

RESEARCH ARTICLE

Optimized phenotype definitions boost GWAS power

Michael Zietz¹, Kathleen LaRow Brown², Undina Gisladdottir², Nicholas P. Tatonetti^{1*}

1 Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, California, United States of America, **2** Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, New York, United States of America

* Nicholas.Tatonetti@cshs.org



OPEN ACCESS

Citation: Zietz M, Brown KL, Gisladdottir U, Tatonetti NP (2026) Optimized phenotype definitions boost GWAS power. PLoS Comput Biol 22(7): e1014431. <https://doi.org/10.1371/journal.pcbi.1014431>

Editor: Lin Hou, Tsinghua University, CHINA

Received: November 19, 2025

Accepted: June 7, 2026

Published: July 1, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1014431>

Copyright: © 2026 Zietz et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Complex diseases are a major challenge, and genetics underlie a large fraction of the risk for these diseases. Observational data are helpful for this research due to large scale, cost-effectiveness, information on many different conditions, and future scalability, but they reflect factors such as healthcare processes, access to care, and broader societal effects like systemic biases. Here, we introduce MaxGCP, a phenotyping method designed to purify the genetic signal in observational data. MaxGCP optimizes a phenotype definition to maximize its coheritability—the genetic covariance between two traits normalized by their phenotypic standard deviations—with the complex trait of interest. Unlike previous phenotype-combination methods, MaxGCP is phenotype-specific, has linear computational complexity in the number of features, and does not require manual feature selection. In an analysis of stroke, we found that MaxGCP boosts study power by more than 13 percent compared to conventional, single-code phenotype definitions. MaxGCP is a powerful tool for genetic discovery in observational data, and we anticipate that it will be broadly useful for studying complex diseases using observational data.

Author summary

Genome-wide association studies (GWAS) seek to identify genetic variants associated with disease risk. However, phenotype definitions derived from electronic health records contain substantial environmental noise that reduces statistical power. We developed MaxGCP, a method that constructs optimized phenotype definitions by combining multiple observed phenotypes (e.g., diagnosis codes) into a single index that maximizes the shared genetic signal with a target disease. MaxGCP requires only summary-level genetic covariance estimates and a phenotypic covariance matrix, and it has an exact, closed-form solution that scales linearly with the number of input phenotypes. In simulations and real-data analyses of stroke and Alzheimer's disease using the UK Biobank, MaxGCP

Data availability statement: All data and code used in this study are publicly available. The MaxGCP software and example datasets can be accessed at the following GitHub repository: <https://github.com/tatonetti-lab/maxgcp>. Any additional supporting data are provided within the manuscript.

Funding: This work was supported by the National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) <https://www.nigms.nih.gov/> grant R35GM131905 awarded to NPT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

consistently improved GWAS sensitivity compared to conventional single-code phenotype definitions, with the strongest gains observed when high-quality genetic covariance estimates were available. MaxGCP is freely available as a Python package and can be applied broadly to improve genetic discovery in observational data.

1. Introduction

The genetic basis of complex disease remains imperfectly characterized despite decades of theoretical and experimental advances and thousands of studies into the genetics of various complex diseases [1]. Piecing apart this causality will involve—among many other things—characterizing the associations between individual genetic variants and disease risk for millions of genetic variants and thousands of phenotypes. Observational data provide a practical source of information about complex diseases due to their abundance, representativeness, low cost to use, and the fact that they have already been collected. However, observational data bring additional challenges such as incompleteness, noise, and bias [2], which reduce study power [3,4].

By prioritizing genetic over environmental causes, careful phenotype definitions are one way to strengthen the genetic signal available in observational data [5–7]. Complex phenotypes are often correlated, leading researchers to analyze multiple phenotypes at once in the hopes of increasing study power [8]. However, combining related phenotypes in an optimal way is challenging, and previous work has dealt primarily with combining phenotypes that are known *a priori* to be related to the phenotype of interest [6,7,9]. Such approaches are inapplicable to phenome-wide analyses of bio-bank datasets, in which thousands of phenotypes are available. Moreover, by requiring features to be pre-selected, such methods are not able to incorporate weak genetic signals across large numbers of seemingly unrelated phenotypes. A method that can be applied to thousands of phenotypes without manual feature selection could increase power for phenome-wide studies. Existing multi-trait methods such as MTAG [10] are phenotype-specific but require pairwise LD score regressions among all features, resulting in quadratic computational complexity that limits scalability to large feature sets.

Here we present MaxGCP, a new statistical method for defining phenotypes in GWAS. MaxGCP fits a linear combination of phenotypes to maximize heritability—the genetic covariance between two traits normalized by their phenotypic standard deviations—with the phenotype of interest. Because MaxGCP optimizes for genetic signal relevant to a specific target trait, irrelevant features receive near-zero weight automatically, eliminating the need for manual feature pre-selection. As an example, a MaxGCP phenotype (y) for stroke could be defined as the linear combination of heart disease, hypertension, and other risk factors (e.g., $y = 1.2 \times \text{heart disease} - 0.1 \times \text{hypertension} + \dots$) that maximizes the heritability between stroke and y . By boosting the genetic signal using information from other phenotypes, this method is able to reduce environmental noise and improve GWAS power.

2. Results

The goal of MaxGCP is to fit a phenotype y to maximize the coheritability between y and z , the target phenotype. In the previous example, $y = 1.2 \times \text{heart disease} - 0.1 \times \text{hypertension} + \dots$, the features are heart disease, hypertension, etc., and the coefficients (i.e., 1.2, -0.1) are optimized by MaxGCP. A general overview of MaxGCP follows.

We chose coheritability as the optimization objective because it directly captures the quantity we seek to maximize in GWAS: the genetic signal relative to total phenotypic noise. Coheritability is proportional to genetic covariance and inversely proportional to phenotypic standard deviation, so maximizing it simultaneously amplifies shared genetic signal and suppresses environmental variance. By contrast, maximizing heritability alone (as in MaxH) would not ensure relevance to the target phenotype, and maximizing genetic correlation would not account for the overall noise level of the index.

Suppose any phenotype, A , can be written as a sum of genetic and environmental components (i.e., $A = g_A + e_A$). The coheritability between two phenotypes, A and B is

$$h_c^2(A, B) = \frac{\text{Cov}(g_A, g_B)}{\sqrt{\text{Var}(A)\text{Var}(B)}}.$$

Define an index y as a linear combination of feature phenotypes $x_1, \dots, x_m$, (i.e., $y = \mathbf{x}^T \beta$). MaxGCP finds the $\hat{\beta}$ that maximizes coheritability between index y and target z . Because the index is linear, MaxGCP has an exact solution. Let $\mathbf{v}$ be a vector of genetic covariances between the features and the target (i.e., $v_i = \text{Cov}(g_{x_i}, g_z)$). Let $\mathbf{P}$ be the covariance matrix of the features (i.e., $P_{ij} = \text{Cov}(x_i, x_j)$). The optimal coefficient vector can be obtained by solving a linear system and is the following:

$$\hat{\beta} = \frac{\mathbf{P}^{-1} \mathbf{v}}{\sqrt{\mathbf{v}^T \mathbf{P}^{-1} \mathbf{v}}}.$$

A derivation is provided in the methods.

2.1. Evaluation using simulated data

We first evaluated MaxGCP using simulated data. In simulation, effect sizes can be known exactly, making it easy to evaluate whether MaxGCP achieves its aims. We simulated 5000 realistic phenotypes (see Methods; [S1 Fig](#)) and computed an optimal index for each phenotype using our method, MaxGCP. MaxGCP phenotypes were more heritable and coheritable than naively defined phenotypes (i.e., presence or absence of a single diagnosis code; [Fig 1](#)). This indicates that MaxGCP is able to prioritize specific genetic signals. When used in GWAS, MaxGCP phenotypes increased power compared to naive phenotypes ([Fig 2](#), [S1 Table](#)). MaxGCP's performance depended on the quality of inputs used to fit the model. Excellent inputs (i.e., noiseless genetic covariance estimates) result in nearly optimal performance, while poor inputs (i.e., noisy estimates) result in worse performance.

