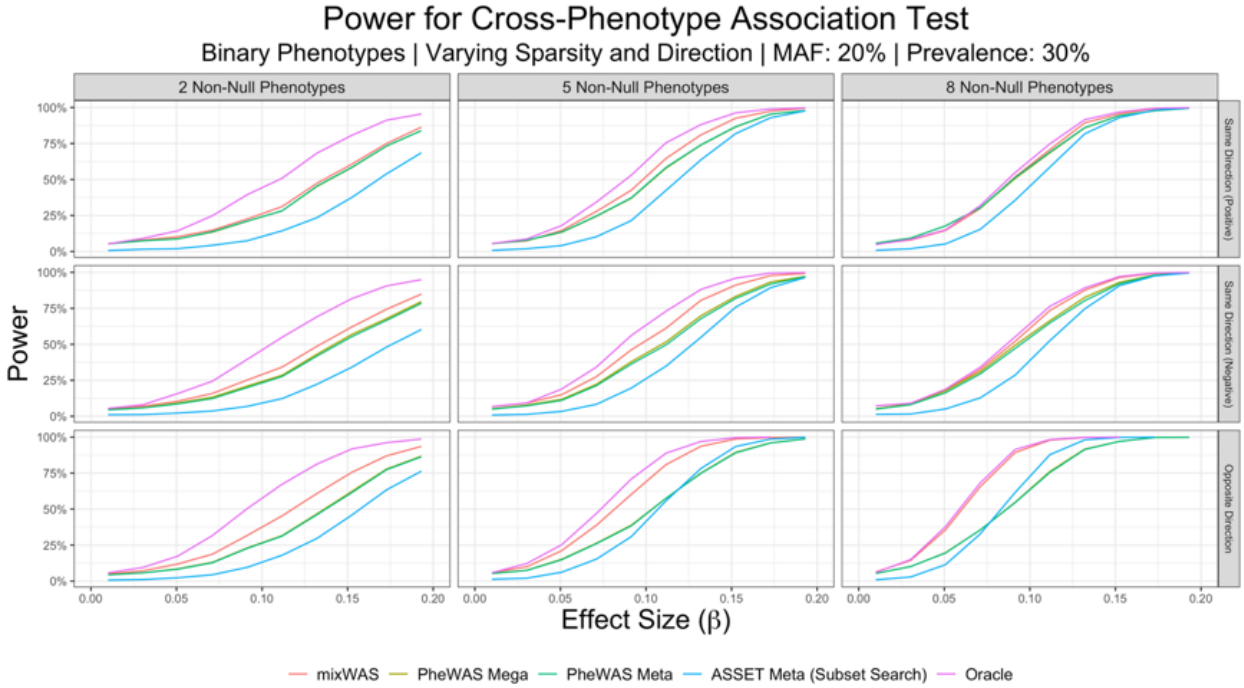
Supplementary Material

Binary Phenotypes: Common and Rare Variant Settings

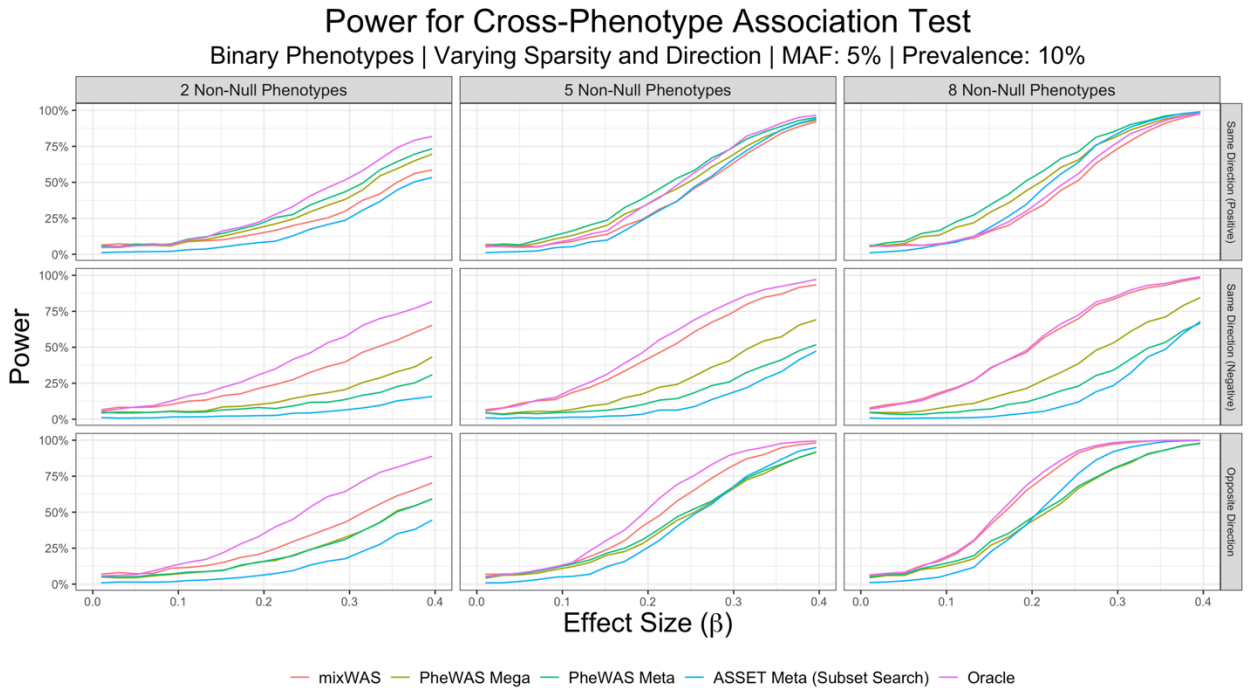
In the additional simulation study, we also consider a method that is more common in pleiotropy analysis, ASSET[47], which exhaustively searches subsets of the phenotypes for significant effects, and is able to account for correlation between the phenotypes introduced by sample overlap. [43]. Since ASSET is only for pleiotropic analyses of binary phenotypes, we consider various simulation settings in which all 8 phenotypes $(\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_8})$ are binary. As outlined above, $q = q_b = 8, n_m = 1000$ subjects at each of $M = 5$ sites. We consider a common variant setting with MAF = 20% and prevalence = 30%, along with a rare variant setting with MAF = 5% and prevalence = 10%. Effect sizes $|\beta|$ considered ranged from [0.01, 0.60]. $\boldsymbol{\Sigma} = \boldsymbol{\Sigma_b}$ is shown in Figure 3a. Only positive correlation is considered in this setting.

Empirical power curves for this pair of binary phenotype simulations are shown in Figure S1. In the common variant/high prevalence setting, mixWAS outperforms all PheWAS methods, including ASSET, which is overly conservative in most settings, as it only accounts for correlation due to sample overlap, rather than correlation induced by the fact that outcomes themselves may be highly correlated across subjects. In the rare variant/low prevalence setting, PheWAS methods are highly impacted by the direction of non-null effects. When effects are all positive (i.e. SNP increases disease prevalence), PheWAS-Meta and mega outperform even the oracle score test.

However, when all SNP effects are negative (SNP decreases disease prevalence) PheWAS-Meta and mega do significantly worse than mixWAS, due to the fact that the direction of these effects is against the direction of the residual correlation (Figure S3a). Similar results can be seen when effects are in opposite directions. The sign of the effect likely impacts PheWAS methods in low prevalence settings due to the fact that logistic regression is biased in problems with heavy class imbalance[48]



a: Common Variant/High Prevalence Setting



b: Rare Variant/Low Prevalence Setting

Figure S1: Empirical power curves for binary phenotype simulations. In the common variant/high prevalence setting, mixWAS outperforms all PheWAS methods, including ASSET, which is overly conservative in most settings. In the rare variant/low prevalence setting, PheWAS methods are highly impacted by the effect direction(s), doing better than mixWAS in the case of all positive effects but performing significantly worse when all effects are negative, or effects are in opposite directions.

