nature portfolio

Corresponding author(s): Jeffrey Barrett

Last updated by author(s): 9/24/2024

# 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<br>*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<br>*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 | Biomarker quantification with Nightingale Health quantification library 2020 |
| Data analysis | All statistical analyses and modeling were performed in R version 4.3.2. Code to reproduce main figures are available from https://github.com/NightingaleHealth/ukb-nightingale-omics-prediction. |

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

For reasons of patient confidentiality and to ensure research is carried out in accordance with the consent it was collected under, data from the three biobanks are available under controlled access. The UK Biobank data are available for approved researchers through the UK Biobank data-access protocol (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access). Data from Estonia Biobank can be accessed through a research application to Institute of Genomics of the University of Tartu (https://genomics.ut.ee/en/content/estonian-biobank). Data from FINRISK and Health 2000 cohorts can be accessed through a research

application to THL Biobank (https://thl.fi/en/web/thl-biobank).
Source data for Figures 1-4 are provided with this paper.

## 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 | We used data from a total of 700,217 individuals (297,151 males, 403,066 females) from three biobanks, UK Biobank (N = 477,078, 216,825 males, 260,253 females), median age 58), Estonian Biobank (N = 190,785, 65,220 males, 125,565 females, median age 43) and Finnish THL Biobank (N = 32,354, 15,106 males, 17,248 females, median age 51).<br><br>Analyses in this paper were all carried out conditional on age and sex. Sex was defined differently in the different cohorts: in UK Biobank, sex at recruitment is taken from the patient's medical record but can subsequently be edited by the participant if they choose. For the Finnish THL biobank, sex was taken from the population registries held by the Digital and Population Data Services Agency. In the Estonian Biobank, sex was extracted from the Estonian National Identity number which is created based on the sex recorded in the Estonian Birth Registry.<br><br>Written informed consent was obtained from all participants. Participants were not offered compensation for participating in this study. |
| Reporting on race, ethnicity, or other socially relevant groupings | No direct adjustment for race, ethnicity, etc. Genetic analyses are adjusted for global principal components. |
| Population characteristics | Details in methods & supplement. |
| Recruitment | UK and Estonian Biobanks likely have healthy volunteer bias. THL Biobanks are intended to be population representative. |
| Ethics oversight | The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee. This research was conducted using the UK Biobank Resource under Application Number 30418. The Estonian Committee on Bioethics and Human Research approved the Estonian Biobank study. Data was accessed with research approval number 1.1-12/2770. The THL biobank cohorts were approved by the Coordinating Ethical Committee of the Helsinki and Uusimaa Hospital District, Finland. Data was accessed with research application number BB2016_86. |

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 | Maximum samples used, no pre-study sample size calculation. Six times bigger than previous equivalent analysis. |
| Data exclusions | Missing metabolomic data, >4 SDs from mean metabolite measurements, health-insurance-fund-only derived diagnoses in EBB |
| Replication | Main findings were replicated in a held out set of UK Biobank, as well as two other large biobanks. |
| Randomization | This was an observational study, and so participants were not randomized to experimental groups. |
| Blinding | For all experiments, data generation and biomarker quantification were carried out blinded to sample phenotype, using randomly assigned pseudonymous sample IDs. |

# 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

*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes

*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication

*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*