

Polyexposure, Polysocial, and Polygenic Scores for Common, Complex Disease Classification and the Creation of a Clinical Exposure Assessment

Alison Motsinger-Reif

`alison.motsinger-reif@nih.gov`

National Institute of Environmental Health Sciences

Dillon Lloyd

Harvard University

Farida Akhtari

National Institute of Environmental Health Sciences

John House

National Institute of Environmental Health Sciences <https://orcid.org/0000-0002-8447-7871>

Chirag Patel

Harvard Medical School <https://orcid.org/0000-0002-8756-8525>

Charles Schmitt

National Institute of Environmental Health Sciences

David Fargo

Lawrence Kirschner

National Institute of Environmental Health Sciences <https://orcid.org/0000-0001-6180-342X>

Janet Hall

National Institute of Environmental Health Sciences

Kyle Walsh

National Institute of Environmental Health Sciences

Joshua Denny

National Institutes of Health <https://orcid.org/0000-0002-3049-7332>

Geoffrey Ginsburg

All of Us Research Program, National Institutes of Health <https://orcid.org/0000-0003-4739-9808>

Richard Woychik

National Institute of Environmental Health Sciences

Article

Keywords:

Posted Date: July 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10284839/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Polyexposure, Polysocial, and Polygenic Scores for Common, Complex Disease Classification and the Creation of a Clinical Exposure Assessment

Dillon T. Lloyd^{1*}, Farida S. Akhtari^{2*}, John S. House², Chirag J. Patel¹, Charles P. Schmitt³, David C. Fargo⁴, Lawrence S. Kirschner⁵, Janet E. Hall⁵, Kyle M. Walsh⁴, Josh C. Denny⁶, Geoffrey S. Ginsburg⁶, Richard P. Woychik⁷, Alison A. Motsinger-Reif^{1#}

* Equal Contribution

Corresponding author

1. Department of Biomedical Informatics, Harvard University, Cambridge, MA, USA
2. Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA
3. Office of the Scientific Director, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA
4. Office of the Director, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA
5. Clinical Research Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA
6. All of Us Research Program, National Institutes of Health, Bethesda, MD, USA
7. Office of the Director, National Institutes of Health, Bethesda, MD, USA

Abstract

Genetic risk, as well as the physical environment and psychosocial factors, affect susceptibility to common, complex diseases. Although portions of patients' social and exposure histories are captured during routine healthcare, these data are rarely collected in a standardized, quantitative format that supports cross-disease risk stratification. We analyzed data from the Personalized Environment and Genes Study (PEGS) for 2,344 participants with genome sequencing who completed three exposome surveys. We derived polyexposure (PXS) and polysocial (PSS) scores from 489 variables and compared disease-classification performance of these scores, polygenic scores (PGS), and classification based on age/sex across nine self-reported, ever-diagnosed diseases. For cross-sectional classification of disease risk, PXS had greater improvement than PGS or PSS compared to models with common clinical factors and demographic variables such as sex and age. The combined PXS+PSS+PGS model had the best performance overall (mean Δ AUC 0.10; maximum Δ AUC 0.27 for asthma). Importantly, the score types identified different individuals in high-score groups, and there was minimal overlap among these groups (PGS-PXS 0.94%; PGS-PSS 1.0%; triple overlap 0.18%, averaged across diseases). Further, PGS-PXS and PGS-PSS correlations were weak (mean $r=0.04$ and 0.03). This independence may enable identification of different individuals at risk of disease compared to existing clinical assessments and could help tailor interventions and care. As a next step in

determining clinical feasibility, we propose a 66-question minimum exposure question set, with 61 questions based on PXS and five on PSS, as a pragmatic framework for clinical collection of standardized data for multiple diseases.

Introduction

In addition to the well-known influence of genetics on disease, the physical environment and psychosocial factors impact disease etiology¹. The exposome, defined as the totality of environmental exposures across the life course^{2,3}, encompasses occupational, lifestyle, psychosocial, and structural socioeconomic factors and is an important contributor to human health and disease⁴. Polyexposure scores (PXS) and polysocial scores (PSS) summarize information on the exposome and are used to classify disease status and estimate disease risk. These scores are analogous to polygenic scores (PGS), which use genetic variation to identify disease susceptibility⁸.