As a sensitivity analysis, we repeated this analysis using a variety of different simulation parameters ([Fig 3](#)). The results of this analysis are intuitive. Increasing the heritability of phenotypes increased GWAS power for all methods. Increasing the genetic correlation between phenotypes had a strong positive effect on MaxGCP GWAS but little effect on naive GWAS. Missingness (fraction of cases randomly switched to controls) reduced performance for all methods. Increasing variance in estimates of genetic covariance decreased MaxGCP performance and had no effect on naive phenotype GWASs, which do not take genetic covariance into account. Overall, the results of this sensitivity analysis were consistent with the previous results: (1) MaxGCP outperformed the naive phenotype (in sensitivity and F1, not in specificity), and (2) noise in the genetic covariance matrix degraded performance (for all metrics).

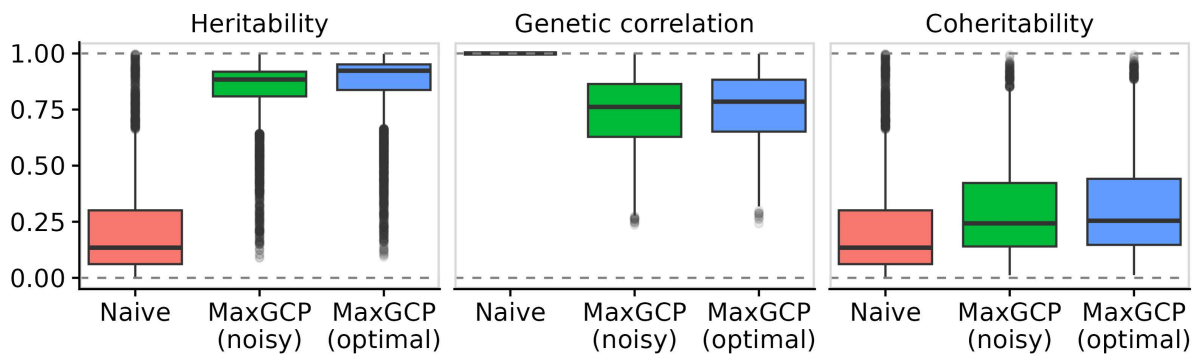


Fig 1. MaxGCP amplifies the genetic signal in simulated data. Shown is a comparison across phenotype definitions for 5000 simulated phenotypes. Naive definitions are single-code binary phenotypes with missingness. MaxGCP definitions are fit using either a noisy or noiseless ("optimal") genetic covariance matrix. Genetic correlation and coheritability are computed relative to the naive definitions.

<https://doi.org/10.1371/journal.pcbi.1014431.g001>

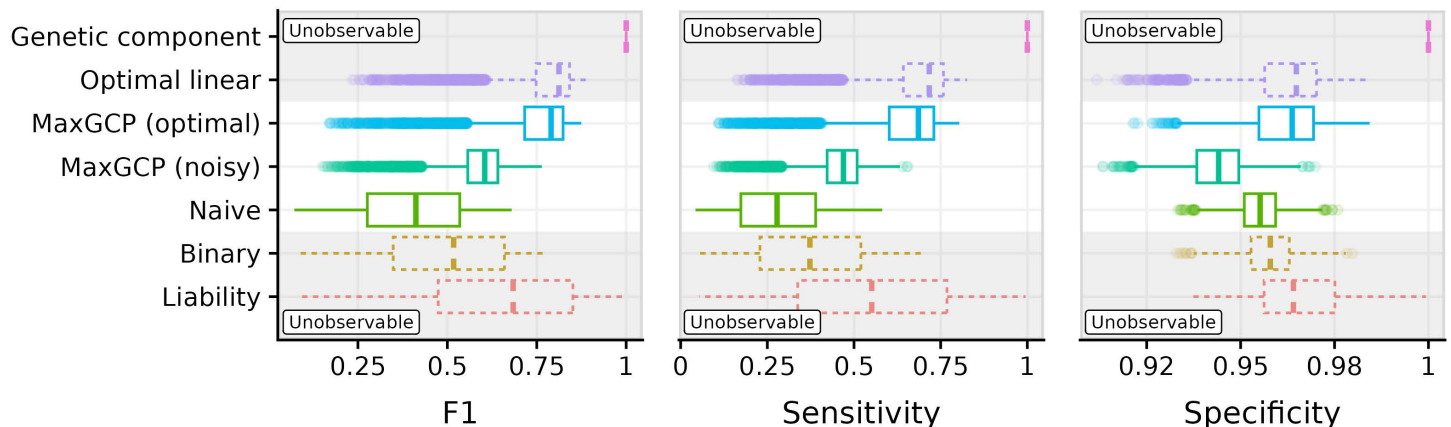


Fig 2. MaxGCP improves GWAS performance in simulated data. Shown are GWAS performance metrics for seven methods to define phenotypes in simulated data. Only naive and MaxGCP can be used in real data. These are indicated by the white strip through the middle of the plot. MaxGCP noisy and optimal provide bounds of reasonable performance, with the difference being the quality of genetic covariance estimates. Performance metrics are computed using a p-value cutoff of 0.05 to declare variants as positive or negative, and associations are compared to those correctly identified by the genetic component GWAS.

<https://doi.org/10.1371/journal.pcbi.1014431.g002>

We used a nominal p-value threshold of 0.05 for the simulation evaluation because the simulation scale (10,000 individuals, 100,000 variants, 1% heritability) yields very few genome-wide significant hits for any method, making comparative evaluation at stringent thresholds unstable. This threshold is applied uniformly across all methods and does not favor any particular approach. We note that MaxGCP operates on the phenotype definition rather than on individual variant associations; consequently, it does not selectively target variants of any particular effect size but uniformly amplifies the genetic signal component of the phenotype.

2.2. Real data

Next, we evaluated MaxGCP using real phenotypes from the UK Biobank. This is a more challenging comparison because the true genetic architecture is not known. Our solution to this problem was to compare our GWAS results to external validation data, in the form of large, high-quality GWAS that do not include samples from the UK Biobank. We

