The phenotype-genotype reference map: Improving biobank data science through replication

Authors

Lisa Bastarache, Sarah Delozier, Anita Pandit, ..., Jacob J. Hughey, Matthew Zawistowski, Josh F. Peterson

Correspondence

lisa.bastarache@vumc.org

The phenotype-genotype reference map (PGRM) is a curated set of GWAS associations that facilitates efficient phenome-wide replication studies with biobank data. The PGRM can be used to detect data corruption, assess model parameters, and explore factors that affect replicability of findings.

Bastarache et al., 2023, The American Journal of Human Genetics *110*, 1522–1533
September 7, 2023

https://doi.org/10.1016/j.ajhg.2023.07.012

CellPress

ARTICLE

The phenotype-genotype reference map: Improving biobank data science through replication

Lisa Bastarache,[1,*] Sarah Delozier,[1] Anita Pandit,[2] Jing He,[1] Adam Lewis,[1] Aubrey C. Annis,[2] Jonathon LeFaive,[2] Joshua C. Denny,[3] Robert J. Carroll,[1] Russ B. Altman,[4] Jacob J. Hughey,[1] Matthew Zawistowski,[2] and Josh F. Peterson[1,5]

Summary

Population-scale biobanks linked to electronic health record data provide vast opportunities to extend our knowledge of human genetics and discover new phenotype-genotype associations. Given their dense phenotype data, biobanks can also facilitate replication studies on a phenome-wide scale. Here, we introduce the phenotype-genotype reference map (PGRM), a set of 5,879 genetic associations from 523 GWAS publications that can be used for high-throughput replication experiments. PGRM phenotypes are standardized as phecodes, ensuring interoperability between biobanks. We applied the PGRM to five ancestry-specific cohorts from four independent biobanks and found evidence of robust replications across a wide array of phenotypes. We show how the PGRM can be used to detect data corruption and to empirically assess parameters for phenome-wide studies. Finally, we use the PGRM to explore factors associated with replicability of GWAS results.

Introduction

Over the last decade, experimental methods used to study the relationship between genetic variants and human disease at population scale have evolved significantly. Early genome-wide association studies (GWASs) used phenotype-specific cohorts with carefully curated exposures and outcomes.[1,2] More recent discovery research has gravitated toward multi-purpose biobanks where phenotypes are defined from a variety of sources, including electronic health records (EHRs).[3–7] As resources with both breadth and depth, biobanks support both GWASs for individual traits and phenome-wide association studies (PheWASs).[8] "High-throughput" methods, extensive catalogs of GWAS results, and precomputed GWAS × PheWAS associations, have led to a more fine-grained understanding of the genetic underpinnings of complex disease.[9–11]

Over time, the datasets used for GWASs have grown in both size and complexity. Multi-purpose biobanks have replaced recruited cohorts, and phenotyping algorithms have replaced physical exams and manual chart review. The effect of these changes on the reproducibility of GWAS results is not well understood. Prior work has shown that high-throughput discovery methods are prone to errors related to ascertainment bias, phenotype misclassification, errors in sample tracking, and missteps in designing analytic pipelines.[12–14] Detecting such problems in biobank data is challenging. While quality control (QC) metrics are routinely applied to the genotype data,[15] analogous metrics are not yet routinely used for phenotype data.

The abundance of phenotype-genotype associations from prior studies, combined with standardized phecode-based phenotypes, makes it possible to study replication on a phenome-wide scale. This can serve three purposes. First, phenome-wide replication studies can be used as a practical tool to assess overall data quality, as has been shown in previous studies.[16–18] Because GWAS results are highly replicable, a new cohort should be able replicate numerous previously discovered associations (given sufficient power) and inability to do so may indicate of data quality issues or an incompatibility between the original and replication analysis.[19–21] Second, phenome-wide replication studies can be used to validate a new analytical tool or assess analytical parameters used in PheWASs. Competing methods can be compared in terms of their capacity to replicate known findings.[22] Finally, these studies can be used to assess the factors associated with replication across heterogeneous cohorts.[23]

To facilitate phenome-wide replication studies across biobanks, we created the phenotype-genotype reference map (PGRM), a set of genotype associations selected from the National Human Genome Research Institute and European Bioinformatics Institute (NHGRI-EBI) GWAS catalog.[24] Best practice guidelines for replication studies emphasize the importance of aligning the phenotype definition, cohort composition, and statistical methods used

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; [2]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA; [3]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; [4]Department of Genetics, Stanford University, Stanford, CA, USA; [5]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
*Correspondence: lisa.bastarache@vumc.org
https://doi.org/10.1016/j.ajhg.2023.07.012.

in the initial study with those of the replication study, as well as ensuring adequate power.[25] To adhere to these recommendations, we manually reviewed all GWAS catalog associations, excluding those that were incompatible with a replication study in general-purpose biobanks and including information necessary for power calculations. Critically, the PGRM represented phenotypes as phecodes—ICD-based phenotypes developed to conduct PheWASs. Phecodes are a popular means of high-throughput phenotyping that have been applied to biobanks across the world.[26]

To validate the PGRM, we attempted to replicate PGRM associations in four biobanks—BioVU, Michigan Genomics Initiative biobank (MGI), UK Biobank (UKB), and BioBank Japan (BBJ)—finding replication was robust for the three genetic ancestries tested and across disease categories. We developed simplified replication measures that can be applied to any EHR-linked biobank cohort. Through a series of replication experiments, we show how the PGRM can be used to detect data corruption and assess modeling assumptions in phenome-wide studies. Finally, we identified factors that predicted replicability of GWAS associations, including expected factors such as reported odds ratio (OR) and p value, and unexpected factors such as disease category and the publication date of the original GWAS. We hope that the PGRM and publicly available source code will enhance the rigor of genetic studies using biobanks.

Material and methods

The PGRM is an annotated reference set of phenotype-genotype associations drawn from the GWAS catalog. In this section, we describe the creation of the PGRM, including retrieving and filtering GWAS catalog associations, mapping diseases to phecodes, normalizing genetic variants, and annotating associations by genetic ancestry and risk alleles. We then explain how we validated the PGRM by measuring replication rates on cohorts drawn from four independent biobanks, and how we used the PGRM to conduct a series of replication experiments. Finally, we describe an analysis using PGRM to explore factors that influence replicability of GWAS results.

Creating the PGRM

Retrieving and filtering GWAS catalog associations

Files from the publicly available GWAS catalog were downloaded on January 4, 2022 ("all associations v1.0.2," "all studies v1.0.2," and "all ancestry data v1.0"). Each row of the "all association" file represents a genotype/phenotype association. We included SNP/phenotype associations that met the following criteria: (1) were based on a single-nucleotide polymorphism (SNP) (as opposed to a haplotype or combination of SNPs), excluding the X or Y chromosome; (2) had a specified OR and confidence intervals; (3) reported a p value of $<5 \times 10^{-8}$; and (4) were based on a binary trait modeled as a logistic regression. We excluded continuous traits like blood pressure by using the OR_or_BETA column (all values < 1), mentions of measurement words (e.g., "increase," "decrease," "ratio") in the 95_CI_TEXT, and specification of continuous trait in the DISEASE_TRAIT or P-VALUE (TEXT) field (e.g., "age of onset").

