nature portfolio

Corresponding author(s): Pui-Yan Kwok

Last updated by author(s): Mar 25, 2025

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 for data collection
Data analysis	Code for genotyping quality control process and statistical analyses is available at our Github (https://github.com/TPMI-Taiwan/).

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

All summary statistics, polygenic risk score (PRS) models, and GWAS results are freely available from the TPMI website (https://pheweb.ibms.sinica.edu.tw/).

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 conducted sex-specific analysis for sex-specific diseases, e.g. prostate cancer for male and female breast cancer for female. Additionally, we included sex as a covariate in our statistical analyses.
Reporting on race, ethnicity, or other socially relevant groupings	The proportion of genetic ancestry was determined using ADMIXTURE, and projected principal component scores with the 1000 Genomes Project as a reference panel were applied to determine individuals' ancestry. As a result, 463,447 participants with high genetic similarity to Han Chinese were included in the main analyses. No other socially relevant groupings variables were included in this study.
Population characteristics	Age was included as a covariate in our analyses. Among the 486,956 participants with both genotype and electronic medical record data in the TPMI cohort, there are 217,595 male participants with an average age of 57.4 ± 17.5 years and 269,361 female participants with an average age of 54.9 ± 17.0 years.
Recruitment	TPMI participants were recruited from 16 partner medical centers, which included 33 affiliated hospitals, beginning in July 2019. The TPMI adopted a strategy of utilizing electronic medical record data from these hospitals in its original form, with personal identifying information removed. Since we are conducting genetic studies, selection bias is unlikely unless genetic markers are associated with participants' willingness to participate.
Ethics oversight	This study was approved by the Institutional Review Boards of Taipei Veterans General Hospital (2020-08-014A), National Taiwan University Hospital (201912110RINC), Tri-Service General Hospital (2-108-05-038), Chang Gung Memorial Hospital (201901731A3), Taipei Medical University Healthcare System (N202001037), Chung Shan Medical University Hospital (CS19035), Taichung Veterans General Hospital (SF19153A), Changhua Christian Hospital (190713), Kaohsiung Medical University Chung-Ho Memorial Hospital (KMUHIRB-SV(II)-20190059), Hualien Tzu Chi Hospital (IRB108-123-A),Far Eastern Memorial Hospital (110073-F), Ditmanson Medical Foundation Chia-Yi Christian Hospital (IRB2021128), Taipei City Hospital (TCHIRB-10912016), Koo Foundation Sun Yat-Sen Cancer Center (20190823A), Cathay General Hospital (CGH-P110041), Fu Jen Catholic University Hospital (FJUH109001) and Academia Sinica (AS-IRB01-18079), Taiwan. Written informed consent was obtained from the subjects in accordance with institutional requirements and the Declaration of Helsinki principles. All collected information was de-identified before statistical data analysis. The analysis with Taiwan Biobank was apporaved by Institutional Review Boards of Academia Sinica (AS-IRB-BM-19014), and the NHIRD analysis with Health and Welfare Data Science Center (HWDC) was approved by Institutional Review Boards of Academia Sinica (AS-IRB-BM-23056). This research has been conducted using the UK Biobank Resource under UK Biobank Main Application 15326. Work with All of Us data was performed using the All of Us Researcher Workbench under the workspace "Duplicate of Prediction of Polygenic Traits'".

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	No sample size calculation was performed. The TPMI recruited as many participants as possible. As of the end of December 2023, there were 565,390 enrolled participants from 16 partner medical centers, which included 33 affiliated hospitals.
Data exclusions	The raw genotypic data underwent quality control measures, and the genetic variants were excluded when they had call rate <0.02, minor allele frequency <0.01, Hardy-Weinberg equilibrium test p-value <1x10-6. We also excluded individuals with overall call rate <0.95, failed heterozygosity check, or inconsistent documented versus genetically determined sex.
Replication	All statistical analyses were replicated, and the reproducibility of our findings was confirmed.
Randomization	Randomization is not applicable in this study.
Blinding	The investigators were blinded to genotype and outcome allocation during data collection.

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

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	This study is not a clinical trial.
Study protocol	This study is not a clinical trial.
Data collection	TPMI participants were recruited from 16 partner medical centers, which included 33 affiliated hospitals, beginning in July 2019. The TPMI adopted a strategy of utilizing electronic medical record data from these hospitals in its original form, with personal identifying information removed. Since we are conducting genetic studies, selection bias is unlikely unless genetic markers are associated with participants' willingness to participate.
Outcomes	Dichotomized disease status was defined by phecodes, which were based on information extracted from the EMR using International Classification of Diseases (ICD) codes. To ensure robustness, cases were defined by having the diagnosis of the relevant condition on two or more clinical visits. We also extracted quantitative traits from the EMR, including anthropometric, vital sign and laboratory measurements, and we excluded the extreme outliers and removed or adjusted the treated and/or medicated measures based on previous research, and the median value was kept if the participant had multiple qualified measures. (Supplementary methods) In this study, we focused on 695 dichotomized phenotypes (phecodes) that had at least 2,000 cases and 24 quantitative traits that were measured in at least 100,000 individuals. These phecodes spanned 17 disease categories, including but not limited to infectious diseases, neoplasms, endocrine/metabolic disorders, and circulatory system diseases. The 24 quantitative traits were categorized into anthropometric, circulatory, hematological, kidney-related, liver-related, and metabolic measurements.

Plants

Seed stocks	not applicable in this study.
Novel plant genotypes	not applicable in this study.
Authentication	not applicable in this study.