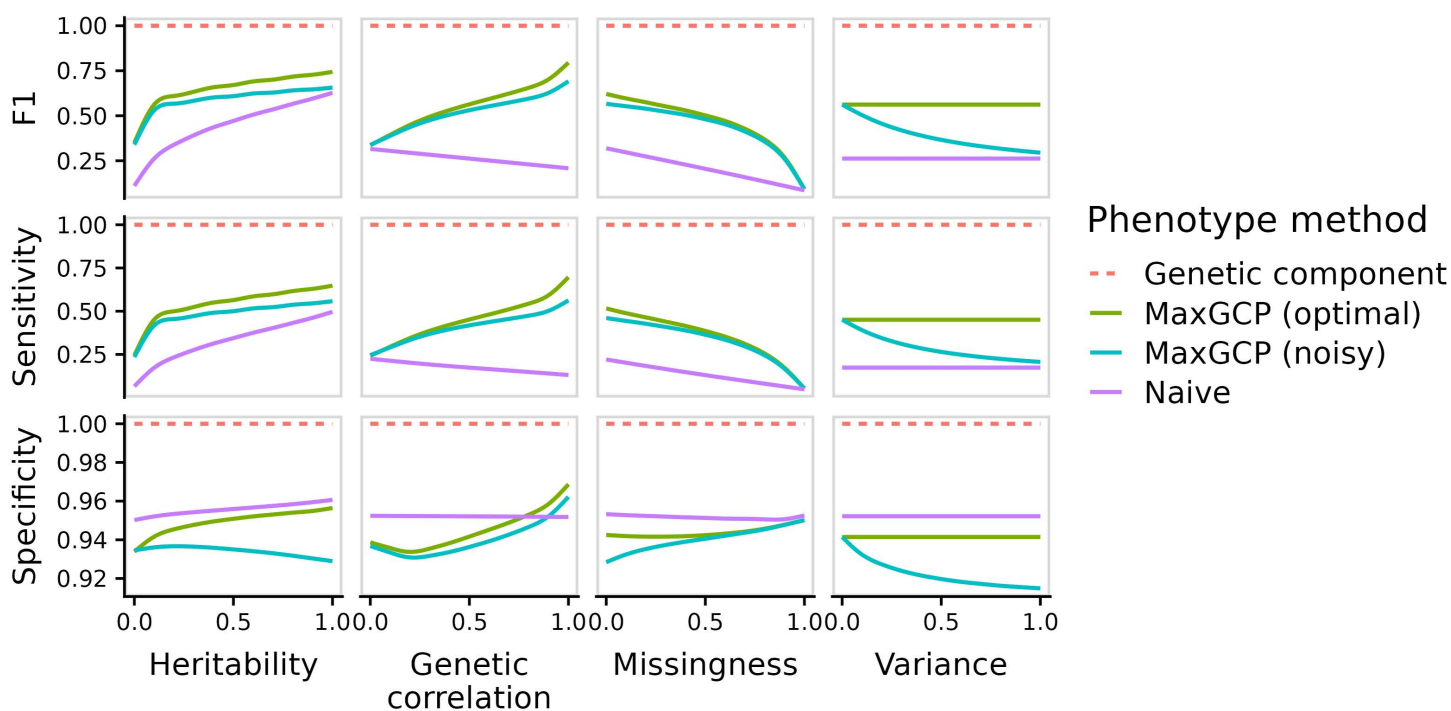


Fig 3. Sensitivity of GWAS performance to simulation parameters. Shown are marginal variations of parameters while all others are held constant at the realistic values. Each simulation parameter setting was simulated across 50 phenotypes and 100 replicates. GWAS performance was evaluated as before: relative to variants correctly identified by the genetic component GWAS.

<https://doi.org/10.1371/journal.pcbi.1014431.g003>

used two GWAS datasets for this task: MEGASTROKE (of which we used the “any stroke”, “ischemic stroke”, “cardio-embolic stroke”, and “small vessel stroke” phenotypes) [11] and IGAP (Alzheimer’s disease) [12]. These studies used clinician-adjudicated phenotypes, which are more reliable than naive observational definitions (e.g., single instances of diagnosis codes from medical records) and, therefore, provide a high-quality reference for GWAS associations. As features for MaxGCP, we used the 50 most common ICD-10 codes in our cohort (see Methods). We ran GWAS using Plink 2 [13] and estimated genetic covariances using SumHer [14] (see Methods for further details).

Because there is an inherent trade-off between sensitivity and specificity, we first compared methods at fixed thresholds. We evaluated sensitivity at 90% specificity and specificity at 10% sensitivity. MaxGCP performed well in this evaluation, with the full-cohort version performing best among all the methods for Alzheimer’s and three of the four stroke types (Fig 4A, Table 1). At a single p-value threshold ($p < 0.05$), MaxGCP improved sensitivity at the expense of specificity for all five phenotypes (Fig 4B, Table 1).

Overall, we draw the same conclusions as in the simulation. First, MaxGCP can improve GWAS sensitivity. Second, the quality of input data affects MaxGCP’s performance. MaxGCP showed excellent performance in the full cohort, but it underperformed in the subsampled cohort, where estimates of genetic covariance are less accurate. Third, MaxGCP improves sensitivity at the expense of specificity. Fourth, MaxGCP improves sensitivity more than it lowers specificity, meaning overall performance is generally superior (e.g., when comparing methods at equal specificity). We note that absolute sensitivity values are low for all methods (e.g., ~12% for MaxGCP vs ~6.5% for naive at $p < 0.05$ for any stroke), reflecting the inherent difficulty of cross-cohort variant replication due to differences in genotyping platforms, imputation strategies, and cohort sizes between the UK Biobank and the reference GWAS. The relevant comparison is therefore the relative improvement, which consistently favors MaxGCP.

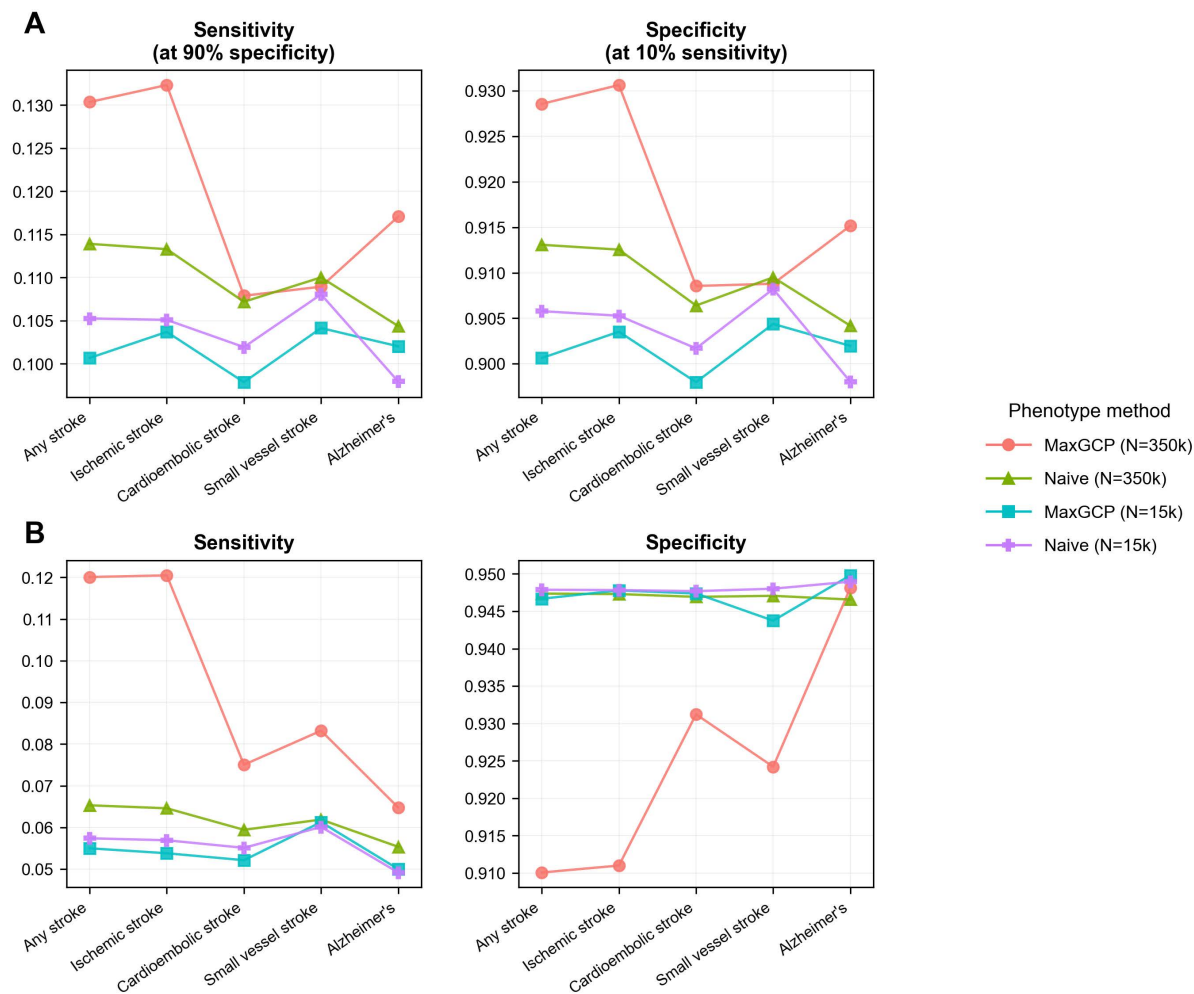


Fig 4. GWAS performance for real data. Phenotype definitions are compared in their ability to match variant associations with the MEGASTROKE and IGAP GWASs. Each analysis was performed in both the full and subsampled cohorts (N = 350k and N = 15k, respectively). A Sensitivity and specificity are evaluated at 90% specificity and 10% sensitivity, respectively, providing a fair comparison across methods. B Sensitivity and specificity are evaluated using a single p-value threshold ($p < 0.05$).

<https://doi.org/10.1371/journal.pcbi.1014431.g004>

MaxGCP prioritizes relevant features, including known risk factors like heart disease for stroke and type 2 diabetes for Alzheimer's (Fig 5 and S2 Table). The top features for stroke phenotypes were cerebral infarction itself (ICD-10 code I63), acute ischemic heart disease (code I24), and pulmonary heart diseases (I27). For Alzheimer's disease, the top MaxGCP features were Alzheimer's disease (G30), pulmonary heart diseases (I27), and type 1 diabetes (E10), with the latter two having negative coefficients.

