A One-Shot Lossless Algorithm for Cross-Cohort Learning in Mixed-Outcomes Analysis

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], Thomas Lumley[11], Marylyn D. Ritchie[12], Jason H. Moore[1], Yong Chen[13]

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 Biostatistics, University of Auckland
12. Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania
13. 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

In cross-cohort studies, integrating diverse datasets, such as electronic health records (EHRs), is both essential and challenging due to cohort-specific variations, distributed data storage, and data privacy concerns. Traditional methods often require data pooling or complex data harmonization, which can reduce efficiency and limit the scope of cross-cohort learning. We introduce mixWAS, a one-shot, lossless algorithm that efficiently integrates distributed EHR datasets via summary statistics. Unlike existing approaches, mixWAS preserves cohort-specific covariate associations and supports simultaneous mixed-outcome analyses. Simulations demonstrate that mixWAS outperforms conventional methods in accuracy and efficiency across various scenarios. Applied to EHR data from seven cohorts in the US, mixWAS identified 4,534 significant cross-cohort genetic associations among traits such as blood lipids, BMI, and circulatory diseases. Validation with an independent UK EHR dataset confirmed 97.7% of these associations, underscoring the algorithm's robustness. By enabling lossless cross-cohort integration, mixWAS improves the precision of multi-outcome analyses and expands the potential for actionable insights in healthcare research.

Introduction

Cross-cohort studies increasingly require robust methods to integrate heterogeneous datasets, especially in contexts where multiple outcomes—such as phenotypic traits, clinical measures, and other health indicators—are examined concurrently. 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), provides a rich source for uncovering complex multi-outcome associations. For example, multi-phenotype associations (MPA), where a single genetic variant is linked to multiple traits, reveal potential shared genes and pathways across diseases. Since patients' genetic data can be mapped to a wide range of clinically relevant phenotypes, these datasets enhance opportunities for comprehensive multi-outcome analysis[1–9].

Genome-wide association studies (GWAS) have systematically identified numerous genetic associations for various diseases and traits[10–12],yet challenges persist in uncovering shared genetic and non-genetic architectures across multiple outcomes[13–15]. Many observed multi-outcome associations [16–18], such as MPA, reveal complex genetic relationships across traits and may provide insights into disease pathogenesis[17,19–22]. Moreover, identifying MPA holds significant clinical values across various domains, including the development of targeted therapies against genes or pathways involved in the development of multiple diseases[23], repurposing existing drugs targeting shared druggable genes across diseases[24], and improving disease risk prediction and screening by incorporating information about share genetic factors across diseases[25,26]. Thus, identifying MPA is a crucial next step towards improved understanding of the genetic architectures of complex diseases.

The integration of multiple EHRs for cross-cohort analysis improves the power to detect MPA and increases the reproducibility of findings. 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[27,28]. Alternative methods have been developed for more efficient statistical tests but have been generally limited to continuous phenotypes or the analysis of individual datasets[22,29]. 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[22,30–33]. However, meta-analyzed PheWAS has been suggested to be less powerful than the pooled analysis when detecting MPA across multiple EHRs[1,33,34].

To address these issues, we introduce *mixWAS*, a one-shot lossless algorithm designed for versatile cross-cohort integration, particularly suited to identifying associations across diverse, mixed-outcome datasets. 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, such as site-specific covariate associations, 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, including site-specific covariate associations commonly 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).

