nature portfolio

Corresponding author(s): Vassy

Last updated by author(s): 2021-7-23

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collected the data.
Data analysis	PRS summary statistics were obtained from the Polygenic Score (PGS Catalog) and Cancer-PRSWeb. Imputation was performed against 1000 Genome Project phase 3 data using the Michigan Imputation Server with ShapeIT (v2.r790) for phasing and Minimac3 for imputation. Unadjusted raw PRS were calculated using PLINK (version 2.0a). PRS distributions were visualized using the density function in R (v4.0.3). R coefficients were calculated with RStudio (v1.1.383) with R (v4.0.3). Code used to adjust the PRS for population structure will be available for download from GitHub prior to publication.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Individual-level MGBB data are available from https://personalizedmedicine.partners.org/Biobank/Default.aspx; however, there are restrictions on this data which was accessed under IRB protocol for this current study so are not publicly available. The majority of the MGBB genotyped samples are deposited in dbGAP as part of

the eMERGE consortium (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001584.v2.p2). Data from the GenoVA Study trial will be made publicly available after trial completion.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	PRS were developed and validated in all 36,423 Mass General Brigham Biobank participants with available genotype data. We show that this sample size was sufficient to replicated the known PRS-disease phenotype associations.
Data exclusions	No data were excluded from the analysis.
Replication	To determine clinical validity, we replicated previously published PRS-disease associations. The analytic validity of the PRS were confirmed by examining the robustness of the calculated PRS across different genotyping platforms and with comparison to genome sequence data.
Randomization	Models used to develop the PRS adjusted for population structure included the first 4 ancestry-informative principal components.
Blinding	Analysts were blinded to the case/control status of biobank participants when calculating each PRS.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Among 36,423 MGBB participants, mean (SD) age was 58.8 (17.1) years (range 9-106), 19,719 (54.1%) were female, and 5706 (15.7%) were of reported race other than white [30,716 (84.3%) white, 1,807 (5.0%) Black, 786 (2.2%) Asian, and 3,113 (8.5%) of other/unknown race]. Among the first 141 GenoVA Study enrollees aged 50-70, 92 (65%) are of non-white reported race/ethnicity, and 49 (35%) currently identify as women.
Recruitment	All patients of affiliated hospitals and clinics are eligible to participate in the MGB Biobank. VA Boston patients aged 50-70 without known diagnoses of the 6 targets diseases are eligible for enrollment. Both the MGB Biobank and GenoVA Study populations have the potential for healthy volunteer bias. The samples used for genome sequencing derive from 22 clinical samples of patients who have undergone clinical sequencing, but these results should not meaningfully bias the comparison of sequence and genotype data for PRS calculation.
Ethics oversight	Analysis of the genomic and MGB Biobank samples and data has been reviewed and approved by the Mass General Brigham IRB. All individuals with genome sequencing data gave consent for clinical genomic screening and all individual data were de-identified. The GenoVA Study is approved by the VA Boston Healthcare System and Harvard Medical School IRB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	The GenoVA Study: ClinicalTrials.gov Identifier: NCT04331535
Study protocol	https://clinicaltrials.gov/ct2/show/NCT04331535
Data collection	GenoVA Study data collection began July 17, 2020 and is ongoing.
Outcomes	The primary outcome of the study is time-to-diagnosis both of undiagnosed prevalent cases of the 6 target conditions and incident cases during the study period. This composite outcome will only include clinically significant diagnoses, as adjudicated by expert clinical chart review.