nature portfolio

Corresponding author(s): Andrey Rzhetsky

Last updated by author(s): Jun 29, 2026

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 extraction from the annual SAS7BDAT releases was performed using custom Python 3.11 scripts (pandas, sqlite3).
Data analysis	Statistical analyses used Python 3.11 (pandas, numpy, scipy, statsmodels, joblib for parallelization, matplotlib 3.8 for figures, and xgboost for the exploratory adjusted T-learner) and R 4.3 (survival for conditional logistic regression; sandwich and lmtest for robust and clustered variance estimation). All custom analysis code is publicly available at https://github.com/benjaminkramer510/BirthOrderPhenome2026.

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

The individual-level Merative MarketScan Commercial Claims and Encounters data are proprietary and cannot be redistributed by the authors. Any qualified academic, government, or commercial researcher or institution may request access directly from Merative (https://www.merative.com/real-world-evidence) by

completing its licensing and data-use-agreement process, for research use consistent with that agreement. The authors held no special access privileges. Summary-level data sufficient to interpret, verify, and extend the findings are provided in Supplementary Tables 1 to 15 and a 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	The administrative enrollment files include a sex variable, which was used as recorded in MarketScan and was incorporated as a covariate in the primary models. Gender identity was not available in the claims data and could not be analyzed. Sex-stratified analyses are reported in Supplementary Table 14. No gender-stratified analyses were performed because gender was not measured.
Reporting on race, ethnicity, or other socially relevant groupings	Race, ethnicity, and other socially relevant groupings are not available in the Merative MarketScan commercial claims extract and were therefore not analyzed.
Population characteristics	The study population comprises 10,270,012 de-identified individuals in 5,135,006 two-child families. Covariate-relevant characteristics include sex (50.9% male first-borns), birth years 1978 to 2013 (mean 1994.7), mean paternal age at birth 31.4 (first-born) and 35.0 (second-born) years and maternal age 29.4 and 32.9 years, predominantly high-urbanicity residence, all four U.S. Census regions, and the full diagnostic spectrum (569 ICD-defined phenotypes). Full characteristics are in Table 1 and the Methods.
Recruitment	No participants were recruited or contacted. The cohort was assembled by algorithmic selection of all eligible two-child families from de-identified secondary administrative claims records, using pre-specified criteria described in the Methods. Enrollment is limited to individuals in employer-sponsored insurance plans, a potential source of selection bias that is acknowledged in the manuscript.
Ethics oversight	The University of Chicago Institutional Review Board reviewed the study and determined it exempt from human subjects review (secondary analysis of de-identified commercial claims data). Informed consent was not applicable.

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 a priori statistical sample-size calculation was performed. We analyzed the entire eligible Merative MarketScan two-child-family cohort (5,135,006 families; 10,270,012 individuals) to maximize statistical power; this is the largest available commercial-claims sibling cohort. Minimum case-count thresholds (at least 500 cases for between-family analyses; at least 100 disease-discordant sibling pairs for within-family analyses) were applied to ensure stable estimation.
Data exclusions	Exclusion criteria were pre-established. Families were required to have at least one inferred parent and exactly two non-parent children; each child required at least 365 days of enrollment visibility, age at least 12 years at last observation, and valid birth-year and observation data. Diseases not meeting the minimum case-count thresholds were not analyzed (418 of 569 diseases met the between-family threshold; 541 met the within-family threshold). No analyzed observations were otherwise excluded.
Replication	All replication and robustness analyses were successful. The primary between-family findings reproduced in direction under state fixed effects ($r > 0.99$), full-sibling restriction ($r = 0.998$), the stricter clinical rematch ($r = 0.93$), and alternative link functions (log-link Poisson and complementary log-log, direction-concordant for all 418 diseases). The between- and within-family designs were positively correlated ($r = 0.66$; 84.7% of Bonferroni-significant between-family hits were directionally concordant).
Randomization	Not applicable. Birth order is fixed by family structure and cannot be randomly assigned. Confounders that covary with birth order (parental age at birth, calendar period, sibling spacing, sex, follow-up duration, ICD coding era, and geography) were controlled by exact and caliper matching and by regression adjustment; the within-family conditional logistic regression additionally absorbs all time-invariant shared family confounders.
Blinding	Not applicable. This is a fully specified observational analysis of administrative data with no experimental intervention or subjective outcome assessment; outcomes are defined algorithmically from ICD codes, so investigator blinding is not relevant.

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

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.