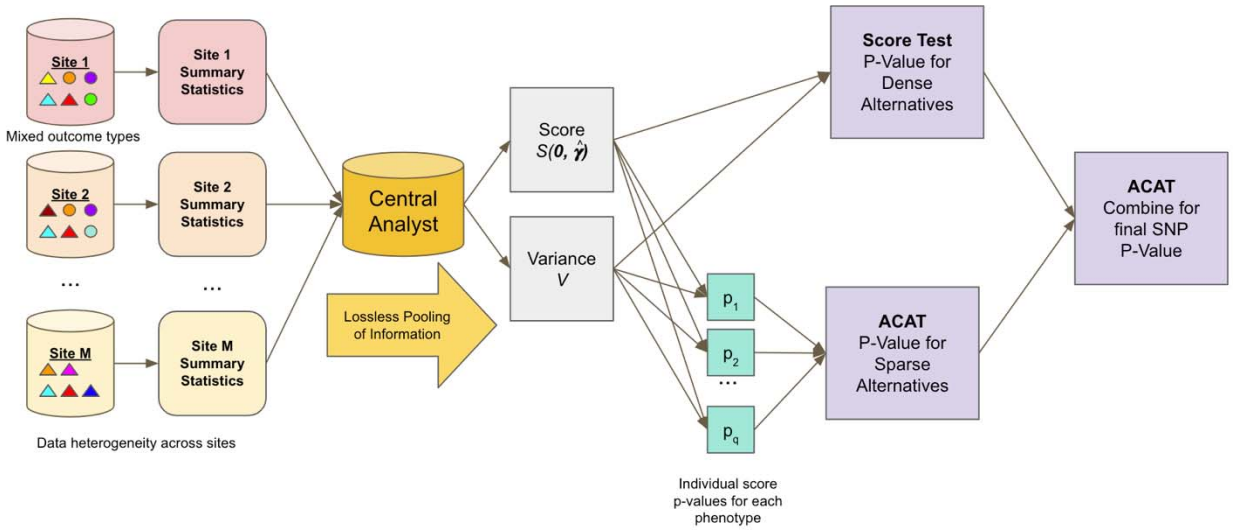

Figure 1: Outline of the mixWAS algorithm. Each site transmits summary level statistics of heterogeneous, mixed-typed phenotypes, specifically the score vector and variance matrix , to a central analyst. The central analyst pools the site-specific score and variance contributions in a lossless manner for use in a score test, which is powerful against dense phenotypes, and a second test, robust to sparse phenotypes, which combines the -values of individual score tests for each of phenotypes using the ACAT method[35]. Finally, these two -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 and PheWAS, 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, where the beta coefficients were combined using meta-analysis. 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 Figure S1). 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 S1). 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).