Although the conditions necessary to ensure replication across cohorts are not well characterized, prior work has indicated that misalignment of study design, cohort composition, and phenotype definition decreases the likelihood of replication. Thus, to maximize the likelihood of replication in a test cohort, we filtered out GWAS catalog associations with characteristics that were not aligned with a general population biobank cohort. We manually extracted and normalized information regarding the phenotype, study cohort, and statistical model by using information from the following columns in the GWAS catalog: DISEASE_TRAIT, P-VALUE (TEXT), STUDY, and BACKGROUND_TRAIT as well as the study title. When necessary, we reviewed the original manuscript to resolve ambiguous information provided in these columns. We excluded associations from the PGRM for the following reasons.

(1) Phenotype misalignment: the catalog includes phenotypes that are qualified by severity (e.g., mild proteinuria), family history (e.g., familial lung cancer), age of onset (e.g., early-onset schizophrenia), and subtype (e.g., ER-breast cancer).
(2) Cohort misalignment: some catalog associations are based on specialized cohorts wherein all study participants share a characteristic, defined on the basis of disease (e.g., diabetic nephropathy in type 1 diabetes), exposure (e.g., Clostridium difficile infection in antibiotics-users), genetics (e.g., breast cancer in BRCA1/2 carriers), sex (e.g., hypertension in males), and young age (e.g., obesity in children). These are referred to as "background traits" in the GWAS catalog, and they are sometimes (though not always) annotated in the column of that name.
(3) Non-standard statistical models: while GWAS results are typically based on logistic regression and additive genotype, the catalog includes recessive or dominant associations models as well those generated from non-standard models (e.g., family-based models, interactions).

Mapping diseases and traits to phecodes

The GWAS catalog diseases and traits are annotated with the Experimental Factor Ontology (EFO).[27] We attempted to annotate all EFO terms present in the filtered list of associations with a matching phecode. We used WikiMedMap to generate candidate matches and manually chose matching phenotypes.[28] Each phenotype in the PGRM is labeled by one of 13 category labels. These labels were derived from the phecode category labels, which were modified for the purposes of this study. The metabolic/endocrine categories split into two separate categories. Because there were very few hematopoietic phenotypes, these were consolidated into the metabolic category. The "mental disorders" category was renamed "psychiatric," and phenotypes for Alzheimer disease and dementia were added to the "neurological" category. The EFO to phecode map can be found in Table S1.

Normalizing and annotating genetic variants

We used the Ensembl REST API (https://rest.ensembl.org/) to annotate each rsID present in the filtered list of associations. We recorded the allele_string, the location (chromosome and start/end position), and reference allele frequencies from gnomAD (EUR, AFR, EAS, SAS, AMR). For each variant, we stored the unique alleles defined in gnomAD and 1000 Genomes. A variant ID was created for each non-multi-allelic variant by concatenating the chromosome, start position, reference allele, and alternate allele (e.g., 1:62782860:T:C). Each row in the catalog was annotated

with a variant ID. Variants with more than two alleles specified were flagged as "multi-allelic" and excluded if the risk allele was ambiguous in the catalog. The annotation process was conducted twice: first with build GRCh37 and then with build GRCh38, so that the PGRM could be used on datasets from either build.

Defining risk alleles

The risk allele is defined as the allele that is associated with risk of the phenotype (i.e., OR > 1). For each association, we annotated the risk allele as reference ("ref") or alternative ("alt") according to the STRONGEST_SNP_RISK_ALLELE reported in the catalog. In cases where the catalog did not report a STRONGEST_SNP_RISK_ALLELE, we determined the risk allele by matching allele frequencies in gnomAD with the risk allele frequency (RAF) reported in the catalog, matching by ancestry. When the RAF and risk allele were absent from the catalog or ambiguous, we labeled the direction as "unknown." Following association testing, we check the allele direction against the results of our five test cohorts. For associations with an unknown direction, if two or more test cohorts replicated the association at $p < 0.05$ with the same direction of effect, or if a single cohort replicated an association with $p < 0.01$, we set the PGRM direction accordingly. Associations without a clear risk allele were labeled with a "?" and treated as "alternate" in subsequent calculations. Each row was annotated with the source of information on the allele direction (e.g., risk allele, RAF, one or more test cohorts, or unknown).

Annotating genetic ancestry

We annotated study cohort ancestry by using the INITIAL SAMPLE SIZE and REPLICATION SAMPLE SIZE in the "all ancestry data" file as well as the P-VALUE (TEXT) column. We used the 1000 Genomes superpopulations ancestry groupings—African (AFR), East Asian (EAS), European (EUR), admixed American (AMR), South Asian (SAS)—as well as "multiple" (for cohorts that included >1 genetic ancestry superpopulation) and "other" (including individuals from founder populations or genetic ancestries not covered in the superpopulations). We excluded catalog associations that were based on multiple ancestries or included founder populations. The number of subjects in the original GWAS were calculated from the INITIAL SAMPLE SIZE and REPLICATION SAMPLE SIZE columns. These counts were also ancestry specific.

Consolidating the PGRM

Each association in the PGRM is unique by phenotype, SNP, and ancestry. When associations were reported multiple times in the catalog, only the association with the lowest p value was included in the consolidated PGRM. The number of times an association appeared in the catalog was stored. The full PGRM is available in Table S2.

Test cohort datasets

All replication experiments were conducted on summary statistics drawn from four biobanks: BioVU, MGI, UKB, and BBJ. BioVU was divided into two test cohorts: one of European genetic ancestry ($BioVU_{Eur}$) and another for African genetic ancestry ($BioVU_{Afr}$). These datasets comprised summary statistics for SNP/phenotype associations present in the PGRM, including betas, standard errors, and p values. The MGI and UKB summary statistics were drawn from existing datasets described in previous publications.[29,30] The BBJ associations were downloaded from http://pheweb.jp and are described by Sakaue et al.[9] For the BioVU cohorts, we generated association statistics by using the run_PGRM_assoc() function from the pgrm R package. The use of the BioVU data was approved by Vanderbilt's institutional review board; because of the retrospective design of the specific study and the use of deidentified data, the board did not require additional informed patient consent. The BioVU biobank has also been described in prior publications.[31] A description of each cohort, including phenotype definitions, genotyping platform, and models, can be found in Table S3.

The BioVU, MGI, and UKB cohorts defined phenotypes via phecodes (version 1.2; https://phewascatalog.org). The BBJ dataset comprised GWAS results for 229 phenotypes, 42 of which are based on ICD-10 codes. We manually identified 59 phenotypes in the BBJ that were exact matches to phecodes in the PGRM (Table S4).

We identified all associations that included subjects from BioVU (or eMERGE), MGI, or UKB by systematically searching for the biobank names in the source manuscripts. We used this annotation in subsequent replication measures to prevent testing for self-replication (i.e., replicating an association from UKB with the UKB cohort) (Table S5). A complete set of association results from all test cohorts with annotations from the PGRM can be found in Table S6.