MaxGCP coefficients are driven by both genetic and phenotypic covariance. Genetic covariance estimates were strongly, though imperfectly, correlated to MaxGCP coefficients (Pearson correlations between 0.54 and 0.94; Fig 6). In other words, the MaxGCP coefficient for a feature (e.g., hypertension) is predicted by the strength of genetic covariance between the feature and the target (e.g. Alzheimer's disease). The fact that this correlation is not perfect suggests that MaxGCP correctly includes information about phenotypic covariances, which in effect minimizes phenotypic variance while maximizing genetic covariance.

Table 1. GWAS performance in real data. “P-value” results were computed by applying a p-value threshold of 0.05. “Fixed” results were computed by applying a fixed threshold to one metric (either 90% specificity or 10% sensitivity) and evaluating performance in the other. In both cases, predicted associations were compared to those made in the corresponding target GWAS (MEGASTROKE or IGAP). These values are the same as shown in Fig 4.

Target	Cohort	Phenotype	P-value		Fixed	
			Sens.	Spec.	Sens.	Spec.
Any stroke	N = 350k	MaxGCP	0.120	0.910	0.130	0.929
		Naive	0.065	0.947	0.114	0.913
	N = 15k	MaxGCP	0.055	0.947	0.101	0.901
		Naive	0.057	0.948	0.105	0.906
Ischemic stroke	N = 350k	MaxGCP	0.120	0.911	0.132	0.931
		Naive	0.065	0.947	0.113	0.913
	N = 15k	MaxGCP	0.054	0.948	0.104	0.904
		Naive	0.057	0.948	0.105	0.905
Cardioembolic stroke	N = 350k	MaxGCP	0.075	0.931	0.108	0.909
		Naive	0.059	0.947	0.107	0.906
	N = 15k	MaxGCP	0.052	0.947	0.098	0.898
		Naive	0.055	0.948	0.102	0.902
Small vessel stroke	N = 350k	MaxGCP	0.083	0.924	0.109	0.909
		Naive	0.062	0.947	0.110	0.909
	N = 15k	MaxGCP	0.061	0.944	0.104	0.904
		Naive	0.060	0.948	0.108	0.908
Alzheimer's	N = 350k	MaxGCP	0.065	0.948	0.117	0.915
		Naive	0.059	0.950	0.115	0.914
	N = 15k	MaxGCP	0.050	0.950	0.102	0.902
		Naive	0.053	0.950	0.104	0.904

<https://doi.org/10.1371/journal.pcbi.1014431.t001>

MaxGCP GWAS identified target peaks that are not present in the naive GWAS but are in the target (S2 Fig and S3 Fig). If MaxGCP functions as expected, then a Manhattan plot of its GWAS results should look more similar to the target (e.g., MEGASTROKE) than the naive phenotype itself (e.g., ICD-10: I63). Moreover, MaxGCP results should not resemble the results for a top feature more than they resemble the target, as this would suggest that MaxGCP increases the number of hits by including false positives. Comparisons of Manhattan plots from this analysis confirm that MaxGCP is performing correctly, providing additional evidence that MaxGCP correctly identifies phenotype-specific information.

2.3. Comparison to MTAG

We compared MaxGCP's performance in real data to MTAG [10]. We were not able to perform a full comparison for two reasons. First, MTAG requires pairwise LD score regressions among all feature phenotypes. As a consequence, the number of LD score regressions required for MTAG grows quadratically in the number of features used. MaxGCP, on the other hand, requires only a linear number. Because of this, the MTAG comparison was not computationally feasible for us using large numbers of features (e.g., ≥ 50). Second, because the IGAP Alzheimer's disease GWAS summary statistics do not include allele frequencies, we excluded Alzheimer's from this comparison.

For this comparison, we used four different phenotype methods: the naive phenotype (a single ICD code), MTAG (with 10 features), MaxGCP (with 10 features), and MaxGCP (unrestricted features). The unrestricted MaxGCP was the model used previously, where features are selected for each phenotype using quality control only. For the fixed feature

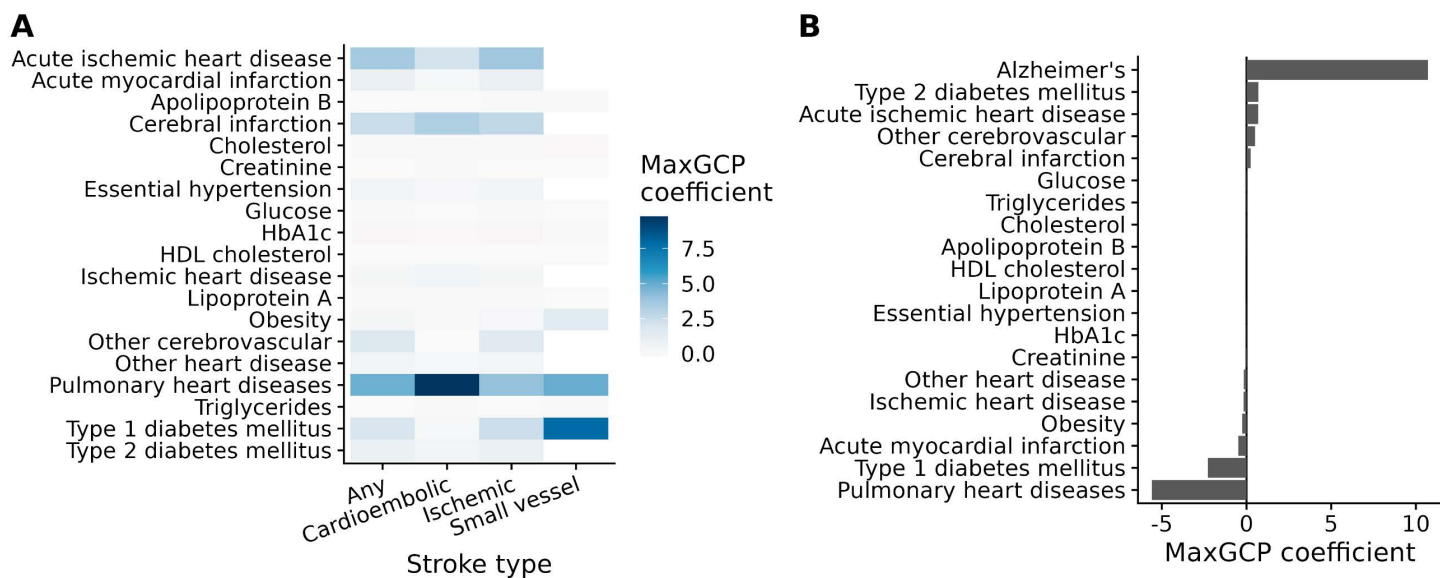


Fig 5. Top MaxGCP coefficients in real data. **A.** Top coefficients for MEGASTROKE phenotypes. **B.** Top coefficients for Alzheimer's disease. Shown are the fitted coefficient values of the MaxGCP model. The corresponding MaxGCP phenotypes are linear combinations of the features with the shown coefficients. Most features received essentially zero weight, so we selectively visualize top coefficients along with known risk factors to present an interpretable view into MaxGCP.