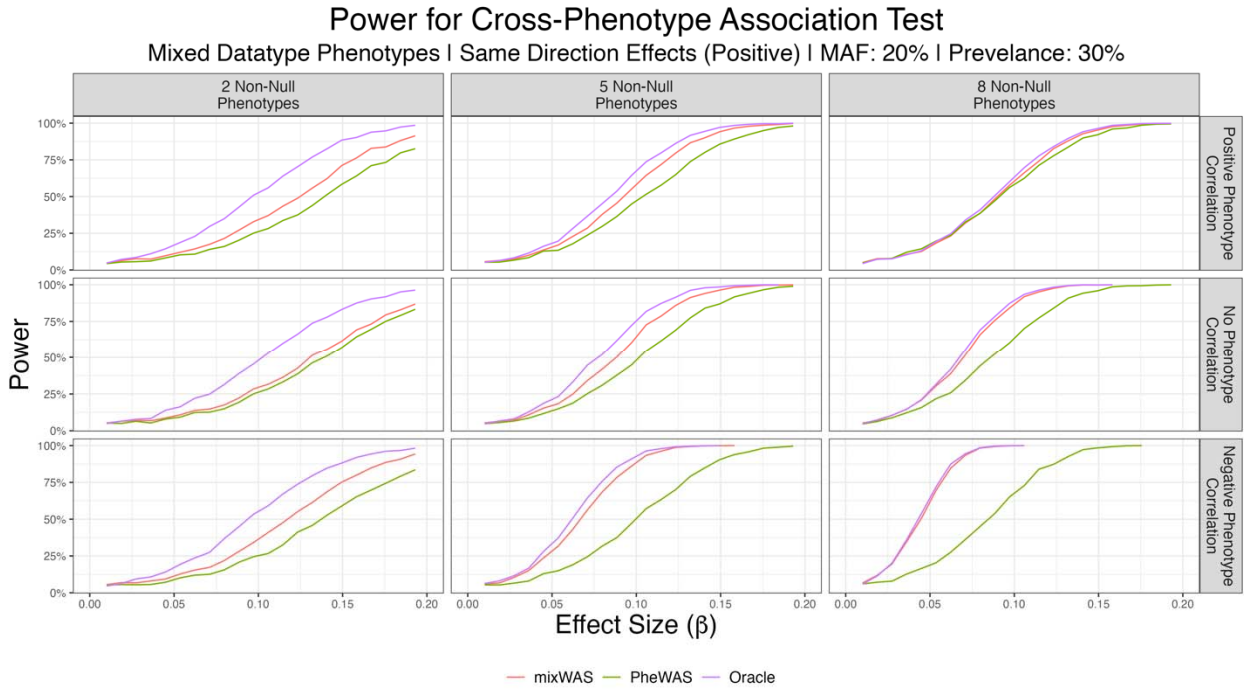

Figure 2: Empirical power curves comparing various cross-phenotype association tests for simulated mixed-type phenotypes. The simulated SNP is positively associated with all phenotypes, while the MPA sparsity (e.g. number of phenotypes with significant association) and correlation between phenotypes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

Detecting MPA across blood lipids levels, BMI, and diseases of circulatory system. Previous research has identified MPA between BMI and coronary heart disease [36,37], blood lipids levels[38], and low-density lipoprotein, triglycerides, and cardiovascular diseases[39]. 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 S2, illustrating heterogeneities among the EHRs, including variations in the number of patients and patterns of missing data.

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 3).