Annotating test cohorts with the PGRM

We annotated the summary statistics of each test dataset by using the annotate_results() function from the pgrm R package. This function merges association results from a test cohort with the PGRM, filtering by the specified ancestry, and annotates the result set with the following information.

(1) Information from the original GWAS, including the accession number, summary statistics, and risk allele direction.
(2) Power for each association based on the number of affected and unaffected individuals from the test cohort, the ancestry-matched allele frequency from gnomAD, and the lower confidence interval from the GWAS catalog. Power is calculated with the genpwr.calc function in the genpwr R package with an alpha = 0.05. We used the lower confidence interval instead of the point estimate to compensate for the "winner's curse."
(3) Replication Boolean value, set to 1 if the test cohort replicated the association. Replication is defined as having $p < 0.05$ and OR in the same direction as reported in the GWAS catalog.
(4) The comparison of the confidence intervals (CIs) from the original study and test cohort. Associations are labeled "overlap" if the CIs from catalog and test association are overlapping, "test_cohort_greater" if the lower CI from the test cohort is higher than the upper CI of the catalog cohort, and "PGRM_greater" if the lower CI from the PGRM is higher than the upper CI of the test cohort.

Calculating replication measures in test cohorts

We calculated overall replication rate (RR_{All}) and powered replication rate (RR_{Power}) by applying the get_RR() function to the annotated test cohorts. We calculated the actual:expected ratio (AER) by applying the get_AER() function to each test cohort. The AER is defined as the total number of replicated associations divided by the sum of the power over all associations, a measure used by Palmer and Pe'er.[32]

Replication experiments

We conducted replication experiments by creating new datasets with modified phenotype files by using the $BioVU_{EUR}$ test cohort. We compared these modified datasets to the dataset generated on

the full $BioVU_{EUR}$ cohort—the "benchmark dataset"—to assess the effects of various phenotype and cohort definitions on replication measures. The following list describes the way we modified the $BioVU_{EUR}$ cohort to generate the test datasets.

Randomized phenotypes

We created randomized cohorts by using the $BioVU_{EUR}$ cohort by randomly shuffling a proportion of the individuals in the phenotype and covariate file but not the genotype file. Shuffling was accomplished with the sample() function in the R base package, without replacement. We created nine additional cohorts, randomizing from 10% to 100%, at 10% intervals.

No exclude ranges

We generated a phenotype file without exclude ranges. The benchmark analysis used exclude ranges.

Inpatient only

We generated a phenotype file by using only ICD codes from the inpatient context. The benchmark analysis used ICD codes from all clinical context (inpatient, outpatient, emergency).

Variable minimum code counts

We defined seven additional phenotype files by using different minimum code counts (MCCs). The MCC is the number of unique dates a phecode occurs in an individual's record for them to be classified as affected individuals (i.e., cases). The benchmark analysis used MCC = 2, and we analyzed phenotype files defined with MCC = 1 and 3–8.

Cohort size

We created randomly selected sub-cohorts from $BioVU_{EUR}$, producing cohorts that were 75%, 50%, and 25% of the size of the original cohort.

Annotated summary statistics were generated for each new dataset and annotated with the run_PGRM_assoc() and annotate_results() functions from the pgrm package. We compared the results of the test datasets against the benchmark $BioVU_{EUR}$ results by using the compare_results_sets() function in the pgrm R package. This function assesses the difference between replication measures of two result sets. A Fisher's exact test compares the results of a benchmark and modified cohort for each measure (RR_{All}, RR_{Power}, and % power).

Cross-ancestry replication experiment

Using the $BioVU_{EUR}$ cohort, we attempted to replicate all associations in the catalog, regardless of ancestry, for phenotypes with at least 100 affected individuals. To do so, we used the annotate_results() function for all five ancestry groups, and we computed the RR_{Power} and AER for each by using the get_RR() and get_AER functions. We also computed the replication measures for associations that were reported in the catalog for a single ancestry versus multiple ancestries. Finally, we compared the likelihood of replication for associations that were discovered in European cohorts alone versus those that were discovered in multiple ancestry cohorts by using a chi squared test.

Assessing the contents of the PGRM with test cohorts

Comparing replications across test cohorts

We assessed the overlap of successful replications across the three European test cohorts, including only associations that were included in all three cohorts, and identified associations that were not replicated in any of the three cohorts.

Comparing ORs of initial versus replication studies

We compared the ORs in the PGRM with those generated in our test cohorts by using a paired t test. Only associations that replicated in the test cohorts were included in the analysis.

Factors of replication

We used a logistic regression model to assess the factors associated with replication across all five test cohorts, including the following independent variables in the model: risk allele frequency, number of affected individuals from the test cohort, lower CI from the catalog, the date of the publication, the number of subjects of the original study, the number of times the association occurred in the catalog, the phenotype category, and the ancestry.

Results

Creating the PGRM

We curated associations for the PGRM that are compatible with population-scale biobanks, specifically those based on adult populations of both sexes that are not explicitly selected for specific diseases or traits (Figure 1). At the time of download (2022-01-04), the GWAS catalog included 320,477 rows, each specifying an association for a phenotype and genotype. To ensure compatibility with phecodes, we excluded associations based on continuous traits (n = 228,853). We also excluded those with missing statistical information in the catalog (n = 63,324); those with a reported $p > 5 \times 10^{-8}$ (n = 11,805); and associations based on genetic variants that were rare (minor allele frequency [MAF] < 1%) or non-standard genetic variants (e.g., haplotypes) (n = 1,530). The remaining 14,897 associations were annotated with 655 EFO terms—the terminology used by the GWAS catalog to represent diseases and traits—162 of which had a corresponding phecode (version 1.2) that exactly matched the disease specified in the GWAS catalog. Associations that did not have a matching phecode were excluded (n = 3,911). We also excluded 2,252 associations that were based on modified phenotypes (e.g., ER-breast cancer; n = 1632) or specialized cohorts (e.g., breast cancer in BRCA1/2 carriers; n = 628) or used non-standard statistical models (n = 207).

Due to differences in allele frequency and linkage disequilibrium (LD) patterns between populations, the majority of GWAS findings are ancestry specific.[33] To facilitate ancestry-match replication studies, we excluded 1,737 associations on the basis of multi-ancestry or founder populations. European ancestry was over-represented in multi-ancestry cohorts: 86% of associations had a majority of European ancestry subjects. We annotated PGRM associations according to genetic ancestry of the initial study, including African (AFR; n = 49), East Asian (EAS; n = 1,215); European (EUR; n = 4,568); Latino/admixed American (AMR; n = 30); and South Asian (SAS; n = 17). 78% of the associations in the PGRM were based on EUR cohorts.