<https://doi.org/10.1371/journal.pcbi.1014431.g005>

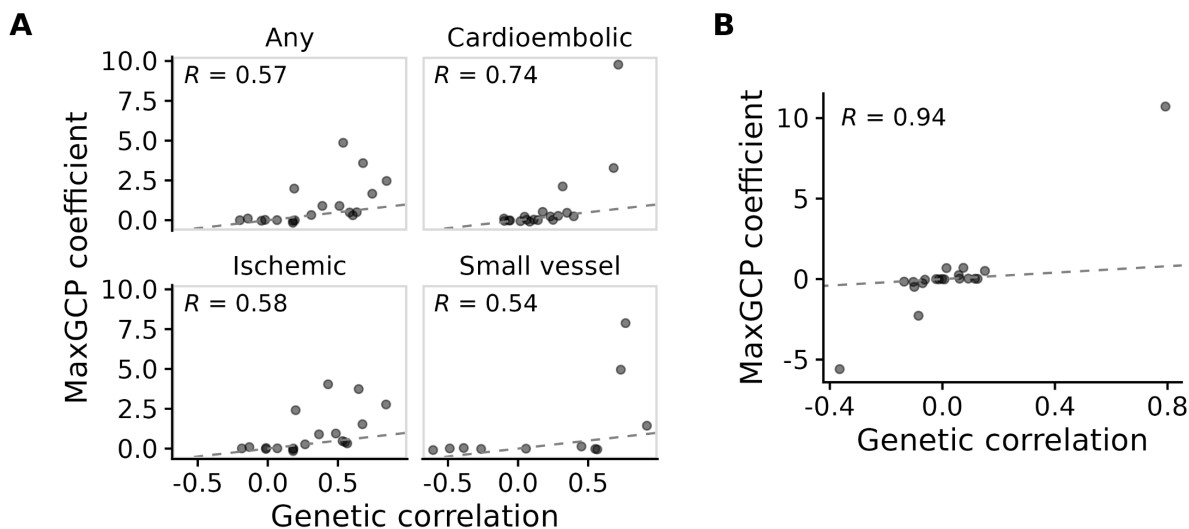


Fig 6. Genetic covariance vs MaxGCP coefficients in real data. Comparison of each feature phenotype's genetic correlation with the target to the coefficient it received in the MaxGCP phenotype. **A.** MEGASTROKE phenotypes. **B.** Alzheimer's disease. Each point represents a single feature phenotype. The y-axis represents the coefficient it received in the MaxGCP phenotype, while the x-axis represents the estimated genetic correlation between the target phenotype (i.e., stroke on the left, Alzheimer's on the right) and the feature phenotype. Shown are the same features as in Fig 5 and S2 Table.

<https://doi.org/10.1371/journal.pcbi.1014431.g006>

comparisons, we selected the 10 features in decreasing order of their absolute genetic correlation to the target phenotype, as this is the most natural and method-agnostic selection criterion given the available information. These features substantially overlap with, but are not identical to, the features receiving the highest weights in the unrestricted MaxGCP

model, because MaxGCP coefficients are correlated with but not identical to genetic correlations (Fig 6). Because the p-value cutoff plays such a large role in performance in this comparison, we evaluated across a range of p-value cutoffs.

We found that MaxGCP performs similarly to or better than the naive phenotype throughout this comparison. Without restricting the number of features used, MaxGCP outperformed MTAG throughout this comparison. With only 10 features, MaxGCP performed similarly or slightly better than MTAG in all comparisons. For example, with a p-value cutoff of 0.05, MTAG outperformed MaxGCP-10 for the All stroke phenotype (MTAG sensitivity at 95% specificity was 0.089, MaxGCP was 0.085) when using 10 features for both methods. However, with a p-value cutoff of 0.005, MaxGCP-10 performed better in the same metric (MTAG: 0.137, MaxGCP 0.144).

Fig 7 shows the results of this comparison. Overall, MaxGCP was the best performing method, particularly when unrestricted in the features it could use. MTAG performed well for all stroke and acute ischemic stroke, but underperformed for cardioembolic stroke. Compared to MTAG, MaxGCP has the advantages of significantly better computational complexity and no dependence on allele frequency information. This allowed us to use far more features for MaxGCP than for MTAG, leading to improved performance and reduced computation time.

3. Discussion

Here we described MaxGCP, a method for defining phenotypes in observational data. MaxGCP takes two inputs—a phenotypic covariance matrix (feature phenotypes × feature phenotypes) and a vector of genetic covariance estimates (feature phenotypes × the target phenotype)—and it fits a linear combination to maximize coheritability with a target phenotype. Our analysis used both real and simulated data and provided two takeaways. First, MaxGCP improves the

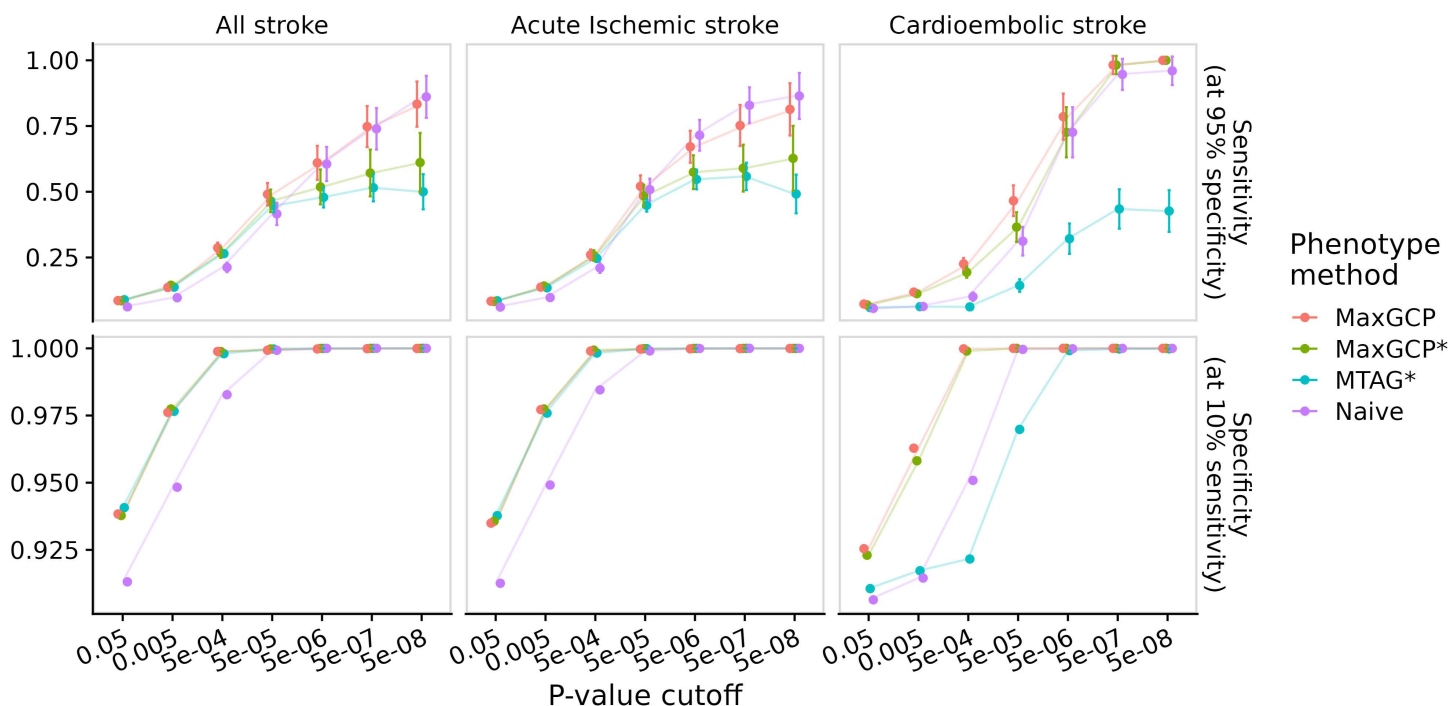


Fig 7. Comparison between MTAG and MaxGCP. Shown are the sensitivities and specificities of MaxGCP and MTAG across three stroke phenotypes and a range of p-value cutoffs (0.05 to 5×10^{-8}). The top row shows sensitivity (at 95% specificity) and the bottom row shows specificity (at 10% sensitivity). Four phenotype methods are compared: MaxGCP (unrestricted, using all available features), MaxGCP* (restricted to 10 features), MTAG* (using the same 10 features), and Naive (a single ICD-10 code). The key result is that MaxGCP (red) performs comparably to or better than MTAG* (teal) across all phenotypes and thresholds.

<https://doi.org/10.1371/journal.pcbi.1014431.g007>

performance of GWAS. It increased power throughout our comparisons. While it reduces specificity, it increases sensitivity more than it reduces specificity, meaning that, overall, it performs better. Second, MaxGCP's performance depends on the quality of inputs used. In real data, MaxGCP approximately doubled GWAS sensitivity relative to naive phenotype definitions (e.g., from 6.5% to 12.0% for any stroke at $p < 0.05$; Fig 4). While both absolute values are low—reflecting the inherent difficulty of cross-cohort variant replication—the relative improvement demonstrates that MaxGCP successfully amplifies genetic signal. However, MaxGCP can reduce power given poor estimates, meaning that quality control is critical when applying MaxGCP.