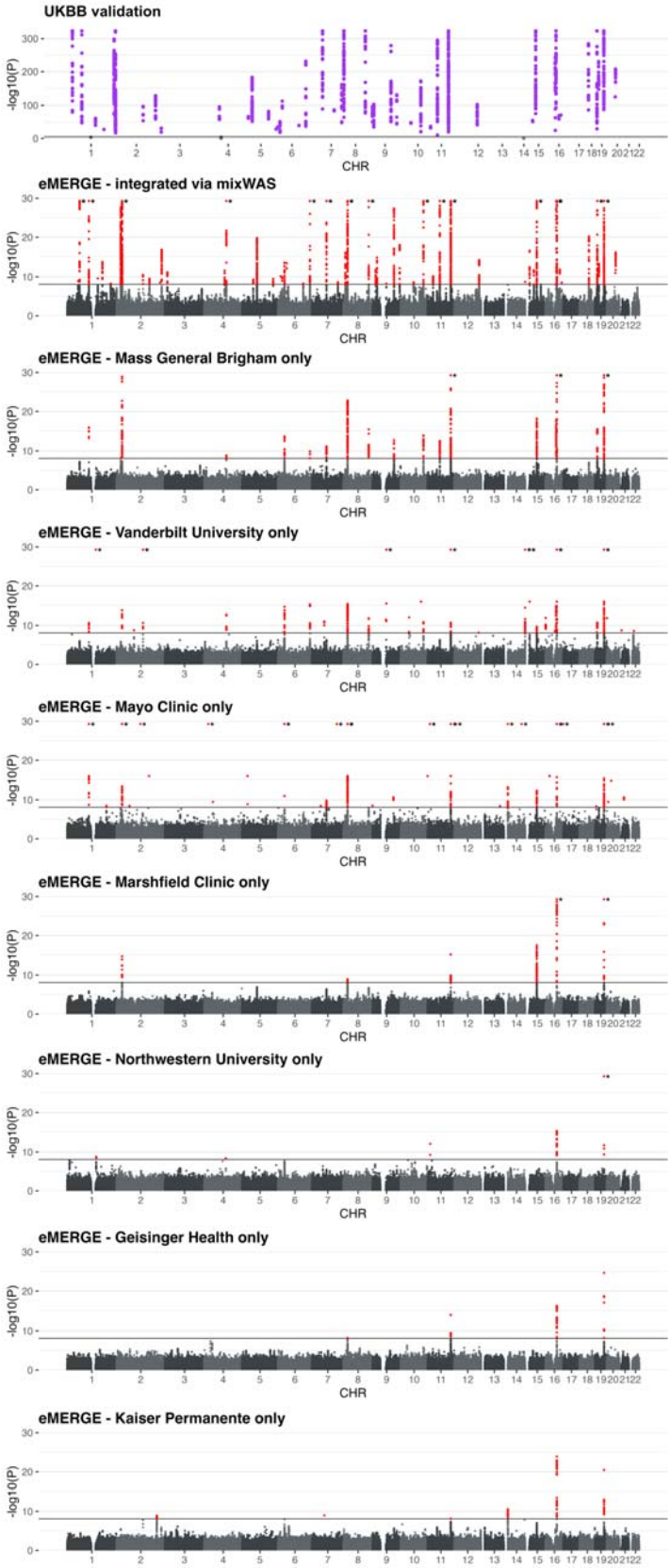

Figure 3. Manhattan plots of MPA in eMERGE single site analysis, eMERGE integrated analysis, and UKBB validation. The association between each SNP and all phenotypes was tested using mixWAS. The resulting p-value of each SNP was -log10(p) transformed and plotted along the y-axis. SNPs with p-values lower 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. 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 4a).

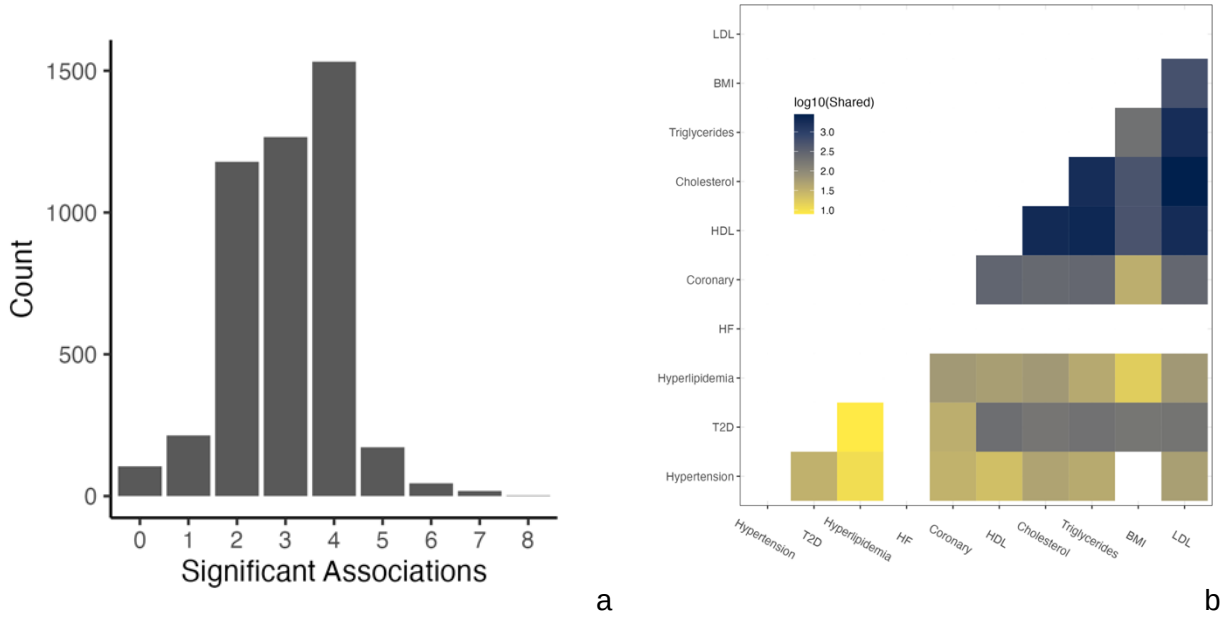

Figure 4. a) The distribution of the number of traits associated with a SNP. Count (y-axis) reflects the number of SNPs with significant associations for varying numbers of phenotypes (x-axis). Each association is determined using regression analysis, while adjusting for the Bonferroni-corrected p-value 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 log10 scale. 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 4b).

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 S3). 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.

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 5. These findings provide additional support for the genetic associations identified through mixWAS.

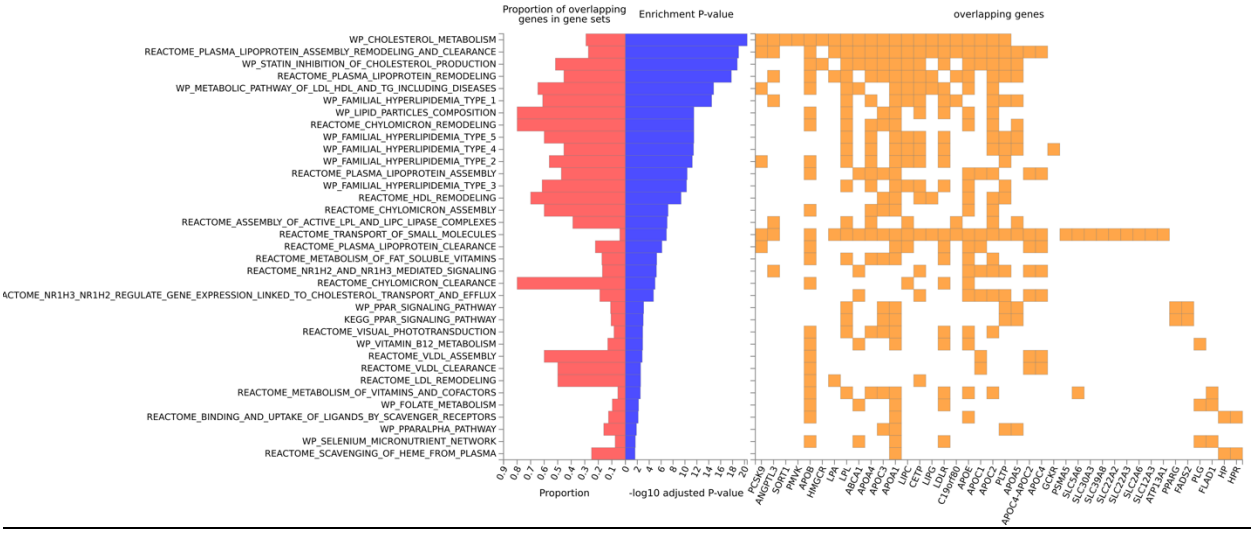

Figure 5. Pathway annotations of mixWAS identified SNPs. The mixWAS identified SNPs were annotated using the MSigDB canonical pathways. The enrichment of a pathway is determined by the number of SNPs associated with the pathway. The pathways are ranked by their enrichment -log10(p-values) from top to 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

The proposed mixWAS Algorithm

mixWAS is designed to jointly test the association between a single SNP and multiple mixed-type phenotypes , across sites, which cannot share individual level patient data across sites due to privacy concerns. We let index individuals, index phenotypes, and index sites. For the individual at the site we denote the SNP by and the phenotypes by . The outcomes of interest may be of differing data types such as binary, continuous, count, or time to event.

Let denote a vector of an individual's covariates, such as age, gender, or ancestry principal component (PCs). As a general framework for phenotype j,

(1)

where denotes the SNP effect, shared across all sites, and 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

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

(2)

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

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

(3)

Following the approach of Li et al[1], 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

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

(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 $\hat{\boldsymbol{\gamma}}_{\boldsymbol{jm}}=\mathrm{argmax}_{\boldsymbol{\gamma_{jm}}}L_{jm}(\mathbf{0},\boldsymbol{\gamma_{jm}})$. 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.

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.

Comparisons with 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 $(\hat{\beta}_1, \ldots, \hat{\beta}_q)$. 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 (Supplemental Materials). 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.

mixWAS was compared against the PheWAS methods described in the PheWAS section, as well as an oracle score test, which is a score test using only de-correlated z −scores 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 Figure S1. 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 (Supplemental Material), 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 for each individual was computed as the median age for each continuous laboratory measure, corresponding to the median laboratory measures used for the analysis For binary diseases, the median age of the ICD-9 code assignments for each individual of 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 (Supplemental Material). 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[42,43], and these SNPs were annotated using canonical pathways from the Human Molecular Signatures Database (MSigDB)[44].