In total, the PGRM (version 0.0.1) comprises 5,879 unique phenotype-genotype associations drawn from 523 independent GWAS publications. These associations capture a wide array of diseases, including 149 unique phecodes from 13 disease categories (Figure S1). The PGRM includes summary statistics from the original study

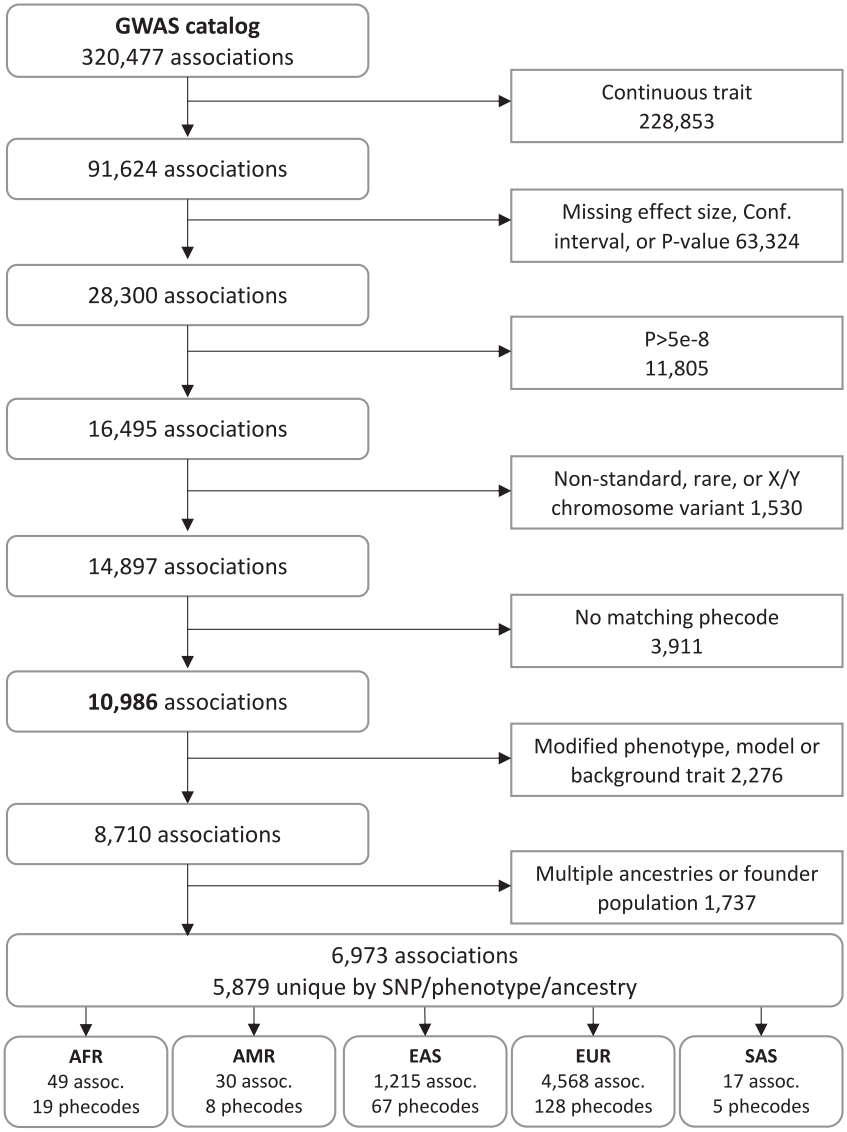

Figure 1. Flow diagram of the creation of the PGRM, beginning with the entire GWAS catalog at the top, to the associations included in the PGRM at the bottom

Rounded boxes on the right show the count of phenotype-genotype associations at each stage, and squares on the left show the number of associations filtered out. The bottom row of the figure shows the number of unique SNP/phenotype associations and unique phecodes included in each subset. Ancestry abbreviations are as follows: AFR, African; AMR, Latino/admixed American; EAS, East Asian; EUR, European; SAS, South Asian.

(ORs, 95% CIs, and p values), the PubMed IDs and publication dates, risk alleles, and variant identifiers for builds hg19 and hg38.

Developing replication measures for QC and comparative studies

To address the lack of QC tools designed for biobank data, we developed replication measures that can be applied to any EHR-linked biobank. We considered an association replicated if it had $p < 0.05$ and a direction of effect consistent with the original study. Power was defined as having power $\geq$ 80% with the lower 95% CI from the catalog to account for winner's curse.[34] The overall replication rate (RR_{All}) and powered replication rate (RR_{Power}) describe the proportion of PGRM associations replicated by a test cohort (see Box 1 for terms and definitions). A third metric, the actual:expected ratio (AER), is a similar measure that includes all PGRM associations, regardless of power. This metric may be more suitable for smaller association sets within the PGRM, as it is found in under-represented ancestries.

Applying replication measures to test cohorts

Replication measures were computed on summary statistics from five test cohorts as described in Table 1. The overall replication rate (RR_{All}) across test cohorts ranged from 37%–61%. The replication measures that took power into account were more consistent across cohorts: the RR_{Power} ranged from 76%–85% and the AER ranged from 0.78–0.94. The AER utilized an average of 3.5 times more associations than the RR_{Power} (Tables 2 and S6). These replication measures were computed after excluding 3,180 associations in the PGRM that were discovered with one or more of the test cohorts. For these non-independent associations, the overall replication rate (RR_{All}) was higher than the RR_{All} of the independent associations (75.1% versus 43.6%), as expected. These non-independent associations were excluded from replication analyses with the test cohorts; over half of the associations derived from BBJ were excluded in this step, reflecting the outsized role this biobank plays in GWASs for the East Asian population.

PGRM replication experiments

We conducted a series of replication experiments to demonstrate applications of the PGRM. In these analyses, we compare the results from the $BioVU_{Eur}$ study (i.e., the "reference cohort") to results generated with the same dataset after some modification.

Detecting data corruption

We hypothesized that corrupted data would negatively impact a cohort's replication rate, resulting in a lower RR_{Power} and AER. We found that randomizing subjects at 10% increments (starting with a cohort at 0% randomization and ending with a fully randomized cohort) decreased the

Box 1. Replication measures and definitions

Replication (r): an association with a p value < 0.05 and OR in the same direction as the original study.

Powered (t_p): an association with power $\geq 80\%$ ($\alpha = 0.05$).

Overall replication rate (RR_{All}): number of replicated associations, r, divided by total number of associations tested, t.

$$RR_{All} = \frac{r}{t}$$

Power replication rate ($RR_{Powered}$): number of replicated powered associations, r_p, divided by total number of powered associations tested, t_p.

$$RR_{Powered} = \frac{r_p}{t_p}$$

Percent powered: number of powered associations, t_p, divided by the number of associations tested, t.

$$\%\ Powered = \frac{t_p}{t}$$

Actual over expected ratio (AER): number of replicated associations, r, divided by the sum of the power estimate.

$$AER = \frac{r}{\sum_{i=1}^{n} Power_i}$$

replication rate monotonically. The RR_{All} and RR_{Power} between 10% increments was statistically significant for all pairs tested. A fully randomized cohort yielded an RR_{Power} of 1.6%, consistent with chance, indicating that this measure is sensitive to data corruption. (Figure 2A; Table S7).

Testing utility of exclude ranges

Phenome-wide analyses such as PheWAS typically exclude unaffected individuals (i.e., controls) with similar conditions to the target phenotype (e.g., the phecode for "migraine" excludes unaffected individuals with "other headache syndromes"). Theoretically, excluding unaffected individuals with similar conditions to the target phenotype should reduce the number of misclassified unaffected individuals, but its effect on replication has never been studied systematically. We hypothesized that replication would be more robust when using those exclude ranges compared than when not using them. We found that replication measures were nominally higher for summary statistics generated with exclude ranges than those without ($RR_{Power} = 76.3\%$ versus 75.0%; $AER = 0.81$ versus 0.79), but these differences were not statistically significant. (Figure 2B; Table S8A).

