nature portfolio

Corresponding author(s): Julian Mutz, Jinchen Li, Bin Li

Last updated by author(s): Aug 31, 2024

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	This research analysed data provided by the UK Biobank via application (No. 103082).
Data analysis	Relevant analyses in this study were conducted using R version 4.2.0 (https://www.r-project.org), PLINK 2.00 alpha (https://www.cog-genomics.org/plink/2.0/), Bcftools (https://samtools.github.io/bcftools/), and LDSC (LD Score) v1.0.1 (https://github.com/bulik/ldsc). No customized code was developed.

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

The data for this research are obtained from the UK Biobank and are publicly available to approved researchers for health-related research (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access). The NMR metabolomic data in the UK Biobank are generated by Nightingale Health and are provided in Category 220 (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=220). The GWAS summary statistics for 325 NMR biomarkers have been deposited in the

NHGRI-EBI GWAS Catalog database with study accession IDs ranging from GCST90445833 – GCST90446157 (Detailed GWAS Catalog assession IDs for each NMR biomarker are provided in the Supplementary Data 21). For example, GWAS summary statistics of acetate has been deposited at http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90445001-GCST90446000/GCST90445833/ . The data for figures in this study are provided in the Source Data file. Source data are provided with this paper. Also, the data that support the findings of this study are available from the corresponding author upon request.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Self-reported sex (UK Biobank Field ID: 31) was used in the analyses. Sex was not specifically considered in the study design as the primary research question was not expected to differ between sexes, making a combined analysis more relevant and straightforward for addressing the aim of the study. The findings from the study were applicable to both sexes.
Reporting on race, ethnicity, or other socially relevant groupings	Samples were not excluded or grouped based on the ethnicity. In the subsection of multivariable Mendelian Randomization analysis, Genome-wide association analyses for NMR biomarkers were conducted among those with the genetic background of white (UK Biobank Field ID: 22006) to avoid potential bias from population stratification.
Population characteristics	The summary statistics of population characteristics (for example, age, ethnicity, sex, alcohol intake frequency, systolic blood pressure, smoking status, body mass index, Townsend deprivation index, frailty index, etc.,) were provided in the Supplementary Note 1.
Recruitment	UK Biobank recruited about 500,000 people aged between 40-69 years in 2006-2010 from across the UK. They have undergone a wide range of physical measures, provided information on their lifestyle and medical history, donated blood, urine and saliva samples for future analysis and agreed to have their health followed up through linkage to their health-related records.
Ethics oversight	UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval. This approval means that researchers do not require separate ethical clearance and can operate under the RTB approval. Details on the ethics and governance framework of the UK Biobank are provided on the website (https://www.ukbiobank.ac.uk/media/0xsbmfmw/egf.pdf). This study has been approved under the UK Biobank application ID 103082.

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

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	Participants recruited from 20 assessment centers in England and Wales (n=234,553) were included in the training dataset, while participants recruited from the two assessment centers in Scotland (n=15,788) were included as an out-of-sample validation dataset, yielding a total sample size of 250,341 participants.
Data exclusions	The exclusion of data primarily pertains to the GWAS section.1. Quality control of samples: A subset of individuals with available whole-genome sequencing data in UK Biobank 200k release was included in the first place. Then, samples failed to pass quality requirements (UKB Field ID: 23093), samples with sex chromosome aneuploidy (UKB Field ID: 22019), samples with discordant genetic sex (UKB Field ID: 22001) and self-reported sex (UKB Field ID: 31) were excluded from further analysis. Finally, samples whose genetic ethnic group belonged to white (UKB Field ID: 22006) were included, resulting in a final sample size of 95,372 individuals for downstream GWAS analysis. 2. Quality control of WGS data: Multiallelic variants were decomposed into biallelic variants using bcftools (v1.15.1). Quality control for SNPs and indels were performed based on following inclusion criteria: (1) alternative alleles with AAscore>0.5; (2) variant sites with the tag "FILTER=PASS"; (3) Hardy-Weinberg P-value>10E-15; (4) genotype missing rate<10%. Further, only common variant sites (MAF>0.1%) were included in GWAS analysis.
Replication	The findings from the training dataset got replicated in the out-of-sample validation dataset. However, external validation cohorts are still warrant given the "Healthy volunteer bias" in the UK Biobank cohort.
Randomization	Not relevant because this is an observational cohort study.
Blinding	Not relevant because this is an observational cohort study.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA