nature portfolio

Corresponding author(s): Jiang Bian

Last updated by author(s): Sep 15, 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	Data was stored in our UF Health Integrated Data Repository database. Our UF Health IDR data follow Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which is an open community data standard.
Data analysis	python (3.7.12), scikit-learn (1.0.2), xgboost (1.5.0), aif360 (0.5.0), scipy (1.7.3), pandas (1.3.5), shap (0.41.0), tetrad-gui (7.5.0), Cytoscape (3.10.1). Custom code is available upor reasonable request.

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

Data from UF Health IDR can be requested through https://idr.ufhealth.org/research-services/data-request-form/. Since the UF Health data is a HIPAA-limited data set, a data use agreement needs to be established with the UF Health IDR research team. The relevant raw data for each figure is provided in the Source Data file.

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	(i) the abstract reports that our experiments and findings apply to female and male. (ii) the sex of participants was recorded in the health care system. (iii) the overall number was reported in the abstract.
Reporting on race, ethnicity, or other socially relevant groupings	The research classifies the race-ethnicity to non-Hispanic White, non-Hispanic Black, Hispanic, and Others. The researchers used the information from health care system (reported by the patient themselves) to classify the race-ethnicity information.
Population characteristics	This is a population with 10,192 patients who were diagnosed by Type 2 Diabetes. 58% are female, and the mean age is 59 years. Of the cohort, 50% were non-Hispanic White, 39% were non-Hispanic Black, 6% were Hispanic, and 5% were other races/ethnicities.
Recruitment	This is a secondary study that uses existing real-world data from the health care system.
Ethics oversight	University of Florida

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	Sample size was determined based on similar studies in this field.
Data exclusions	The patients who have no ZIP codes are excluded.
Replication	To verify the reproducibility of our findings, bootstrapping experiements are performed and the results are reported in the supplements. All models and model parameters are stored.
Randomization	In the current study, we included patients who were (1) aged 18 and older, (2) had a T2D diagnosis, identified as having at least one inpatient or outpatient T2D diagnosis (using ICD-9 codes 250.x0 or 250.x2, or ICD-10 code E11) and ≥ 1 glucose-lowering drug prescription in (a case finding algorithm previously validated in EHRs with a positive predictive value [PPV] >94%), and (3) had at least one encounter during both baseline period and the follow up year.
Blinding	All samples used in this study are de-identified.

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