Assessing the effect of missing phenotype data

Some biobanks include only a subset of ICD codes. For example, until recently the UK biobank included only inpatient ICD codes. However, the effect of missing outpatient codes in a biobank cohort has never formally been assessed. We tested the hypothesis that phenotypes defined with inpatient ICD codes alone and excluding codes from the outpatient context would decrease both the power and replication rate of a cohort. Using only inpatient codes did significantly reduce the number of powered associations, from 853 to 377, and significantly decreased the RR_{All} (29.5% versus 41.4%; $p = 4.0 \times 10^{-22}$), but the RR_{Power} was not significantly different (Table S8B).

Comparing minimum code count (MCC) thresholds

The PheWAS R package defines affected individuals as those with at least two instances of a phecode or a minimum code count (MCC) of 2. Prior studies have found an MCC of 2 maximized phenotype accuracy by balancing

Table 1. Description of biobank test cohorts

Test cohort	Source	Genetic ancestry	Cohort size	PGRM associations tested	Unique phenotypes tested
$BioVU_{EUR}$	BioVU	EUR	62,777	3,268	106
MGI	Michigan Genomics Initiative	EUR	51,393	4,117	109
UKB	UK Biobank	EUR	407,202	2,238	81
$BioVU_{AFR}$	BioVU	AFR	12,142	31	14
BBJ	BioBank Japan	EAS	178,726	384	26

Table 2. Replication measures for biobank test cohorts

Test cohort	% power	RR_{All}	RR_{Power}	AER
$BioVU_{EUR}$	26.0%	41.4%	76.3% (651 of 853)	0.81
MGI	22.5%	36.9%	76.1% (706 of 928)	0.79
UKB	38.1%	56.9%	85.2% (727 of 853)	0.94
$BioVU_{AFR}$	45.2%	61.3%	78.6% (11 of 14)	0.94
BBJ	57.0%	53.4%	76.7% (168 of 219)	0.78

Full results can be found in Table S6.

precision and recall, but these studies were based on a limited number of phenotypes.[26] We used the PGRM to assess the effect of different MCC on a phenome-wide scale. Our results showed a tradeoff between power and replication with increasing MCC. The number of powered associations was highest at MCC of 1 (n = 1,126) and lowest at MCC = 8 (n = 458) due to the reduction in affected individuals. The decrease in powered associations was most precipitous from MCC of 1–2, where 273 associations lost power. This decrease was statistically significant from MCC of 1–4. RR_{Power} increased significantly from MCC of 1–2 (67.5%–76.3%) and continued to increase with ascending MCC, though the stepwise differences were small and not statistically significant. MCC of 2 yielded the most replications overall (n = 1,354), suggesting that this threshold strikes a balance between power and phenotype accuracy (Figures 2C and 2D; Table S9).

Testing effect of cohort size on replication rate

We found that replication rates were influenced by cohort size, most likely because of a thresholding effect of power calculations. The full $BioVU_{EUR}$ (n = 62,777) yields a RR_{Power} of 76.3%. The RR_{Power} was 74.7% when a random 25% was excluded from the cohort, 69.8% when half the cohort was excluded, and 65.8% when 75% was excluded. The AE ratio also decreased (0.81, 0.79, 0.78, and 0.75). (Table S10)

Cross-ancestry replication experiment

We assessed cross-ancestry replicability by using our $BioVU_{EUR}$ cohort. We found a RR_{Power} of 195 of 390 (50.0%) of associations from non-European discovery cohorts (compared with 76.3% in the matched analysis, reported above). Associations that were discovered in both European and non-European cohorts were more likely to replicate than those discovered in European ancestry cohorts alone ($RR_{Powered}$ of 95.1% versus 76.3%, respectively; $p = 6.2 \times 10^{-4}$). (Tables S11 and S12)

Assessing replicability with biobank test cohorts

We used our biobank test cohorts to assess the replicability of PGRM associations. First, we looked for associations that were "replication-resistant" (i.e., associations that did not replicate in multiple cohorts). Of the 393 associations that were sufficiently powered in all three European ancestry cohorts, 22 (5.6%) did not replicate in all three cohorts. Replication-resistant associations were not restricted to any one phenotype or phenotype category (see Table S13 for a list of non-replicated associations). 86% of associations (n = 339) replicated in at least two cohorts, and 66% (n = 258) replicated in all three (Figure S2).

Comparing ORs

Prior work has shown that GWAS findings are upwardly biased in terms of reported ORs.[32] Therefore, we hypothesized that ORs in the PGRM would be higher than those derived from the test cohort replications. For the 4,373 associations that replicated in the test cohorts, we found that ORs were significantly higher in the PGRM than the test cohort (t test $p = 2.1 \times 10^{-18}$; 95% OR = 0.05 [0.039–0.061]) (Figure S3).

Assessing factors of replication

Using multiple logistic regression, we assessed the factors of PGRM associations that were correlated with replication in our test cohorts. As expected, factors related to the strength of the association and power were significantly associated with replication. Higher OR and lower p values in the original study were positively correlated with replication as were the number of affected and unaffected individuals in the test cohort ($p < 0.001$; Table 3).

We found that the number of times an association was reported in the catalog was positively correlated with replication ($p = 7.5 \times 10^{-3}$; OR = 1.15 [1.04–1.27]), as has been previously reported.[16] The size of the original cohort was inversely correlated with replication ($p = 6.0 \times 10^{13}$, OR = 0.96 [0.95–0.97]). The date the association was published was negatively correlated with replication, such that the later the association was first published, the less likely it was to replicate ($p = 1.6 \times 10^{-13}$; OR = 0.94 [0.92–0.95] per year, after 2005). Eight disease categories were significantly less likely to replicate when compared with neoplasm phenotypes (the largest category of association and the category with the highest replication rate), including digestive, endocrine, genitourinary, metabolic, psychiatric, respiratory, and sense organs. Psychiatric disorders had a notably low likelihood of replicating ($p = 8 \times 10^{-50}$; OR = 0.15 [0.12–0.19]).

