MVP gwPheWAS Basic Information

These summary statistics are the result of the VA Million Veteran Program (MVP) gwPheWAS effort (collaboration between the US Departments of Veterans Affairs and Energy). For full details, please see our publication in Science and our methods preprint on BioRxiv. For additional questions please contact the corresponding authors of these publications or email MVP_gwPheWAS@va.gov.

This analysis was run twice in order to respond to evolving recommendations and guidelines around determination of genetic ancestry, race, and ethnicity. We used our previous "harmonized ancestry and race/ethnicity" (HARE) definition initially then followed with genetically inferred ancestry (GIA). Both sets of results are available for download via the dbGaP ftp site, however, the GIA results are the focus of the main manuscript and only these results have been processed into dbGaP's browsable website and viewer. We refined our analysis plan based on lessons learned from the HARE analysis therefore there are slight differences between the analyses, which are listed below. Detailed methods for the HARE analysis are attached as an appendix as they are not included in the Science publication.

	HARE	GIA
Quantitative Traits	Raw and inverse-normalized trait values analyzed	Only Inverse-normalized trait values analyzed
Sex-specific Traits	Not analyzed	Analyzed and restricted to only the specific biological sex affected

The VA considers several conditions to be stigmatizing or sensitive therefore have not been included. These include HIV status, sickle cell disease status, and some outcomes relating to substance use/abuse. As the latter fell within the phecode category of "Mental Disorders", none of the phecodes in this category have been deposited into dbGaP. A future release is anticipated to include phecodes from this category which have been approved for distribution. Additionally, in order to reduce privacy and reidentification concerns, only population-specific analyses with at least 500 cases were uploaded to dbGaP despite a cutoff of 200 cases for inclusion in the meta-analysis.

Please see Data Dictionary for additional information on phecode definitions and case/control/total counts. If there are any discrepancies between files with regards to these counts, the summary statistics results file is correct. This is also helpful for determining which tar file(s) contain your phenotype(s) of interest.

Please see our GitHuB repository for detailed methods information and code (https://github.com/exascale-genomics/SAIGE-GPU/).

Key References

Verma A, Huffman JE, Rodriguez A, Conery M, Lui M, et al. Diversity and Scale: Genetic Architecture of 2,068 Traits in the VA Million Veteran Program. Science. 2024 July 19
doi: 10.1126/science.adj1182

Rodriguez A, Kim Y, et al. Accelerating Genome- and Phenome-Wide Association Studies using GPUs - A case study using data from the Million Veteran Program. bioRxiv [Preprint]. 2024 May 22 doi: 10.1101/2024.05.17.594583. PMID: 38826407

Fang H, et al. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. Am J Hum Genet. 2019 Oct 3;105(4):763-772. doi: 10.1016/j.ajhg.2019.08.012. PMID: 31564439

Methods gwPheWAS [HARE]

Million Veteran Program (MVP)
The VA Million Veteran Program (MVP) is a national cohort launched in 2011 designed to study the contributions of genetics, lifestyle, and military exposures to health and disease among US Veterans[1]. Blood biospecimens were collected for DNA isolation and genotyping, and the biorepository was linked with the VA EHR, which includes diagnosis codes (International Classification of Diseases ninth revision [ICD-9] and tenth revision [ICD-10]), laboratory measures, and detailed survey questionnaires collected at the time of enrollment for all Veterans followed in the healthcare system up to September 2019.

Genotyping, Quality Control, and Imputation
Specimen collection and genotype quality control have been described in detail before[2,3]. In brief, blood specimens were collected at recruitment sites across the country then shipped within 24 hours to the VA Central Biorepository in Boston, MA for processing and storage. Study participants were genotyped using a customized Affymetrix Axiom biobank array (the MVP 1.0 Genotyping Array), containing over 730,000 variants. Duplicate samples were excluded as well as samples with observed heterozygosity greater than the expected heterozygosity, missing genotype call rate greater than 2.5%, or incongruence between sex inferred from genetic information and gender extracted from phenotype data. Probes with high missingness (>20%), those that were monomorphic, or those with a Hardy Weinberg Equilibrium p<1E-06 in both the overall cohort and within one of the 3 major HARE groups (non-Hispanic White, non-Hispanic Black, or Hispanic/Latino). See below for HARE methods.

Population-specific principal components (PCs) were computed using EIGENSOFT v.6[4].

Genetic imputation was performed to a hybrid imputation panel comprised of the African Genome Resources panel (https://imputation.sanger.ac.uk/?about=1#referencepanels) and 1000G p3v5[5] using SHAPEIT4 (v 4.1.3)[6], and Minimac4[7].