Discussion

Recent initiatives have made EHR-linked genetic data increasingly available for genomics research, providing extensive, well-characterized phenotype data that opens unprecedented opportunities for cross-cohort learning. The integration of genetic data with detailed clinical records from multiple health systems, institutions, and population studies allows for a more powerful and reproducible examination of genetic associations with multiple diseases and traits, illuminating potential shared genetic architectures among diverse phenotypes. However, despite these opportunities, significant data-sharing restrictions and methodological challenges limit the full potential of multi-EHR analyses in identifying multi-phenotype associations (MPA) and other cross-cohort genetic insights.

To overcome these barriers, we developed *mixWAS*, a computationally efficient, one-shot, lossless method designed for flexible integration of summary statistics across multiple datasets.

This method enables robust identification of genetic variants associated with multiple phenotypes—both binary and continuous—and facilitates a comprehensive exploration of shared genetic variants underlying various traits. Beyond MPA detection, *mixWAS* is applicable to a wide range of multi-outcome studies across distributed data, underscoring its utility as a versatile tool for cross-cohort learning in complex multi-EHR settings.

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, a scenario often seen in real-world disease phenotypes[45]. In addition, type 1 errors were well controlled in all settings, as reflected by the power when the simulated effects equal to zero (Figure 2). 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 S2 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 3). 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 3). 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 4a).

Among the traits, lipid levels (including LDL, HDL, Cholesterol, and Triglycerides) shared the largest number of associated genetic variants, followed by BMI (Figure 4b). 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 S3). 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 5). 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. Li, R. *et al.* Lossless integration of multiple electronic health records for identifying pleiotropy using summary statistics. *Nat Commun* **12**, 1–10 (2021).
2. 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).
3. 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).
4. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
5. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
6. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2–S8 (2017).
7. 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).
8. Kho, A. N. *et al.* Electronic medical records for genetic research: Results of the eMERGE consortium. *Sci Transl Med* **3**, (2011).
9. 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).
10. 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).
11. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).
12. 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).
13. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009).

14. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
15. 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).
16. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, (2014).
17. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51**, 1339–1348 (2019).
18. 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).
19. 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).
20. Chun, S. *et al.* Leveraging pleiotropy to discover and interpret GWAS results for sleep-associated traits. *PLoS Genet* **18**, e1010557 (2022).
21. Zhang, X. *et al.* Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and nervous system disorders. *Nat Commun* **13**, (2022).
22. 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).
23. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).
24. Nabirotchkin, S. *et al.* Next-generation drug repurposing using human genetics and network biology[Formula presented]. *Current Opinion in Pharmacology* vol. 51 78–92 Preprint at https://doi.org/10.1016/j.coph.2019.12.004 (2020).
25. Bellou, E., Stevenson-Hoare, J. & Escott-Price, V. Polygenic risk and pleiotropy in neurodegenerative diseases. *Neurobiology of Disease* vol. 142 Preprint at https://doi.org/10.1016/j.nbd.2020.104953 (2020).
26. Li, C., Yang, C., Gelernter, J. & Zhao, H. Improving genetic risk prediction by leveraging pleiotropy. *Hum Genet* **133**, 639–650 (2014).
27. 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).
28. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
29. 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).
30. Johnson, E. C. *et al.* A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry* **7**, 1032–1045 (2020).
31. 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).

32. Diogo, D. *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nat Commun* **9**, (2018).
33. Deflaux, N. *et al.* Demonstrating paths for unlocking the value of cloud genomics through cross cohort analysis. *Nat Commun* **14**, (2023).
34. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* **50**, 229–237 (2018).
35. 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).
36. 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).
37. 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).
38. 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).
39. 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).
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. 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).
43. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1–11 (2017).
44. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
45. Elgart, M. *et al.* Correlations between complex human phenotypes vary by genetic background, gender, and environment. *Cell Rep Med* **3**, 100844 (2022).