Discussion

The PGRM is a large set of phenotype-genotype associations that can be used to conduct replication studies at a

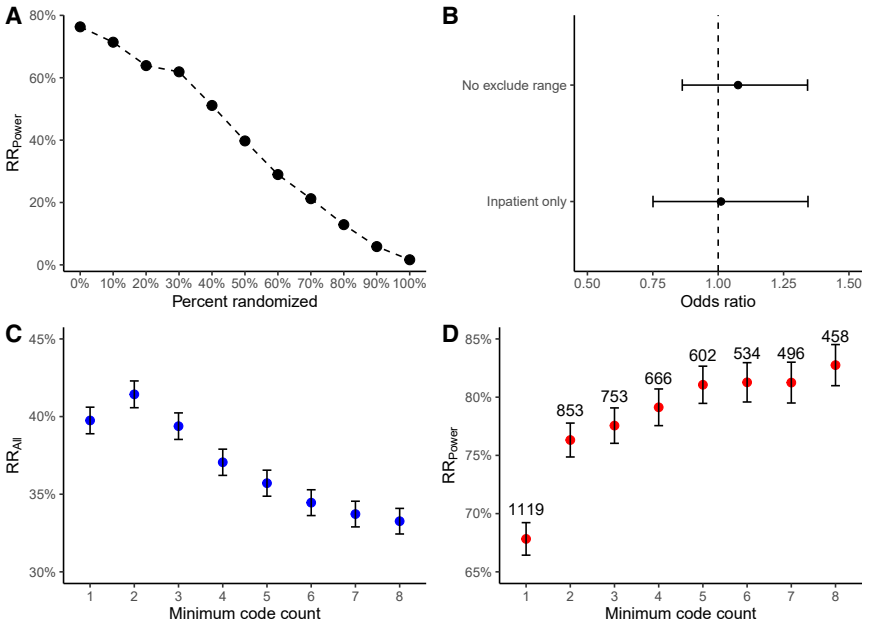

Figure 2. Replication experiments conducted with the PGRM

(A) The replication rate for the BioVU cohort was calculated after permuting subject IDs in the genotype file. The RR_{Power}, shown as black circles, decreased significantly with every additional 10% of subjects that were randomized.

(B) No significant difference was detected for the odds of replicating in an analysis that did not use exclude ranges (top line) or when only inpatient codes were included (bottom line) when compared to a reference dataset.

(C) The RR_{All} for varying minimum code count (MCC) thresholds are shown as blue dots. The maximum RR_{All} was observed with MCC = 2.

(D) The RR_{Power} for varying MCC, shown as red dots. Above the RR_{Power}, the total number of associations in each analysis are labeled. The maximum RR_{Power} was observed for MCC = 8; however, this analysis only included 485 associations. The number of powered associations decreased with increasing MCC.

phenome-wide scale. The use of phecodes in the PGRM ensures interoperability with international ICD standards and a familiar context for researchers who work with EHR-linked biobanks. Each PGRM association includes information about the ancestry of the original cohort and summary statistics to facilitate power calculations. In total, the PGRM includes 5,879 replication candidates (unique by SNP, phenotype, and ancestry) drawn from 523 GWASs and spanning 149 distinct diseases across the medical phenome.

Given the high replicability of GWAS findings, we anticipated that PGRM associations would robustly replicate in our five mature biobank cohorts. Indeed, the overall phenome-wide replication rates were higher than previously reported, most likely in part because of the larger cohorts tested in this study.[16] The majority (86%) of phenotypes replicated at least once. We also found replication rates were comparable across tested ancestry groups. The powered replication rate was strikingly consistent across test cohorts, averaging 79% across all cohorts with a range from 76%–85%.

Through a series of demonstration studies, we showed how the PGRM can be used to assess data quality and analytical assumptions. We found that replication measures are sensitive to random error, indicating that the PGRM may be used to detect problems such as inadvertent shuffling of subject identifiers that may occur in the transfer and processing of large datasets. We also showed how the PGRM can be used to test the effect of different analytical parameters used in PheWAS, such as exclude ranges and minimum code count threshold, and assess the effect of missing phenotype data.

The PGRM offers a way for researchers to assess phenome-wide parameters in their own datasets. This is important because the optimal settings for a phenome-wide analysis will differ depending on the goals of the analysis and the characteristics of the underlying dataset. Some analyses might benefit from enhanced precision and others from increased power. Furthermore, biobank cohorts (and sub-cohorts therein) differ in terms of longitudinality and density, which means that the effect of various parameters such as MCC will most likely differ as well. Therefore, it is likely that the optimal parameters for an analysis are study dependent. While replication measures do not tell the whole story, the PGRM is at least one way of generating empirical evidence for decisions analysts must make when modeling phenome-wide data.

When interpreting PGRM measures, investigators should be aware that there are multiple reasons beyond data quality for why a test cohort may yield a low replication rate. According to best practice guidelines, a replication study should be conducted with the "same or very similar" phenotype and a "similar" cohort.[25] But there is much complexity to unpack in this seemingly simple statement. How might we ensure that two phenotypes are indeed the *same* when diagnostic criteria for disease may differ over time and place and, more generally, the notion of similarity depends on point of view, context, and purpose?[35] The question of what makes cohorts *similar* is at least as vexing. Cohorts may differ in terms of age structure, presence of co-morbidities, and environmental exposures, which in turn may influence replicability.[36] In an effort align the original and replication studies, we excluded from the PGRM associations for modified phenotypes (e.g., ER-positive breast cancer) and cohorts with background traits (e.g., non-obese children). However, because there is no consistent framework used to specify these attributes, important details may not be present in the GWAS catalog. Moreover, the effect of background traits for most phenotype-genotype associations is poorly

Table 3. Factors associated the replication of association from the GWAS catalog

Original association attributes	Odds ratio	95% CI	p value
Effect size[a]	1.24	1.18–1.31	9.0×10^{-17}
p value[b]	1.07	1.06–1.08	1.1×10^{-92}
Risk allele frequency	1.16	0.97–1.39	0.103
Cohort size[c]	0.96	0.95–0.97	6.0×10^{-13}
GWAS publication date[d]	0.94	0.92–0.95	1.6×10^{-13}
Number of times in catalog	1.15	1.04–1.27	7.5×10^{-3}
Test cohort attributes			
Cases[e]	1.16	1.14–1.18	1.6×10^{-60}
Controls[f]	1.02	1.02–1.02	6.4×10^{24}
Category			
Neoplasms	reference	–	–
Circulatory	1.01	0.83–1.22	0.91
Dermatologic	0.60	0.49–0.74	1.4×10^{-6}
Digestive	0.67	0.57–0.78	2.1×10^{-7}
Endocrine	0.91	0.74–1.13	0.40
Genitourinary	0.86	0.59–1.26	0.44
Infectious disease	0.42	0.22–0.80	8.0×10^{-3}
Metabolic/heme	1.24	0.73–2.09	0.42
Musculoskeletal	0.43	0.33–0.57	6.5×10^{-10}
Neurological	0.61	0.50–0.73	1.9×10^{-7}
Psychiatric	0.15	0.12–0.19	8.4×10^{-50}
Respiratory	0.42	0.34–0.51	4.5×10^{-17}
Sense organs	0.50	0.39–0.63	5.1×10^{-9}

[a]Effect size = log(lower 95% confidence interval) reported in the catalog by 0.1 increase.
[b]−log10(p value).
[c]Number of subjects in original study, by 10,000 subjects.
[d]In years.
[e]Cases in test cohort; per 1,000 cases.
[f]Controls in test cohort; per 10,000 controls.

understood. Indeed, non-replication may be instrumental in revealing the scaffolding that supports an observed association.[37] Further replication experiments might help further specify the best practice guidelines by determining which aspects of phenotypes and cohorts are the most important to align for replication.