Phenotype-specificity is the key feature that distinguishes MaxGCP from most previous methods. Large, multi-phenotype datasets like the UK Biobank include many phenotypes that are not obviously related to one another (e.g., motor vehicle collisions and hair color). Previous methods like MaxH also optimize a linear phenotype definition for genetic signal, but they do not restrict to a specific genetic signal. A nonspecific method like MaxH cannot be naively applied to data like the UK Biobank, because a maximally heritable phenotype definition may not have any relation to the specific phenotype under consideration. For example, it may be the case that a combination of eye color and height produces the most heritable phenotype definition in a dataset, but this is not helpful for a researcher interested in the genetics of stroke. Unlike previous methods such as MaxH, MaxGCP can be applied without pre-selecting features based on relevance.

MaxGCP can be applied broadly, but its application merits some consideration. First, the quality of inputs is very important to MaxGCP performance. In particular, this means that only high-quality genetic covariance estimates should be used. Estimates that imply out-of-bounds values (e.g., heritability less than zero) or have large standard errors (e.g., > 2), should be excluded. In our analysis of real data, we used the most common ICD codes and large cohorts to give reliable genetic covariance estimates, and we excluded features according to the above criteria. However, applying MaxGCP to thousands of features without additional or more stringent quality control may lead to poor fits due to the combined effects of many noisy estimates. There are likely additional filters, adjustments, and quality control steps that could be used to further improve performance in such cases. Overall, researchers applying MaxGCP should pay careful attention to the quality of genetic covariance estimates used as inputs to the model and should remove features that do not have high-quality estimates.

MaxGCP may not be appropriate in all settings. Specifically, we caution against using MaxGCP when (1) genetic covariance estimates are unreliable, such as when cohort sizes are small (we recommend $N > 50,000$ based on our subsampling analysis) or when feature phenotypes are rare, yielding noisy estimates; (2) the target phenotype has low genetic correlation with available feature phenotypes, in which case MaxGCP has little information to exploit; or (3) the goal is confirmatory analysis where specificity is prioritized over sensitivity, as MaxGCP trades specificity for sensitivity gains. In such cases, conventional single-code phenotype definitions may be preferable. MaxGCP is also inherently robust to the number of input features: irrelevant features receive near-zero weight because they have low genetic covariance with the target trait (Fig 6), and quality control on genetic covariance estimates naturally limits the effective feature set. This is consistent with recent work showing that adding genetically uninformative traits to multi-trait GWAS does not monotonically increase power [15]. We recommend including all features with reliable genetic covariance estimates rather than attempting to optimize the feature count.

We compared MaxGCP to MTAG, a previous method that is phenotype-specific. This comparison was not straightforward due to differences in the methods. For one, the MTAG software does not separate features and targets in the same way MaxGCP does. When running MTAG, one specifies a set of GWAS summary statistic files, and MTAG generates MTAG summary statistics corresponding to every input file. This means that MTAG must compute pairwise LD score regressions among all phenotypes. Conversely, MaxGCP only requires genetic correlation estimates between the features and target, so the number of LD score regressions needed grows linearly, not quadratically, in the number of features.

Due to high computational costs, MTAG could not be run for large numbers (≥ 50) of phenotypes in this study. This particularly complicated the analysis because we needed to select which features should be included. For example, when

the target phenotype was acute ischemic stroke, the features were the 10 phenotypes with the largest absolute genetic correlation to the target (i.e., acute ischemic stroke). While MaxGCP coefficients are correlated with genetic covariance estimates, this is not necessarily an optimal feature selection mechanism. Overall, we found that MaxGCP performed similarly to or better than MTAG. In addition, MaxGCP is faster, with its main computational cost growing linearly in the number of features, while MTAG grows quadratically due to pairwise LD score regressions.

MaxGCP's performance varied across phenotypes in the real data analysis, with the strongest improvements observed for any stroke and ischemic stroke, and more modest gains for Alzheimer's disease. Several factors likely contribute to this variation. First, the stroke phenotypes have higher genetic correlations with available feature phenotypes (e.g., cerebral infarction, ischemic heart disease, hypertension), providing MaxGCP with more informative signal to combine. Second, the genetic architecture of stroke subtypes involves substantial pleiotropy with cardiovascular risk factors that are well-represented among common ICD-10 codes, whereas Alzheimer's disease has a more distinct genetic basis dominated by the *APOE* locus, with fewer genetically correlated features in the available feature set. Third, stroke-related codes are more prevalent in the UK Biobank than Alzheimer's-related codes, leading to more reliable genetic covariance estimates for stroke features.

MaxGCP relies on estimated genetic covariances, which may be biased by indirect genetic effects. For example, consider the relationship between type 2 diabetes and the ZIP code that one inhabits. ZIP code does not have a genetic basis, but neighborhood segregation along ancestry lines creates an association between genetics and ZIP code. As a consequence, a study which does not fully adjust for the effects of ancestry could find a genetic correlation between ZIP code and type 2 diabetes, even though the relationship is not causally genetic. MaxGCP coefficients are correlated with genetic covariance estimates, so biases like this will bias MaxGCP. To mitigate such biases, we recommend several practical steps: (1) using LD score regression intercepts as a diagnostic, where large intercepts suggest confounding and the corresponding genetic covariance estimates should be down-weighted or excluded; (2) restricting to genetically homogeneous cohorts or including sufficient principal components as covariates in the upstream GWAS; (3) examining MaxGCP coefficients for biologically implausible features that may indicate confounding (e.g., socioeconomic proxies receiving large weights); and (4) comparing MaxGCP results across ancestry-stratified subgroups when possible. Future work could address this more directly, for example by devising a method in which MaxGCP penalizes genetic covariance estimates that are likely biased by population stratification.

As is a common approach in GWAS, we primarily considered binarized variant associations using p-value cutoffs. Here, we were interested in validating the method, so our choice of p-value cutoff was arbitrary. We used a larger (i.e., more lenient) p-value cutoff (0.05) to ensure that enough hits were present to evaluate the method. Accordingly, the sensitivity and specificity values reported here may not reflect performance under typical GWAS cutoffs (i.e., 5×10^{-8}). However, as shown in various studies over the last 15 years, the majority of SNP heritability is not captured by genome-wide significant variants but by many variants with small effects that do not reach statistical significance in current cohorts [16,17]. Thus, while we use more lenient p-value cutoffs, our results may still be informative about the variants/effects underlying the true genetic architecture of the phenotypes.

In terms of computational requirements, MaxGCP's optimization step is negligible: solving a linear system of dimension m (number of features) completes in under one second for $m=50$ on a standard laptop. The dominant cost is the upstream estimation of genetic covariances, which requires one LD score regression per feature for MaxGCP ($O(m)$) versus one per feature pair for MTAG ($O(m^2)$). For $m=50$ features, this translates to 50 regressions for MaxGCP versus 1,275 for MTAG. Each LD score regression takes approximately 5–10 minutes on a single CPU core, so the wall-clock time difference is substantial: 4–8 hours for MaxGCP versus 100–200 hours for MTAG (before parallelization). This ($O(m)$) vs ($O(m^2)$) scaling is the primary reason we were unable to run MTAG with 50 or more features in this study.

The sensitivity-specificity tradeoff observed with MaxGCP warrants discussion in terms of practical GWAS applications. In genetic discovery settings—where the goal is to identify novel loci for downstream validation—sensitivity is typically

prioritized over specificity, as false positives can be filtered through replication and fine-mapping. MaxGCP is therefore best suited for discovery-oriented analyses. For confirmatory analyses where specificity is paramount, users may prefer conventional phenotype definitions or apply more stringent p-value cutoffs.

Although we evaluated MaxGCP using binary ICD-10 code features, the method is formulated for any phenotypes with estimable genetic and phenotypic covariances. The mathematical framework applies equally to continuous phenotypes such as biomarker levels and quantitative traits. We used binary features because they are the most common phenotype representation in biobank EHR data, but continuous trait applications are a natural extension. Furthermore, MaxGCP produces a single optimized phenotype definition per target, so a MaxGCP-enhanced phenome-wide association study (PheWAS) would involve the same number of hypothesis tests as a conventional PheWAS, and standard multiple testing corrections (e.g., Bonferroni, FDR) can be applied in the usual way.