PXS aggregate information on occupational, residential, and lifestyle exposures, and PSS⁵⁻⁷ summarize information on the social environment. PXS studies typically focus on one condition, such as type 2 diabetes^{9,10}, and emphasize modifiable lifestyle exposures such as diet, exercise, and smoking status, often without distinguishing these from structural factors such as housing and education, which are difficult or impossible to modify. In this study, we explicitly separate modifiable exposures (PXS) from structural factors (PSS) to clarify their distinct contributions to disease.

Although PXS and PSS have demonstrated utility in epidemiologic research^{11,12}, their integration into clinical practice remains limited^{13,14}. Routine care often includes the collection of social history and screening for selected social risks¹⁵, but these assessments are heterogeneous across settings and typically do not capture a broad range of residential, occupational, behavioral, and structural exposures in a reusable, quantitative format. Occupational hazards are assessed using job-exposure matrices¹⁶, with broader occupational exposure assessments largely restricted to acute events or specific populations^{17,18,19}. As a result, although clinicians may document relevant contextual information, there is no standardized framework for translating this information into cross-disease risk classification.

To address these gaps, we leverage data from the Personalized Environment and Genes Study (PEGS), a diverse, North Carolina-based cohort that collects detailed information on physical,

lifestyle, and structural exposures with three questionnaires. Using these data, we compared the performance of PGS, PXS, and PSS and their combination for cross-sectional risk classification of nine self-reported, ever-diagnosed diseases. We determined whether these scores identify distinct high-risk groups and measured the overlap in high-score groups identified by each score type and their combinations. Using the results of the analysis and the features included in the models, we derived a minimum exposure question set to complement existing structural and genetic tools, determining its clinical feasibility as part of the development process.

This work investigates whether standardized questionnaire-derived exposure and social scores capture risk information that is complementary to existing measures of inherited susceptibility and demographics across multiple phenotypes, not to position PXS as a clinical gold standard or replace a conventional social history. Our central premise is that PXS and PSS scores provide information distinct from and complementary to PGS and are translationally relevant, as exposures inform disease prevention and intervention. This work thus focuses on the classification and stratification of disease risk and not prognosis.

Results

Data-driven Exposure Selection Yields Distinct Polyexposure and Polysocial Scores across Diseases

Score development is outlined in Figure 1. We examined nine phenotypes, not because they share a single mechanism, but because they provide a test case for whether an exposure questionnaire can generate clinically complementary information across outcomes that differ in diagnostic standard, expected use case, and actionability. From a total of 489 exposures, the number and category of exposures selected for use in disease-specific PSS and PXS varied across phenotypes (Figure 2A).

For PSS, the number of selected features ranged from 13 for fibroids to eight for hypertension and lower gastrointestinal (GI) polyps. Three exposures (maternal education level, paternal education level, and participant housing type) were selected for all diseases. Of the 17 exposures selected, three were unique to a single disease.

For PXS, the number of selected exposures ranged from 58 for hypertension to 33 for ovarian cysts. The highest number of features unique to a single phenotype was 20 for lower GI polyps,

followed by 16 each for migraine and fibroids (Figure 2A). Sleep disrupted by pain was selected for seven diseases, and exercise level was selected for five diseases. Of the 235 selected exposures, 123 were unique to a single disease (Figure 2B).

Polyexposure Scores, Polysocial Scores, and Polygenic Scores Classify Diseases across Biological Systems

To quantify classification performance, we compared the area under the curve (AUC) for PGS, PSS, PXS, and their combinations with the AUC of a baseline model containing factors commonly included in clinical assessments (age, sex, physical activity, and BMI), denoted as Δ AUC (Figure 3A), with a larger Δ AUC indicating improved performance. Across diseases, the largest average increase in AUC over the baseline model was observed from PXS alone (mean Δ AUC 0.08). Smaller average increases (0.02 and 0.02) were observed for PSS and PGS. The combined PXS+PSS+PGS model had the best performance overall (mean Δ AUC 0.10). For this combined model, asthma had the largest Δ AUC (0.27), and lower GI polyps had the smallest (0.02). For PXS alone, asthma (0.25) had the largest Δ AUC, and hypertension (0.03) had the smallest. For PSS, migraine (0.06) had the largest Δ AUC, and iron-deficient anemia (-0.004) had the lowest. For two-score models, PXS+PGS had the highest Δ AUC (0.07), followed by PXS+PSS (0.06) and PSS+PGS (0.04). Additionally, we used McFadden's pseudo R^2 (Figure S1) to assess goodness of fit for single and multi-score models, adjusting for model complexity. The results indicate that most of the improvement in classification performance was driven by exposure information, with more modest benefit from genetic and structural-social information.