We used the PGRM to study the GWAS catalog itself, identifying factors that influence the replicability of its findings. Unsurprisingly, we found factors relating to power (i.e., reported effect size, size of the replication cohort) were positive correlated with replicability. We also observed less intuitive associations. First, replication was inversely correlated with the size of the discovery cohort. This surprising finding may be related to what statistician Xiao-Li Meng calls the "big data paradox," whereby larger datasets are less likely to yield confidence intervals that encompass the true estimate.[38,39] The big data paradox is not caused by the size of the dataset per se but rather the tradeoff that is often made between data quantity and quality. Indeed, a study of Alzheimer disease found that larger cohort studies tend to use noisier phenotypes, which meaningfully altered phenotype-genotype associations.[40] Another study showed that population controls, which are commonly used in large biobanks and consortium studies, can produce spurious associations through exposure-linked selection bias.[41] Questions remain regarding the relevancy of the big data paradox in genetic research, but our finding suggests that, while the GWAS paradigm nudges researchers toward to quantity seeking ever larger cohorts for discovery work, an accompanying focus on data quality would be beneficial. Second, associations from more recent publications were less likely to replicate, independent of power and the size of the discovery cohort. The observed decrease in replicability over time may be related to the increasing reliance on non-representative cohorts in GWASs or the recent concerns that GWAS has become less transparent and methodologically rigorous over time; such hypotheses may be the subject of future replication studies.[42,43]

While replication occurred across all 13 disease categories tested, not all disease categories were equally likely to replicate. Psychiatric phenotypes were an outlier in this analysis, with a 0.15 odds of replicating relative to neoplasm phenotypes. This strikingly low replication rate, which was observed across all test cohorts, may be in part attributable to challenges capturing psychiatric phenotypes at scale with EHR data.[44,45] Conversely, highly replicable GWAS results could suggest high standardization regarding capture of disease phenotypes (e.g., neoplasms) and/or those phenotypes with strong links to causal alleles.[46]

The PGRM is not without limitations. First, non-European ancestry associations are under-represented in the PGRM, a reflection of the European ancestry bias of GWASs.[47] While the GWAS catalog includes associations based on multi-ancestry cohorts, the majority of these associations (86%) were based on predominantly European-ancestry cohorts and were not included in the PGRM. We hope that calls to increase diversity of the populations engaged in genetic research will lead to a more complete picture of genetic associations in under-studied ancestries so that the PGRM can expand to include more diverse ancestry associations.[48,49] While ancestry matching is necessary for computing accurate power calculations, our cross-ancestry analysis suggests that the PGRM is useful even when ancestry matching is not possible. The cross-ancestry replication rate was 50%—significantly lower than that of the matched-ancestry analysis, but far higher than chance expectations. Moreover, the PGRM may be useful for identifying cross-ancestry associations, which prior research suggests are more likely to be causal.[50] A full vetting of the cross-ancestry replicability is beyond the scope of this paper, but these results shows that the PGRM can be useful in exploring this important topic further.[51]

Second, the PGRM is limited to phecode-based phenotypes; continuous traits, which make up the majority of the GWAS catalog, are not currently included. Including continuous traits would require mapping these traits to a standardized vocabulary that can be translated across biobanks. While standards such as LOINC have been developed, there are challenges to applying these seamlessly to real world data.[52] However, future iterations of the PGRM may seek to include these traits in the map. The PGRM could also evolve to use data sources beyond ICD codes, such as survey responses. Each addition will require a new map, like the one we created for BBJ phenotypes in this analysis, but existing features will still be applicable.

Third, while the PGRM provides a means for detecting random error and testing phenotype definitions, it is not suitable for making direct comparisons regarding data quality across biobank cohorts. This limitation relates not only to the multiple and complex causes of non-replication but also the influence of cohort size on replication measures. We observed lower replication rates in a random subset of the BioVU cohort, most likely because of a thresholding effect in power calculations (large cohorts are more likely to have associations that exceed the number of affected individuals needed to reach maximum power). Thus, researchers who use the PGRM to assess data quality should be aware that smaller cohorts may produce lower replication rates than were found in this study. Furthermore, studies using the PGRM to compare analytic methods across two datasets should use cohorts of the same size.

Replication is a powerful way to assess the data integrity of large, complex biobank datasets and to better understand analytical assumptions used in modeling phenome-wide data. To facilitate the use of the PGRM, we created a publicly available R package that allows the PGRM to be incorporated into existing biobank QC pipelines or used to conduct replication experiments. We hope that the development of this resource can help maintain the high standards of replicability in the biobank era.

Data and code availability

The PGRM is available as an R package at https://github.com/PheWAS/pgrm. The package includes functions to annotate datasets with the PGRM and calculate replication measures. The R package also includes all summary statistics from test cohorts to facilitate comparative studies. The PGRM and summary statistics from the five test cohorts used in this paper are also available in Tables S1 and S6, respectively.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.07.012.

Acknowledgments

This work was supported by the National Library of Medicine grant R01-LM010685 and National Center for Advancing Translational Sciences grant 5UL1TR002243-03. The authors acknowledge the Michigan Genomics Initiative participants; Precision Health at the University of Michigan; the University of Michigan Medical School Central Biorepository; the University of Michigan Advanced Genomics Core for providing data and specimen storage, management, processing, and distribution services; and the Center for Statistical Genetics in the Department of Biostatistics at the School of Public Health for genotype data curation, imputation, and management in support of the research reported in this publication.

Declaration of interests

Vanderbilt University Medical Center licensed PheWAS on Vanderbilt's DNA biobank to Nashville Biosciences. L.B. and J.C.D. receive a portion of those royalty payments. L.B. and R.B.A. are advisors to the UK Biobank. R.B.A. is an advisor to All of Us, the Swiss Personalized Health Network, and the Danish National Genome Center. R.B.A. is an advisor and stockholder of Personalis, as well as a stockholder of 23andme, and is a paid advisor to Myome, Invitae, and BridgeBio.

