nature portfolio

Corresponding author(s): Qian Di

Last updated by author(s): Apr 28, 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. All data were already collected.
Data analysis	Large language models analysis was conducted in Python 3.10.8 with Pytorch 2.1.0, vllm 0.5.5. Statistical analyses were performed in R 4.4.2. The Python and R codes and package requirements for main analyses are available at https://github.com/AI4HEALTH-LAB-THU/LLM-Aging.

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 were obtained from the UK Biobank (UKB), the National Health and Nutrition Examination Survey (NHANES), the China Health and Retirement Longitudinal Study (CHARLS), Chinese Longitudinal Healthy Longevity Survey (CLHLS), the China Family Panel Studies (CFPS), and the Northwest China Real-world and Population-based Cohort (NCRP). Researchers can register to access the 5 nationally population data employed in this study via the UK Biobank Access Management System

(https://bbams.ndph.ox.ac.uk/ams/), the NHANES database (https://wwwn.cdc.gov/nchs/nhanes/default.aspx), the CHARLS database (https://charls.pku.edu.cn/), the CLHLS database (https://opendata.pku.edu.cn/dataverse/pku/), and the CFPS database (https://www.isss.pku.edu.cn/cfps/sjzx/gksj/index.htm). For the NCRP cohort over 9.6 million participants, we welcome collaboration from over the world to maximize the usage of this data. Data are inaccessible for public download due to data safety and privacy concerns. Researchers at academic or non-profit institutions may request access by submitting to the corresponding author (Qian Di) a research proposal that specifies study aims, analysis plan, and data security statement. Proposals are reviewed on a rolling basis and applicants will be informed of the outcome (approval, revision request or denial) within six to eight weeks. Access may be declined if the proposed study falls outside NCRP objectives, fails to meet ethical or data-protection standards or duplicates existing work. Approved investigators may conduct analyses on offline servers at Tsinghua University or People's Hospital of Xinjiang Uygur Autonomous Region. A fully de-identified offline mirror for secure remote access is under development, and detailed information will be available in the forthcoming cohort profile publication.

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	In this study, we used the term sex to inidicate the biological attribute. Information on sex was based on self-report. Within our study populations, females comprised 54.3% in UKB, 48.3% in NHANES, 52.5% in NCRP, 50.8% in CHARLS, 61.6% in CLHLS, and 51.0% in CFPS. In our statistical analyses, sex was included as a covariate.
Reporting on race, ethnicity, or other socially relevant groupings	In this study, we incorporated three socio-related variables: income level, educational level, and occupation. The term 'income' was used to represent income level (low level, middle level, and high level), 'education' for educational level (college, below undergraduate, and uneducated), and 'employment' for occupation (in paid employment, other jobs, retired, and unemployed). These variables were all based on self-report. We included these variables when using large language models to predict overall and organ-specific ages, and subsequently used them as covariates in our statistical analyses.
Population characteristics	We utilized 6 population cohorts in this study, including the UKB (n=489,391, age range 38-73 years, 54.3% female), the NHANES (n=2,009, age range 50-75 years, 48.3%female) the NCRP (n=9,633,240, age range 18-110 years, 52.5% female), the CHARLS (n=17,870, age range 40-70 years, 50.8% female), the CLHLS (n=33,173, age range 80-123 years, 61.6% female), and the CFPS (n=19,242, age range 40-70 years, 51.0% female), in totaling 10,194,925 participants.
Recruitment	Participants were recruited by the UK Biobank, the National Health and Nutrition Examination Survey, the China Health and Retirement Longitudinal Study, the Chinese Longitudinal Healthy Longevity Survey, the China Family Panel Studies, and the Northwest China Real-world and Population-based Cohort.
Ethics oversight	The UKB: Ethical information (http://www.ukbiobank.ac.uk/ethics/) and the full protocol (https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf) about UKB are available online. The NHANES: The research protocols were approved by the National Center for Health Statistics (NCHS) Research Ethics Review Board. Informed consent was obtained from all participants. The NCRP: The Tsinghua University Science and Technology Ethics Committee (Medicine) approved the entirely of the project and related protocols (THU01-20240123). The CHARLS: Ethical approval for all the CHARLS waves was granted by the Institutional Review Board at Peking University. The IRB approval number for the main household survey, including anthropometrics, is IRB00001052-11015; the IRB approval number for biomarker collection, was IRB00001052-11014. The CLHLS: The research ethics committees of Peking University (IRB00001052–13074) approved the study. All participants were provided with informed consent. The CFPS: The CFPS has approval from the Biomedical Committee of Peking University to obtain and disseminate data from the participants (IRB00001052-14010).

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 the relevance to the research objective and the data availability from the UK Biobank, the National Health and Nutrition Examination Survey, the China Health and Retirement Longitudinal Study, the Chinese Longitudinal Healthy Longevity Survey, the China Family Panel Studies, and the Northwest China Real-world and Population-based Cohort. These datasets collectively provide a large and diverse sample, offering sufficient statistical power for analyses and ensuring that findings are generalizable across populations from different sociodemographic backgrounds.
Data exclusions	In the UKB, we excluded participants who lacked an entire category of indicators in our definition of health information (for instance, those with no information on dietary habits were excluded; see the Methods). In NHANES, we excluded participants under 50 and over 75 years of age. In CHARLS and CFPS, we excluded participants under 40 and over 70 years of age. In CLHLS, we excluded participants under 80 years old. In NCRP, we excluded participants under 18 and over 110 years of age.

Replication	We conducted our primary analysis in the UKB, NHANES, and replicated the analysis across 4 independent cohorts: CHARLS, CLHLS, CFPS, and NCRP, to provide additional validation and demonstrate the generalizability of our framework for predicting individual multidimensional aging based on large language models. The replications conducted in the NCRP, CHARLS, and CFPS populations obtained similar results with primary analyses. However, opposite results were observed in the CLHLS. This may be due to the fact that CLHLS primarily comprises individuals aged 80 and above, and large language models may have limited capability in aging assessment for vulnerable groups such as the elderly. This has been discussed in the main text.
Randomization	This study was designed as a prospective cohort study, where all participants were enrolled based on predefined eligibility criteria and observed over time without any intervention or allocation into different treatment arms. Therefore, randomization was not applicable, as the study aimed to observe natural outcomes rather than compare effects between randomized groups.
Blinding	This study was designed as a prospective observational cohort study without any intervention or group allocation. Since all participants were observed under real-world conditions and no comparative treatment assignment was involved, blinding was not applicable.

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