We also examined net reclassification improvement (NRI), which determines whether a model moves cases into a higher risk category and non-cases into a lower risk category relative to the clinical model. Across the study-defined probability categories, the PXS-alone model yielded greater average reclassification than the PSS-alone or PGS-alone model (mean NRI 0.14 vs 0.06 and 0.05), and the PXS+PSS+PGS model had the best performance (mean NRI 0.17). The largest improvement from this combined model was observed for asthma (NRI 0.47), indicating meaningful reclassification beyond the covariate-only model within the probability strata used in this study (Figure S2).

To determine alignment of the probability predicted by each model with true likelihood, we assessed model calibration across scores and their combinations. Following logistic recalibration,

calibration intercept values closer to zero indicated better calibration (Figure 3D). Calibration across outcomes had a mean difference of 0.01, suggesting similar calibration across models. Using asthma as a representative example, the calibration intercept for the PXS+PSS+PGS model (0.21) indicated modest deviation from perfect calibration, while calibration curves showed limited dispersion of predicted probabilities across the probability distribution (Figure S3).

We assessed the odds ratios (ORs) and statistical significance of each score and their combinations compared to the baseline clinical model (Figure 3E). Across multi-score models, the OR of the PXS term was always statistically significant after limiting false discovery to 10%. PGS did not meet this significance threshold for high cholesterol, asthma, iron-deficient anemia, and ovarian cysts. PXS alone had an average OR of 1.95 across phenotypes, with the highest OR for asthma (2.27), and the lowest for high cholesterol (1.67). PSS alone had an average OR of 1.36, and the combination of PXS, PSS, and PGS had an average OR of 1.88.

To characterize the goodness of fit of models containing PXS, PSS, PGS, and their combinations compared to age and sex (or age alone), we calculated McFadden's adjusted pseudo R^2 , which measures the likelihood of a fitted model over the null model, adjusting for model complexity (Figure 3E). The addition of PXS increased the average pseudo R^2 by 0.06, compared to 0.012 for PSS alone and 0.007 for PGS alone. When adding PXS, the largest increase was for asthma (0.11) and the smallest was for high cholesterol (0.03). When adding PSS, the largest increase was for type 2 diabetes (0.03) and the smallest was for fibroids (0.002). The PXS+PSS+PGS model (0.07) had the largest average increase in pseudo R^2 ; the largest increase observed for asthma (0.12), and the lowest increase was observed for high cholesterol (0.04).

Polyexposure and Polysocial Scores Identify Distinct Vulnerable Populations beyond Polygenic Scores

To examine the relationships among PXS, PSS, and PGS, we conducted pairwise correlation analyses across diseases. PXS exhibited high correlation (mean $r = 0.32$) (Figure 4A); the strongest correlation was between migraine and ovarian cysts ($r = 0.63$), and the weakest was between migraine and lower GI polyps ($r = 0.07$). Disease-specific PSS were strongly correlated across diseases (mean $r = 0.41$) (Figure 4B); the strongest correlation was between type 2 diabetes and hypertension ($r = 0.95$), and the weakest was between hypertension and lower GI

polyps ($r = 0.007$). Notably, for PSS, several pairs were negatively correlated, including fibroids and lower GI polyps ($r = -0.41$) and ovarian cysts and lower GI polyps ($r = -0.21$). Weak correlations were observed for PGS (mean $r = 0.01$); the strongest correlation was between type 2 diabetes and fibroids ($r = 0.32$), and the weakest was between type 2 diabetes and lower GI polyps (-0.21) (Figure 4C). When examining different score types for the same disease (Figure 4D), on average, there were weak correlations between PSS and PGS (mean $r = 0.03$) and PXS and PGS (mean $r = 0.04$). The strongest correlations were between PXS and PSS (mean $r = 0.22$).

