nature portfolio

Corresponding author(s): ALison Motsinger-Reif

Last updated by author(s): Jul 9, 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	REDCap was used to collect answer to the questionnaires. The questionnaires are publicly available here: https://www.niehs.nih.gov/research/atniehs/labs/crb/studies/pegs/about/data
Data analysis	Open source R software was used for analysis, and all code is available on Github, as described in the manuscript.

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

Individual-level PEGS data cannot be deposited in an unrestricted public repository because the cohort includes human participant questionnaire data, health information, biospecimen-derived data, genetic/genomic data, and geospatial information. PEGS data are shared through a controlled-access process. PEGS maintains a non-PII/de-identified dataset and a more restricted PII/geolocation-containing dataset, each governed by a distinct Data Use Agreement. The non-PII

dataset includes individual-level survey, exposure, health, omics-derived, polygenic-score, metadata, data-dictionary, and supporting documentation resources. The restricted PII/geolocation version includes residential geolocation details and related information needed for justified geospatial linkages. Qualified academic, governmental, nonprofit, and commercial researchers may apply for access through the NIEHS PEGS Access System. Applications are reviewed for consistency with participant consent, NIH IRB approvals, scientific validity, data necessity, privacy/re-identification risk, and adequacy of data protections.

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	Sex was collected and reported for PEGS participants. The enrolled PEGS cohort is approximately 62.3% female and 37.6% male, and the survey-specific demographic table reports sex as counts and percentages. WGS quality control included consistency checks between self-reported and genotype-inferred sex. Sex was included as a covariate in disease models where appropriate; for sex-specific phenotypes, fibroids and ovarian cysts, models were adjusted for age alone. Gender identity is not reported separately in the manuscript.
Reporting on race, ethnicity, or other socially relevant groupings	Race and ethnicity were reported for cohort characterization. The enrolled PEGS cohort is racially and ethnically diverse, with approximately 63% of participants identifying as White, 27.6% as Black, 4.5% as another race, and 5.1% as Hispanic. Race/ethnicity were used for descriptive characterization and WGS quality-control checks, not as proxies for socioeconomic status. Socioeconomic and structural factors were handled separately through variables such as parental education, participant education, household income, and housing type, which were used to construct polysocial scores.
Population characteristics	The study used data from the Personalized Environment and Genes Study, an ongoing cohort of adults in North Carolina conducted by NIEHS. PEGS collects information on genetics, diet, lifestyle, medications, diseases, environmental exposures, and social/demographic factors through questionnaires, genomic data, and related health information. The enrolled cohort includes 19,445 participants and is demographically similar to North Carolina's population, with an average age of approximately 50.2 years. The analytic validation cohort included 2,344 participants with WGS data who completed all three PEGS surveys.
Recruitment	Participants were recruited into the Personalized Environment and Genes Study (PEGS) over approximately the last 20 years at the National Institute of Environmental Health Sciences (NIEHS). PEGS is an ongoing cohort study of adults in North Carolina conducted by NIEHS, and the present analysis used existing PEGS Data Freeze 3.1 data rather than newly recruiting participants for this study. No new recruitment or enrollment was performed specifically for the current analysis. The analytic sample was restricted to existing PEGS participants with the required survey data and, for validation analyses, available WGS data; therefore, selection bias may arise from long-term cohort participation, completion of all three PEGS surveys, and availability of WGS data.
Ethics oversight	The PEGS study protocol, informed consent forms, recruitment materials, participant materials, human data sharing, genomic data sharing, and publications were approved by the National Institutes of Health Institutional Review Board: NIH IRB #04E0053, protocol 04-E-0053. The study is also listed under ClinicalTrials.gov identifier NCT00341237. All PEGS participants provided written informed consent for use of questionnaire data, biospecimens, genetic/genomic data, geospatially derived exposure information, and related health information for research. The current analyses used de-identified PEGS Data Freeze 3.1 data.

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 formal a priori sample-size calculation is reported. Sample size was determined by availability of PEGS participants with the required questionnaire and WGS data and by adequate case counts for the selected disease phenotypes. PXS/PSS feature selection used a derivation dataset of PEGS participants without WGS data (n = 432). Score comparison and classification analyses used PEGS participants with WGS data who completed all three PEGS surveys (n = 2,344), randomly split into training (n = 1,758) and test (n = 586) sets.
Data exclusions	Participants were excluded according to WGS quality-control and analytic-completeness criteria. WGS filtering excluded seven participants whose samples may have represented DNA mixtures or potential aneuploidies, 120 participants who were second-degree relatives or closer, and three pairs of duplicate samples. Additional stringent filtering excluded participants with discrepancies between self-reported and WGS-inferred ancestry/race, WGS outliers, related individuals, and participants missing key demographic data such as age or sex. The validation analysis was limited to participants with WGS data who completed all three PEGS surveys. PGS models with less than 70% of variants available were excluded, and palindromic variants were removed.
Replication	The study included internal validation but not an independent external replication cohort. Feature selection for PXS/PSS was performed in a derivation dataset without WGS data, while classification performance was evaluated in a held-out WGS validation dataset split into training

and test sets. Bootstrap resampling was used to quantify uncertainty for AUC and calibration metrics. The manuscript states that external validation, recalibration, and prospective evaluation are needed before clinical implementation.

Randomization

No participants were allocated to experimental groups; this was an observational, cross-sectional cohort analysis. The WGS validation dataset was randomly split into training and test sets for model evaluation. Covariates such as age and sex were controlled in regression models as appropriate.

Blinding

Blinding to group allocation was not applicable because there was no intervention or experimental allocation. Data were obtained from existing PEGS questionnaire, genomic, and health-related data. Disease status was based on self-reported ever-diagnosed disease history, and analyses were performed using prespecified computational workflows.

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

Not applicable as a clinical trial; this was an observational cohort analysis. The PEGS cohort protocol is listed under ClinicalTrials.gov identifier NCT00341237.

Study protocol

The PEGS study protocol was approved by the NIH IRB under IRB #04E0053, protocol 04-E-0053. PEGS data access and supporting study information are available through the NIEHS PEGS Access System. No interventional trial protocol applies.

Data collection

Data were collected in PEGS, an ongoing cohort study of adults in North Carolina conducted by NIEHS. PEGS participants completed health, exposure, social/demographic, occupational/hobby, lifestyle, diet, medication, physical activity, stress, sleep, and infection questionnaires and provided biospecimens. WGS was performed by the Broad Institute for PEGS participants with the most complete survey data. The current analysis used PEGS Data Freeze 3.1.

Outcomes

The analyzed outcomes were self-reported, ever-diagnosed disease phenotypes: asthma, high cholesterol, fibroids, hypertension, iron-deficient anemia, lower GI polyps, migraine, ovarian cysts, and type 2 diabetes. Case/control status was based on participant report of ever having been diagnosed by a clinician. The study's primary analytic outcomes were classification and stratification metrics for PXS, PSS, PGS, and their combinations, including AUC, ΔAUC, NRI, calibration, McFadden's adjusted pseudo-R^2, odds ratios, score correlations, and high-/low-score group overlap.

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.