Received: April 19, 2023
Accepted: July 27, 2023
Published: August 21, 2023

References

1. Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. Nat. Rev. Methods Primers *1*, 59.
2. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
3. Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. *70*, 214–223.
4. All of Us Research Program Investigators, Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E. (2019). The 'All of Us' Research Program. N. Engl. J. Med. *381*, 668–676.
5. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44*, 1137–1147.
6. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. *12*, e1001779.
7. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al. (2017). Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. *27*. S2–S8.
8. Bastarache, L., Denny, J.C., and Roden, D.M. (2022). Phenome-Wide Association Studies. JAMA *327*, 75–76.
9. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nat. Genet. *53*, 1415–1424.
10. Neale lab - aUK Biobank. Neale lab. http://www.nealelab.is/uk-biobank.
11. Karczewski, K.J., Solomonson, M., Chao, K.R., Goodrich, J.K., Tiao, G., Lu, W., Riley-Gillis, B.M., Tsai, E.A., Kim, H.I., Zheng, X., et al. (2022). Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genom. *2*, 100168. https://doi.org/10.1016/j.xgen.2022.100168.
12. Zuvich, R.L., Armstrong, L.L., Bielinski, S.J., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., et al. (2011). Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality: Pitfalls of Merging GWAS Data: Lessons Learned. Genet. Epidemiol. *35*, 887–898.
13. Colhoun, H.M., McKeigue, P.M., and Davey Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. Lancet *361*, 865–872.
14. DeLozier, S., Bland, S., McPheeters, M., Wells, Q., Farber-Eger, E., Bejan, C.A., Fabbri, D., Rosenbloom, T., Roden, D., Johnson, K.B., et al. (2021). Phenotyping coronavirus disease 2019 during a global health pandemic: Lessons learned from the characterization of an early cohort. J. Biomed. Inform. *117*, 103777.
15. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
16. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. *31*, 1102–1110.
17. Denaxas, S., Gonzalez-Izquierdo, A., Direk, K., Fitzpatrick, N.K., Fatemifar, G., Banerjee, A., Dobson, R.J.B., Howe, L.J., Kuan, V., Lumbers, R.T., et al. (2019). UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J. Am. Med. Inform. Assoc. *26*, 1545–1559.
18. Zhou, W., Kanai, M., Wu, K.H.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., et al. (2022). Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genom. *2*, 100192.
19. O'Sullivan, J.W., and Ioannidis, J.P.A. (2021). Reproducibility in the UK biobank of genome-wide significant signals discovered in earlier genome-wide association studies. Sci. Rep. *11*, 18625.
20. Marigorta, U.M., Rodríguez, J.A., Gibson, G., and Navarro, A. (2018). Replicability and Prediction: Lessons and Challenges from GWAS. Trends Genet. *34*, 504–517.
21. Huffman, J.E. (2018). Examining the current standards for genetic discovery and replication in the era of mega-biobanks. Nat. Commun. *9*, 5054.
22. Hughey, J.J., Rhoades, S.D., Fu, D.Y., Bastarache, L., Denny, J.C., and Chen, Q. (2019). Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. BMC Genom. *20*, 805.
23. National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Board on Research Data and Information; Division on Engineering and Physical Sciences; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Analytics; Division on Earth and Life Studies; Nuclear and Radiation Studies Board; Division of Behavioral and Social Sciences and Education; Committee on National Statistics; Board on Behavioral, Cognitive, and Sensory Sciences; and Committee on Reproducibility and Replicability in Science (2019). Reproducibility and Replicability in Science (National Academies Press (US)).
24. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*. D1005 D1012.
25. NCI-NHGRI Working Group on Replication in Association Studies, Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., et al. (2007). Replicating genotype-phenotype associations. Nature *447*, 655–660.
26. Bastarache, L. (2021). Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annu. Rev. Biomed. Data Sci. *4*, 1–19.

27. Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., and Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. Bioinformatics *26*, 1112–1118.
28. Sulieman, L., Wu, P., Denny, J., and Bastarache, L. (2019). WikiMedMap: Expanding the Phenotyping Mapping Toolbox Using Wikipedia. Preprint at bioRxiv. https://doi.org/10.1101/727792.
29. Zawistowski, M., Fritsche, L.G., Pandit, A., Vanderwerff, B., Patil, S., Schmidt, E.M., VandeHaar, P., Willer, C.J., Brummett, C.M., Kheterpal, S., et al. (2023). The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. Cell Genom. *3*, 100257.
30. Gagliano Taliun, S.A., VandeHaar, P., Boughton, A.P., Welch, R.P., Taliun, D., Schmidt, E.M., Zhou, W., Nielsen, J.B., Willer, C.J., Lee, S., et al. (2020). Exploring and visualizing large-scale genetic associations by using PheWeb. Nat. Genet. *52*, 550–552.
31. Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys, D.R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin. Pharmacol. Ther. *84*, 362–369.
32. Palmer, C., and Pe'er, I. (2017). Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. PLoS Genet. *13*, e1006916.
33. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.
34. Xiao, R., and Boehnke, M. (2011). Quantifying and correcting for the winner's curse in quantitative-trait association studies. Genet. Epidemiol. *35*, 133–138.
35. Chesterman, A. (2004). Where Is Similarity? In Similarity and Difference in Translation (Guaraldi), pp. 63–75.
36. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. Elife *9*, e48376.
37. Gruber, P.J. (2019). Genetic association studies: Is non-replication failure or progress? J. Thorac. Cardiovasc. Surg. *157*. e399–e400.
38. Meng, X.-L. (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. Ann. Appl. Stat. *12*, 685–726.
39. Msaouel, P. (2022). The Big Data Paradox in Clinical Practice. Cancer Invest. *40*, 567–576.
40. Escott-Price, V., and Hardy, J. (2022). Genome-wide association studies for Alzheimer's disease: bigger is not always better. Brain Commun. *4*, fcac125.
41. Duchen, D., Vergara, C., Thio, C.L., Kundu, P., Chatterjee, N., Thomas, D.L., Wojcik, G.L., and Duggal, P. (2023). Pathogen exposure misclassification can bias association signals in GWAS of infectious diseases when using population-based common control subjects. Am. J. Hum. Genet. *110*, 336–348.
42. Burt, C., and Munafò, M. (2021). Has GWAS lost its status as a paragon of open science? PLoS Biol. *19*, e3001242.
43. Munafò, M.R., Tilling, K., Taylor, A.E., Evans, D.M., and Davey Smith, G. (2018). Collider scope: when selection bias can substantially influence observed associations. Int. J. Epidemiol. *47*, 226–235.
44. Curtis, D. (2021). Analysis of 50,000 exome-sequenced UK Biobank subjects fails to identify genes influencing probability of developing a mood disorder resulting in psychiatric referral. J. Affect. Disord. *281*, 216–219.
45. Li, Z., Kormilitzin, A., Fernandes, M., Vaci, N., Liu, Q., Newby, D., Goodday, S., Smith, T., Nevado-Holgado, A.J., and Winchester, L. (2022). Validation of UK Biobank data for mental health outcomes: A pilot study using secondary care electronic health records. Int. J. Med. Inform. *160*, 104704.
46. Waters, K.M., Le Marchand, L., Kolonel, L.N., Monroe, K.R., Stram, D.O., Henderson, B.E., and Haiman, C.A. (2009). Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. Cancer Epidemiol. Biomarkers Prev. *18*, 1285–1289.
47. Morales, J., Welter, D., Bowler, E.H., Cerezo, M., Harris, L.W., McMahon, A.C., Hall, P., Junkins, H.A., Milano, A., Hastings, E., et al. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. *19*, 21.
48. Manolio, T.A. (2019). Using the Data We Have: Improving Diversity in Genomic Research. Am. J. Hum. Genet. *105*, 233–236.
49. Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A.R., and Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. Nat. Med. *28*, 243–250.
50. Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.Y., Popejoy, A.B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R.J., Brick, L., et al. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell *179*, 589–603.
51. Marigorta, U.M., and Navarro, A. (2013). High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. PLoS Genet. *9*, e1003566.
52. Goldstein, J.A., Weinstock, J.S., Bastarache, L.A., Larach, D.B., Fritsche, L.G., Schmidt, E.M., Brummett, C.M., Kheterpal, S., Abecasis, G.R., Denny, J.C., and Zawistowski, M. (2020). LabWAS: Novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. PLoS Genet. *16*, e1009077.