We used ICD-10 diagnosis codes as features because they are universally available in biobank-scale EHR data, making our results broadly reproducible. However, MaxGCP can accept any phenotype with estimable genetic covariance as a feature, including laboratory values, medication records, NLP-derived phenotypes, or curated phenotype algorithms, and richer feature sets may further improve performance. Our analysis was restricted to White British participants from the UK Biobank to reduce population stratification in genetic covariance estimates. Extending MaxGCP to multi-ancestry cohorts is an important future direction that would require ancestry-stratified analyses or cross-ancestry genetic covariance estimation methods.

A major limitation of the current analysis is that the reference GWAS are not well-captured by either naive or MaxGCP phenotypes (i.e., the overall sensitivity was low). We directly compared associated variants, but this is far from a perfect method of evaluation. In particular, differences in sequencing technology and imputation can lead to differences in GWAS results. As a potential future direction, we could consider whether GWAS associations are colocated with one another, as this would provide a more robust evaluation of the method and reduce the effects of unrelated technical choices in the production of reference GWAS. In future work, we also plan to evaluate MaxGCP using a more diverse set of phenotypes, as the current evaluation was limited to stroke and Alzheimer's disease.

4. Methods

MaxGCP seeks to boost the power of GWAS by prioritizing the genetic component of a phenotype definition. To simplify this work, we focus on linear combinations of feature phenotypes, which we refer to here as indices. The heritability, coheritability, and genetic correlation of an index can be written in terms of the feature phenotypes. This fact allows us to optimize coheritability efficiently. Define an index (y) as a linear combination of (feature) phenotypes ($x_1, \dots, x_m$), where β is an arbitrary coefficient vector.

$$y = \mathbf{x}^T \beta$$

Denote the genetic and phenotypic covariance matrices of the feature phenotypes as $\mathbf{G}$ and $\mathbf{P}$, respectively (i.e., $G_{ij} = \text{Cov}(g_{x_i}, g_{x_j})$ and $P_{ij} = \text{Cov}(x_i, x_j)$). Denote the vector of genetic covariances between z and each feature phenotype as $\mathbf{v}$ (i.e., $v_i = \text{Cov}(g_{x_i}, g_z)$). Using this notation, the definitions of heritability, genetic correlation, and coheritability are as follows:

$$h^2(\mathbf{x}^T \beta) = \frac{\beta^T \mathbf{G} \beta}{\beta^T \mathbf{P} \beta} \quad (1)$$

$$r_g(z, \mathbf{x}^T \beta) = \frac{\beta^T \mathbf{v}}{\sqrt{\text{Var}(g_z) \beta^T \mathbf{G} \beta}} \quad (2)$$

$$h_c^2(z, \mathbf{x}^T \beta) = \frac{\beta^T \mathbf{v}}{\sqrt{\text{Var}(z) \beta^T \mathbf{P} \beta}} \quad (3)$$

4.1. Coheritability optimization

To boost the genetic signal of an index, we seek to amplify genetic covariance while minimizing phenotypic variance. Coheritability is the appropriate objective for this task, as it is proportional to genetic covariance and inversely proportional to phenotypic variance. We seek to find coefficients that maximize the coheritability between y and z (i.e., $\hat{\beta} = \arg\max_{\beta} h_c^2(\mathbf{x}^T \beta, z)$). Note that the coheritability does not change based on a scaling of β . Accordingly, we can require that β be scaled so that the resulting index has unit variance. This allows us to simplify the problem to the following: maximize (w.r.t. β) $\beta^T \mathbf{v}$ subject to $\beta^T \mathbf{P} \beta = 1$. This can be solved using a Lagrange multiplier approach.

$$\begin{aligned} \mathcal{L} &= \beta^T \mathbf{v} - \lambda(\beta^T \mathbf{P} \beta - 1) \\ \nabla_{\beta} \mathcal{L} &= \mathbf{v} - 2\lambda \mathbf{P} \beta \\ \nabla_{\lambda} \mathcal{L} &= \beta^T \mathbf{P} \beta - 1 \end{aligned}$$

We set the gradient equal to zero and solve.

$$\mathbf{P} \hat{\beta} = \frac{\mathbf{v}}{2\lambda}$$

Applying the constraint, we can find λ :

$$1 = \hat{\beta}^T \mathbf{P} \hat{\beta} = \frac{\hat{\beta}^T \mathbf{v}}{2\lambda} \Rightarrow \lambda = \frac{\hat{\beta}^T \mathbf{v}}{2}$$

This allows us to rewrite the gradient equation:

$$\mathbf{P} \hat{\beta} = \frac{\mathbf{v}}{\hat{\beta}^T \mathbf{v}}$$

The phenotypic covariance matrix $\mathbf{P}$ is positive semi-definite by construction and is typically positive definite in practice unless features are perfectly collinear. In our analysis, the ICD-10 code features are correlated but not perfectly collinear, ensuring invertibility. If near-singularity were encountered (e.g., with very large or redundant feature sets), standard regularization approaches such as ridge-type shrinkage ($\mathbf{P} + \lambda \mathbf{I}$) could be applied. Our Python implementation uses `numpy.linalg.solve`, which raises an error for singular matrices, alerting the user. We scale the result to ensure the index has unit variance, and obtain the solution as follows:

$$\hat{\beta} = \frac{\mathbf{P}^{-1} \mathbf{v}}{\sqrt{\mathbf{v}^T \mathbf{P}^{-1} \mathbf{v}}}. \quad (4)$$

To summarize, a set of coefficients $\hat{\beta}$ that maximize the coheritability between $\mathbf{x}^T \beta$ and an arbitrary phenotype z can be determined by solving the linear system $\mathbf{P} \beta = \mathbf{v}$ and scaling the resulting coefficients by $\sqrt{\beta^T \mathbf{P} \beta}$.

4.2. Implementation

As shown above, MaxGCP is a straightforward operation that can be computed by libraries in nearly every programming language. We implemented this in a Python package [18] in order to pair the method with related functionality. Specifically, we added summarization functions that compute the heritability, genetic correlation, coheritability, and genetic loss given

fitted coefficients, using [Equations 1, 2, and 3](#), as well as methods to optimize heritability or genetic correlation instead of coheritability.

4.3. Simulation analysis

Real genotypes from the UK Biobank—10,000 (White British, QC-passing) individuals and 100,000 (imputed, HapMap3, QC-passing) variants—were used as the foundation for the simulation. We simulated 100 sets of 50 phenotypes, each with 5000 cases and controls, with a 1% mean heritability, 50% of variants having shared effects, zero mean phenotypic covariance, and 10% missingness using PhenotypeSimulator [19]. These parameters were calibrated against real data: we estimated the heritability, genetic correlation, and phenotypic correlation of the 20 most common ICD-10 codes in the UK Biobank and set simulation parameters to produce distributions matching or more conservative than the observed values (see [S1 Text](#); [S1 Fig](#)). For comparisons, positives were defined as variants simulated to be causal that were associated with nominal statistical significance ($p < 0.05$) in the genetic component GWAS. Negatives were defined as the opposite: non-causal variants with no statistically significant association detected in the genetic component GWAS. We qualified on the genetic component GWAS because this is the optimal GWAS phenotype and allows us to isolate the relevant performance of different methods from undesired randomness in the simulation.

We compared seven phenotype definitions in the simulation. The first two represent the simulation—the simulated continuous phenotypic liability (“liability”), and the binarized liability (“binary”)—and are not observable in real data. The binarized liability with missingness (“naive”) represents our best attempt at a realistic phenotype. MaxGCP models were fit using both noisy and noiseless genetic covariance matrices (“MaxGCP noisy” and “MaxGCP optimal”). Finally, we included two phenotypes that can only be observed in simulation. The optimal linear combination of phenotypes (“optimal linear”) was determined by directly regressing genetic components against phenotypes, and it represents the upper bound for the performance of a linear phenotype definition. The genetic component of each phenotype (“genetic component”) is included as it defines perfect performance.

