nature research

Corresponding author(s): Uri Kartoun

Last updated by author(s): Oct 4, 2021

Reporting Summary

Nature Research 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 Research 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	We analyzed 3 independent data sources: the Explorys Dataset, Mass General Brigham (MGB), and the UK Biobank (UKBB). The Explorys Dataset is comprised of the healthcare data of over 21 million individuals, pooled from different healthcare systems with distinct electronic health records (EHRs) that have been previously used for medical research. Data were statistically de-identified, standardized, and normalized using common ontologies and made searchable after being uploaded to a Health Insurance Portability and Accountability Act-enabled platform. The data included EHR entries for all patients who were seen between January 1, 1999, and December 31, 2020. MGB is a large healthcare network serving the New England region of the US. We utilized the Community Care Cohort Project, an EHR dataset comprising over 520,000 individuals who received care at any of the 7 academic and community hospitals in MGB. The UKBB is a prospective cohort of over 500,000 participants enrolled during 2006–2010. Briefly, approximately 9.2 million individuals aged 40–69 years living within 25 miles of 22 assessment centers in the UK were invited, and 5.4% participated in the baseline assessment. Questionnaires and physical measures were collected at recruitment, and all participants are followed for outcomes through linkage to national health-related datasets.
Data analysis	All analyses were performed using R version 3.6, including the “survival,” “rms,” “data.table,” and “prodlim” packages.

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The institutional review boards of Mass General Brigham (MGB) and IBM approved this study and its methods, including the EHR cohort assembly using the Explorys Dataset, data extraction, and analyses. MGB data contain potentially identifying information and may not be shared publicly. Explorys data can be made available through a commercial license (for details see: https://www.ibm.com/downloads/cas/4P0QB9JN). We are indebted to the UKBB and its participants who provided biological samples and data for this analysis (UKBB Applications #7089 and #50658). All UKBB participants provided written informed consent. The UK Biobank was approved by the UK Biobank Research Ethics Committee (reference# 11/NW/0382).

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	For the AF scenario sample sizes are: Explorys (N = 4,750,660), UKBB (N = 445,329), MGB (N = 174,644). For the ASCVD scenario sample sizes are: Explorys (N = 3,328,992), UKBB (N = 408,154), MGB (N = 198,184). Full details of the cohort construction and sample sizes for the 3 datasets are shown in Table 1 and SUPPLEMENTARY TABLES I–VI.
Data exclusions	In all 3 datasets, individuals with missing data for AF risk estimation at baseline were excluded. We defined the ASCVD analysis set analogously, with exclusion of individuals with missing data needed to calculate the PCE score. In the UKBB we excluded all enrolled individuals who decided at a later point to withdraw consent. Full details of the cohort construction for the 3 datasets are shown in SUPPLEMENTARY TABLES I–VI.
Replication	We computed incidence rates for each outcome, reported per 1,000 patient years (1K PY). For each risk score and subgroup, we assessed the association between the risk score and its respective outcome using Cox proportional hazards regression, with 5-year AF as the outcome of interest for CHARGE-AF and 10-year ASCVD as the outcome of interest for PCE. Hazard ratios were scaled by the within-sample standard deviation (SD) of the linear predictor of each score for comparability (Standardized Hazard Ratio [SHR]). Therefore, the SHR reflects the relative increase in event hazard observed with a 1-SD increase in the respective linear predictor. We also assessed the discrimination of each score by calculating Harrell's c-index. We compared calibration slopes, defined as the beta coefficient of a univariable Cox proportional hazards model with the prediction target as the outcome and the linear predictor of the respective risk score as the sole covariate, where an optimally calibrated slope has a value of one.
Randomization	NA
Blinding	NA

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	Age, sex, race, and smoking status were defined using EHR fields in Explorys and MGB and were self-reported at the initial assessment visit in UKBB. Height, weight, blood pressure, total cholesterol, and high-density lipoprotein cholesterol values were measured relative to baseline in all 3 datasets. For patients with multiple eligible values in the baseline period, only the most recent was used. Smoking status was classified as present or absent and race was classified as White or Black. Patients who indicated themselves as Black (possibly with one or more other race types) were considered Black for risk calculations, and White otherwise. The presence of clinical comorbidities was ascertained using diagnostic (International Classification of Diseases-9th [ICD-9] and -10th [ICD-10] revisions) and procedural (Current Procedural Terminology, CPT) codes, either extracted from the EHR (Explorys and MGB), or from linked national health record data (UK Biobank). All covariates were used in accordance with the CHARGE-AF and PCE definitions. Clinical factor definitions of all covariates appear in SUPPLEMENTARY TABLE VII.
Recruitment	To ensure adequate data ascertainment and follow-up, we included in Explorys individuals with at least two outpatient encounters greater than or equal to 2 years apart. Individuals in the MGB dataset had at least one pair of primary care office visits 1-3 years apart. We included all individuals who enrolled in the UKBB study. In Explorys, the start of follow-up was defined as the first encounter following the second qualifying outpatient encounter. In MGB, the start of follow-up was defined as the second office visit of the earliest qualifying pair. In UKBB, as an enrollment-based resource, start of follow-up was the date of the initial assessment visit. The primary outcomes were 5-year incident AF (for the AF Subsets), and 10-year incident ASCVD (for the ASCVD Subsets). Incident AF was defined using a modified version of a previously validated EHR-based AF ascertainment algorithm (positive predictive value 92%), in which electrocardiographic criteria were not used given the absence of electrocardiogram reports in the Explorys Dataset. Incident ASCVD was defined as a composite of myocardial infarction (MI) and stroke, each defined using previously published sets of diagnosis codes. Outcome definitions are shown in SUPPLEMENTARY TABLE VII.
Ethics oversight	The institutional review boards of Mass General Brigham (MGB) and IBM. The UK Biobank was approved by the UK Biobank Research Ethics Committee (reference# 11/NW/0382).

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