nature portfolio

Corresponding author(s): Riccardo E Marioni and Catalina A Vallejos

Last updated by author(s): Feb 10, 2023

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	Data collection was performed by Generation Scotland (GS) and Cooperative Health Research in the Region of Augsburg (KORA). For further information please contact access@generationscotland.org (GS) and kora-studienzentrum@helmholtz-muenchen.de (KORA). Data cannot be publicly disclosed due to them containing information that could compromise participant consent and confidentiality.
Data analysis	All analysis code was written by the authors and is publicly available on GitHub in the following repositories: https://github.com/marioni-group/episcores-diabetes-prediction https://github.com/marioni-group/MethylPipeR https://github.com/marioni-group/MethylPipeR-UI R packages: • BART (version 2.9) • gbm (version 2.1.8) • glmnet (version 4.1-1) • randomForestSRC (version 2.11.0)

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

According to the terms of consent for GS participants, access to data must be reviewed by the GS Access Committee. Applications should be made to access@generationscotland.org.
Applications for access to KORA should be made using the KORA.PASST system (http://epi.helmholtz-muenchen.de/).

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	We used 9,537 participants from GS for whom we had DNA methylation data. For model validation, we used data from 1,451 participants in the KORA S4 study with DNA methylation and incident T2D data available. Sample sizes for training, testing and validation of the statistical models were determined by the cohort sizes
Data exclusions	Exclusions were made based on DNAm quality control and missingness in health data (outlined under the Generation Scotland section in Methods)
Replication	Replication of model performance was performed through validation in an external cohort, KORA.
Randomization	This study used observational data from a population-based cohort
Blinding	This study used observational data from a population-based cohort

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	Generation Scotland: the Scottish Family Health Study (GS) is a large, family-structured, population-based cohort study of > 24,000 individuals from across Scotland. Of these individuals, 9,537 had DNA methylation measures available. Full summary statistics are shown in Table 1 and Supplementary Table 4. KORA is a research platform performing population-based surveys and subsequent follow-ups in the region of Augsburg in Southern Germany. This study used a subsample of the 1,451 participants of the KORA S4 study with DNA methylation and incident T2D data available. Summary statistics are provided in Supplementary Table 2.

Recruitment	Recruitment for Generation Scotland: the Scottish Family Health Study (GS) took place between 2006 and 2011 with a clinical visit where detailed health, cognitive, and lifestyle information was collected along with biological samples (blood, urine, saliva). Recruitment for KORA S4 took place between 1999 and 2001. Each participant completed a health questionnaire, providing details on health status and medication. Blood samples were also taken for assaying of omics data.
Ethics oversight	All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20-ES-0021), providing generic ethical approval for a wide range of uses within medical research. The KORA studies were approved by the Ethics Committee of the Bavarian Medical Association (Bayerische Landesärztekammer; S4: #99186) and were conducted according to the principles expressed in the Declaration of Helsinki. All study participants gave their written informed consent.

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