Healthy Controls

Genetic biobanks, like the UK Biobank, often only contain controls with none of the diseases/phenotypes of interest, and cases for each specific subject. A subject may be a case for multiple phenotypes, but the absence of a subject being a case for a particular phenotype does not make them a control for that phenotype–their disease status is simply unknown, and no subject can ever be both a case one phenotype and a control for a different phenotype.

To mimic the structure of these real-life genetic databases, we designed a simulation with healthy controls. 8 binary phenotypes were generated in the same manner as the common variant, high prevalence binary phenotypes above, except that controls were designated as subjects for whom Y_{ijm}= 0 for $j \in \{1, \dots, 8\}$. Subjects who were a case for at least one phenotype had their control status for other phenotypes replaced by an *NA* indicating disease status unknown. A stronger correlation structure $\boldsymbol{\Sigma} = \boldsymbol{\Sigma_b}$ was utilized in this simulation, as shown in Figure S3b. Note that due to the perfect split between cases and controls, an even stronger correlation is induced between phenotypes than what is shown in Figure 3b. Additionally, since subjects who were cases for some subset of phenotypes could be controls for other phenotypes in previous simulations, but under the healthy control setting cannot be controls for other phenotypes, the effective number of subjects is drastically reduced compared to other simulation settings.

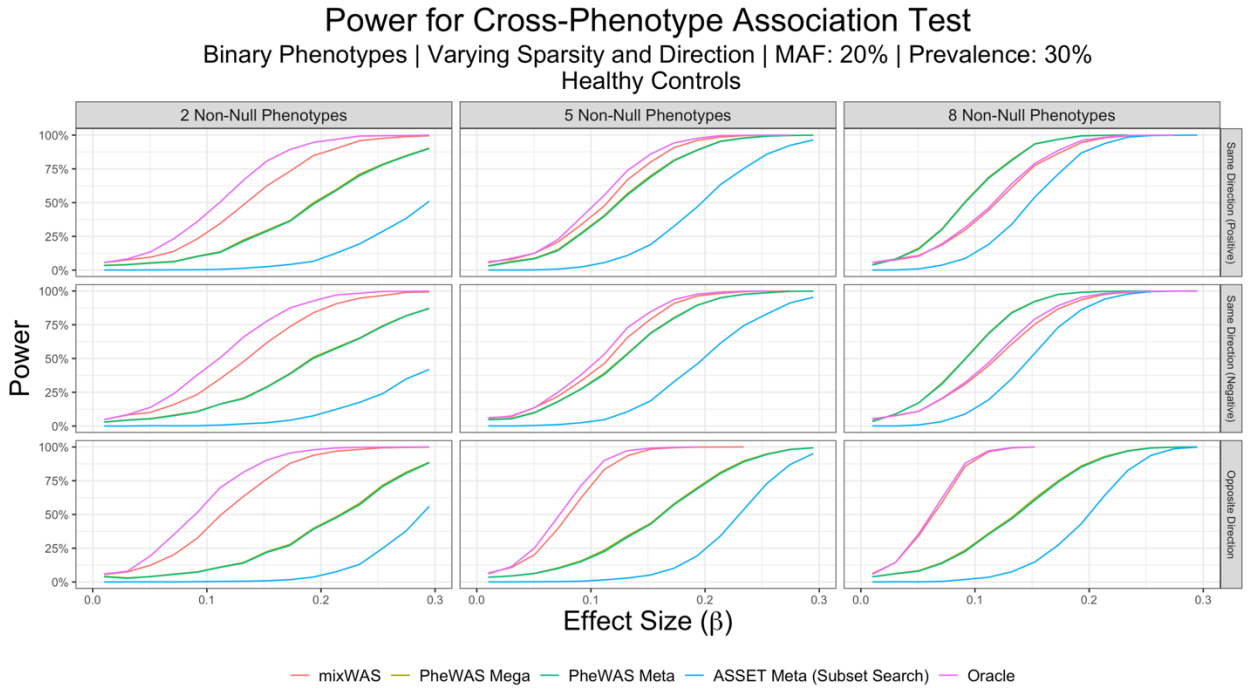
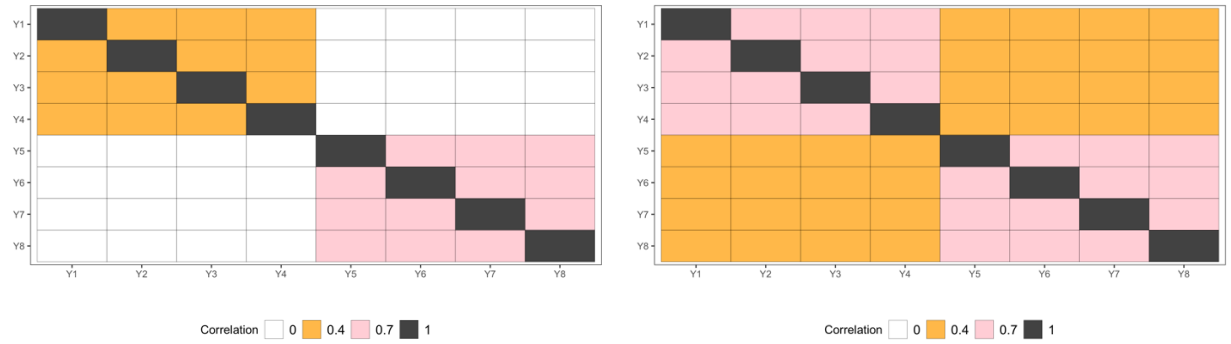


Figure S2: Empirical power curves comparing various cross-phenotype association tests for simulated binary phenotypes using only healthy controls to compare against diseased cases. mixWAS significantly outperforms ASSET in every setting, as the latter method does not account for correlation between phenotypes except for correlation induced by sample overlap. PheWAS-Mega/meta obtain the highest power when no-sparsity is present, but power is again affected by the direction of effects.

Empirical power curves are shown in Figure S2. mixWAS significantly outperforms ASSET in every setting, as the latter method does not account for correlation between phenotypes except for correlation induced by sample overlap. PheWAS-Mega/meta obtain the highest power when no sparsity is present, but power is again affected by the direction of effects, with power being worst under mixed-direction phenotypes. This result is likely explained once again by the smaller effective sample size introduced by limiting our previous simulated sample to include only health controls as well as the bias of logistic regression in class imbalanced problems[48].



a: Common and Rare Variant Settings

b: Health Controls Setting

Figure S3: Correlation matrices for binary phenotype data type, supplied to R function rmvbin() in the bindata package[43].

Algorithm 1: mixWAS

Data: SNP $\boldsymbol{X}_m$, Phenotypes $(\boldsymbol{Y}_{1m}, ..., \boldsymbol{Y}_{qm})$, Covariates $\boldsymbol{Z}_m$ at each site $m \in \{1, ..., M\}$
Result: p-value for SNP testing $H_0 : \boldsymbol{\beta} = 0$

for *Site m from 1 to M* **do**
  for *Phenotype j from 1 to q* **do**
    Fit the reduced version of the model in equation 1, assuming $\beta_j = 0$, to estimate $\hat{\boldsymbol{\gamma}}_{jm}$
    Compute the j^{th} component of the score vector $S_m(\mathbf{0}, \hat{\boldsymbol{\gamma}}_m)$ based on the corresponding phenotype data type. Note that if Phenotype j is not available at site m, the score contribution is 0.
  end
  Compute V_m, the score variance matrix for all phenotypes at the site.
end
Send all summary statistics (site-specific score/variance matrices, $S_m(\mathbf{0}, \hat{\boldsymbol{\gamma}}_m), V_m$ to a central server.
Compute test statistic for distributed score test in Equation (8) and corresponding p-value p_{score} from a χ^2_q distribution.
Compute z-scores via Equation (9) and corresponding p-values $p_1, ..., p_q$ from $N(0, 1)$.
Apply aggregated Cauchy association test (ACAT, Equation (10)) to find p_{ACAT}.
Using the ACAT method, combine p_{score} and p_{ACAT} to get a final SNP p-value, as in Equation (11).

Algorithm S1: Outline of mixWAS algorithm