4.4. Real data analysis

We used real data from the UK Biobank. Because we used the data to estimate genetic covariances, we applied the following QC filters: only White British individuals, excluded mismatched sex, outliers for heterozygosity or missingness, sex chromosome aneuploidy, relatedness to other participants, or those not included in the computation of genetic principal components. We used only imputed HapMap3 variants and excluded variants with high missingness. For feature phenotypes, we used ICD-10 diagnoses. We used SumHer [14] to estimate the genetic covariance between phenotypes. The GWAS inputs to SumHer were computed using Plink 2 with age, sex, and 10 genetic principal components as covariates. For each MaxGCP phenotype, we included the 50 most common ICD-10 codes in our cohort as features. This threshold balances informativeness against estimation noise: genetic covariance estimation via LD score regression becomes unreliable for low-prevalence binary phenotypes [20,21], and recent work on trait selection in multi-trait GWAS has shown that including genetically uninformative traits does not monotonically increase power [15]. After estimating genetic covariances between each feature and each target, we excluded features that had invalid estimates for a given target. We verified that neither the MEGASTROKE consortium nor the IGAP study included UK Biobank participants, ensuring no sample overlap between the discovery and validation datasets. The SumHer tagging files used for genetic covariance estimation were computed from European-ancestry samples, consistent with our White British cohort. The simulation code and all analysis scripts are available in the GitHub repository, and the MaxGCP optimization is fully deterministic given fixed inputs. Extensive details about data collection and preparation are available in the supplement.

Supporting information

S1 Text. Supplementary methods. Details of UK Biobank data processing and simulation parameters.
(DOCX)

S1 Fig. Distributions of heritability and correlation in real and simulated data. Shown is a comparison between simulation and the 20 most common ICD-10 codes in the UK Biobank, in terms of heritability, genetic correlation, and phenotypic correlation. The simulated phenotypes were generated using PhenotypeSimulator with expected heritability 0.01, genetic correlation 0.5, and phenotypic correlation 0. These simulation parameters were chosen so that the resulting simulation is similar or more conservative than the real data shown.
(TIF)

S2 Fig. Comparison of GWAS results for stroke. Shown are GWAS summary statistics for the target, MaxGCP phenotype, naive phenotype, and top feature for the MEGASTROKE “Any stroke” phenotype. MEGASTROKE shows the MEGASTROKE “Any stroke” GWAS summary statistics. MaxGCP shows the corresponding MaxGCP phenotype. Stroke shows the naive equivalent phenotype, defined as a single occurrence of any ICD-10 code starting with I63. The final facet shows GWAS summary statistics for pulmonary heart disease, the top feature in MaxGCP besides stroke itself.
(TIF)

S3 Fig. Comparison of GWAS results for Alzheimer’s disease. Shown are GWAS summary statistics for the target, MaxGCP phenotype, naive phenotype, and top feature for the IGAP Alzheimer’s disease phenotype. IGAP shows the IGAP GWAS summary statistics directly. MaxGCP shows the corresponding MaxGCP phenotype. Alzheimer’s shows the naive equivalent phenotype, defined as a single occurrence of any ICD-10 code starting with G30. The final facet shows GWAS summary statistics for type 2 diabetes mellitus, the top feature in MaxGCP besides Alzheimer’s disease itself (ICD-10 code G30).
(TIF)

S1 Table. GWAS performance of each phenotype definition in simulation. Each reported metric is the median of per-phenotype values. A p-value cutoff of 0.05 was used to declare variants as predicted positives or negatives, and associations were compared to those correctly identified by the genetic component GWAS, making it perform perfectly by construction. Methods marked by an asterisk (*) are unobservable in real data.
(DOCX)

S2 Table. MaxGCP coefficients for each target phenotype. Shown are selected features for each target phenotype. Each MaxGCP phenotype had a slightly different set of input features due to the QC process, in which noisy genetic covariance estimates were removed. Genetic correlations and standard errors were estimated using SumHer and the LDAK-Thin model.
(DOCX)

Author contributions

Conceptualization: Michael Zietz, Nicholas P Tatonetti.

Data curation: Michael Zietz.

Formal analysis: Michael Zietz.

Funding acquisition: Nicholas P Tatonetti.

Investigation: Michael Zietz, Nicholas P Tatonetti.

Methodology: Michael Zietz, Kathleen LaRow Brown, Undina Gisladdottir.

Project administration: Nicholas P Tatonetti.

Resources: Undina Gisladdottir, Nicholas P Tatonetti.

Software: Michael Zietz, Nicholas P Tatonetti.

Supervision: Nicholas P Tatonetti.

Validation: Michael Zietz.

Visualization: Michael Zietz.

Writing – original draft: Michael Zietz, Kathleen LaRow Brown, Undina Gisladdottir, Nicholas P Tatonetti.

Writing – review & editing: Kathleen LaRow Brown, Undina Gisladdottir, Nicholas P Tatonetti.

References

- Wainschtein P, Jain D, Zheng Z, TOPMed Anthropometry Working Group, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Cupples LA, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet.* 2022;54(3):263–73. <https://doi.org/10.1038/s41588-021-00997-7> PMID: [35256806](#)
- Hripsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc.* 2013;20(1):117–21. <https://doi.org/10.1136/amiajnl-2012-001145> PMID: [22955496](#)
- Sagreiya H, Altman RB. The utility of general purpose versus specialty clinical databases for research: warfarin dose estimation from extracted clinical variables. *J Biomed Inform.* 2010;43(5):747–51. <https://doi.org/10.1016/j.jbi.2010.03.014> PMID: [20363365](#)
- Duan R, Cao M, Wu Y, Huang J, Denny JC, Xu H, et al. An Empirical Study for Impacts of Measurement Errors on EHR based Association Studies. *AMIA Annu Symp Proc.* 2017;2016:1764–73. PMID: [28269935](#)
- Smith HF. A Discriminant Function For Plant Selection. *Annals of Eugenics.* 1936;7(3):240–50. <https://doi.org/10.1111/j.1469-1809.1936.tb02143.x>
- Ott J, Rabinowitz D. A principal-components approach based on heritability for combining phenotype information. *Hum Hered.* 1999;49(2):106–11. <https://doi.org/10.1159/000022854> PMID: [10077732](#)
- Jin J, Zhou J, Cho MH, Lange C, Lutz S, Silverman EK, et al. Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. *Human Heredity.* 2015;79(2):93–104.
- Aschard H, Vilhjálmsson BJ, Greliche N, Morange P-E, Trégouët D-A, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet.* 2014;94(5):662–76. <https://doi.org/10.1016/j.ajhg.2014.03.016> PMID: [24746957](#)
- Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, Raby B, et al. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat Appl Genet Mol Biol.* 2004;3:Article17. <https://doi.org/10.2202/1544-6115.1067> PMID: [16646795](#)
- Turley P, Walters RK, Maghziyan O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet.* 2018;50(2):229–37. <https://doi.org/10.1038/s41588-017-0009-4> PMID: [29292387](#)
- Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multi-ancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet.* 2018;50(4):524–37. <https://doi.org/10.1038/s41588-018-0058-3> PMID: [29531354](#)
- Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet.* 2013;45(12):1452–8. <https://doi.org/10.1038/ng.2802> PMID: [24162737](#)
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7. <https://doi.org/10.1186/s13742-015-0047-8> PMID: [25722852](#)
- Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat Genet.* 2019;51(2):277–84. <https://doi.org/10.1038/s41588-018-0279-5> PMID: [30510236](#)
- Suzuki Y, Ménager H, Brancotte B, Vernet R, Nerin C, Boetto C, et al. Trait selection strategy in multi-trait GWAS: Boosting SNP discoverability. *HGG Adv.* 2024;5(3):100319. <https://doi.org/10.1016/j.xhgg.2024.100319> PMID: [38872309](#)
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics.* 2010;42(7):565–9.
- Liu X, Li YI, Pritchard JK. Trans effects on gene expression can drive omnigenic inheritance. *Cell.* 2019;177(4):1022–34.
- Michael Zietz and Nicholas Tatonetti. Maximum genetic component phenotyping. <https://github.com/tatonetti-lab/maxgcp>, 2023.
- Meyer HV, Birney E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics.* 2018;34(17):2951–6. <https://doi.org/10.1093/bioinformatics/bty197> PMID: [29617944](#)

20. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, R Day F, Po-Ru L, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan L, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–41, 2015.
21. Ge T, Chen C-Y, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet*. 2017;13(4):e1006711. <https://doi.org/10.1371/journal.pgen.1006711> PMID: [28388634](https://pubmed.ncbi.nlm.nih.gov/28388634/)