Ancestry assignment
The harmonized race/ethnicity and genetic ancestry (HARE) approach, developed by MVP, was used to assign individuals to ancestral groups[8]. This machine learning algorithm leverages information from both the self-reported race/ethnicity data from the MVP Baseline survey and genotype data to categorized Veterans into four mutually exclusive groups: (1) non-Hispanic White (EUR), (2) non-Hispanic Black (AFR), (3) Hispanic or Latino (HIS), or (4) Asian (ASN).

Phenotype Data
EHR-derived clinical outcomes (PheCodes)
The clinical outcome from EHR was defined by phecodes curated by the MCP Data Core[9]. Each phecode represents ICD codes grouped into clinically relevant phenotypes for clinical studies. Using this approach, all ICD codes for all Veterans in MVP were extracted and each assigned a phenotype defined by a phecode. ICD-9 and ICD-10 codes were mapped to 1,876 phecodes, as previously described[10]. For each phecode, participants with ≥2 phecode-mapped ICD-9 or ICD-10 codes were defined as cases, whereas those with no instance of a phecode-mapped ICD-9 or ICD-10 code were defined as controls. Based on our previous simulation studies of ICD EHR data, populations where the phecode comprises < 200 cases or controls were more likely to result in spurious results, and we thus applied this threshold in each of the four HARE-defined ancestry groups.

Laboratory measurements
For quantitative traits, we calculated the minimum, maximum, and mean value across all visits for each participant and analyzed each resulting phenotype. Only quantitative traits with data for more than 1000 individuals within each HARE-defined ancestry group were included in the analyses. The remaining 69 laboratory measurements that passed quality control were normalized using a rank-based inverse-normal transformation. We additionally filtered values are greater than six standard deviations from the mean in order to remove extreme outliers.

Survey Questions
The two surveys (questionnaires) for MVP, as noted previously, were designed to augment data that are contained in the electronic health record of each participant[1]. As with other study activities and all study materials sent to participants, these documents were approved by the VA Central IRB. As participants are enrolled, informed consent and HIPAA authorization forms are scanned by field site staff and sent to the CERC, to be checked for accuracy and completeness, and the data are entered in GenISIS. Conceptually, the MVP Baseline Survey was designed to collect information regarding demographics, family pedigree, health status, lifestyle habits, military experience, medical history, family history of specific illnesses, and physical features. The MVP Lifestyle Survey contains questions from validated instruments in domains selected to provide information on sleep and exercise habits, environmental exposures, dietary habits, and sense of well-being.

Genetic association analyses

Within each HARE-defined ancestry group (AFR, ASN, EUR, HIS), genetic variants were tested for their association with the trait of interest using generalized linear mixed models to account for participant relatedness using a GPU-optimized version of the SAIGE package[11] implemented on the U.S. Department of Energy Summit supercomputer. Directly genotyped variants were used for step 1 of SAIGE. Imputed genetic dosages were used for step 2 of SAIGE. Variants were only included in the GWAS if they had an imputation quality > 0.3 and a minor allele count (MAC) > 20 within the relevant HARE-defined ancestry group. Analyses were adjusted for age, sex, and 10 ancestry-specific genetic principal components.

Post-GWAS quality control

GWAS results were filtered using a custom R script loosely based on EasyQC[12]. Sanity checks were implemented to remove variants with missing values for major summary statistics (effect size, standard error, etc) or with unreasonable values (p-values or allele frequencies with values >1 or <0). Additionally, variants were removed that were monomorphic, poorly imputed (r2<0.3) or very rare (minor allele frequency < 0.0001) in just the subset of individuals included in the GWAS.

Meta-Analysis

Multi-ancestry meta-analysis was performed using the inverse-variance weighted method as implemented in GWAMA[13]. Meta-analysis results then underwent the same quality control procedures as the GWAS results. Imputation quality filters were not implemented however an additional filter was added to exclude variants that were specific to only one HARE-defined ancestry group.

References

1. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol 2016;70:214–23.

2. Klarin D, Damrauer SM, Cho K, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet 2018;50(11):1514–23.

3. Hunter-Zinck H, Shi Y, Li M, et al. Genotyping Array Design and Data Quality Control in the Million Veteran Program. Am J Hum Genet 2020;106(4):535–48.

4. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904–9.

5. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature 2015;526(7571):68–74.

6. Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun 2019;10(1):5436.

7. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012;44(8):955–9.

8. Fang H, Hui Q, Lynch J, et al. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. Am J Hum Genet 2019;105(4):763–72.

9. Song RJ, Ho YL, Schubert P, et al. Phenome-wide association of 1809 phenotypes and COVID-19 disease progression in the Veterans Health Administration Million Veteran Program. PLoS One 2021;16(5):e0251651.

10. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;26(9):1205–10.

11. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet 2018;50(9):1335–41.

12. Winkler TW, Day FR, Croteau-Chonka DC, et al. Quality control and conduct of genome-wide association meta-analyses. Nat Protoc 2014;9(5):1192–212.

13. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 2010;11:288.