We assessed the concordance of high- and low-score groups across scores and diseases identified by PGS, PXS, and PSS. In high-score groups, a small fraction of individuals was uniquely identified by any single score. On average, among individuals in the top 10% of a score, 7.9%, 7.3%, and 7.3% were uniquely identified by PGS, PXS, or PSS, respectively, indicating independence across scores (Figure 4E). Overlap between the high-score PGS and the high-score PXS groups was 0.94%; overlap between the high-score PGS and high-score PSS was 1%; overlap between the high-score PSS and high-score PXS groups was 1.6%. Triple overlap was limited, with a mean 0.18% of individuals simultaneously classified as high score by all three scores. Similar patterns were observed for low-score groups (Figure 4F). A chi-square test of independence indicated no statistically significant difference in the high-score groups identified by PXS and PGS. This result implies that high-score PXS and PSS groups are independent of high-score PGS groups.

Developing a Framework for Clinical Exposure Assessment Using Self-Reported Exposome Information

We developed overall PXS and PSS using all available exposome data and created a framework for a proposed minimum exposure question set that includes social and physical exposome information. From an initial 489 exposures, 235 features were selected for inclusion in at least one disease-specific PXS (Figure 5A). Of the selected exposures, 123 were selected once. One exposure (amount of sleep) was selected seven times (Figure 5B), and none were selected for all nine diseases. For PSS, three structural exposures (maternal education level, paternal education level, and housing type) were selected for all nine diseases, and two exposures were selected for one disease.

We then developed disease-specific scores using exposures selected between one and 10 times across diseases, for a total of six distinct PXS and 10 distinct PSS (Figure S4). To better understand the number of exposures needed to build scores capable of outperforming PGS, we combined the results of our iterative score development with the correlation structure of the PEGS Internal and External Exposome surveys. We assessed correlations between exposures selected for at least one disease-specific PXS ($n = 235$). We classified exposures into the following categories: UV/Sun, Health History, Occupational and Hobby Exposures, Lifestyle Exposures, Indoor/Residential, and Diet. Exposures in the same category had an average correlation of 0.14 (Figure S5). The UV/Sun exposure category had the strongest correlation (0.33), and Indoor/Residential exposures had the weakest correlation (0.10). Next, we applied the Kaiser rule within these categories to estimate the number of effective features, defined as the number of eigenvalues of the correlation matrix greater than one. The mean reduction of features in a category was 71%. The largest reduction from total features to effective features was observed for the categories of diet (56, 72% reduction), lifestyle factors (31, 76% reduction), and occupational exposures (29, 73% reduction) (Figure 5C).

For the proposed minimum exposure question set, we used the number of effective features for each category. The common PXS included 61 exposures, and the common PSS included five exposures (Figure 5D) from the following categories: diet (number of questions = 22), indoor/housing (11), Lifestyle (10), Occupational/Hobby (8), Health History (7), and UV and sun exposure (3). The exposures selected for each category are listed in Tables S1 and S2.

We constructed single disease-agnostic PXS and PSS using the selected exposures and evaluated them in models containing baseline clinical covariates (age, sex, physical activity, and BMI) and combined-score models. Across diseases, the common PXS had a mean ΔAUC of 0.07 over these clinical covariates alone (Figure 5E). The PXS for asthma had the best performance ($\Delta AUC = 0.17$), while the PXS for high cholesterol ($\Delta AUC = 0.01$) had the poorest performance. The average ΔAUC for PSS and disease-specific PGS was 0.02 and 0.03, respectively. A model containing the common PXS, PSS, and disease-specific PGS had the highest average ΔAUC (0.09). $\Delta McFadden$'s pseudo R^2 reflected similar patterns (Figure S6). NRI, compared to the base model containing age and sex (or age) alone, as well as model calibration, were computed for the disease-agnostic PXS and PSS (Figures S7-8).

To evaluate the goodness of fit, we calculated McFadden's adjusted pseudo- R^2 for models including the disease agnostic PXS and PSS, the disease specific PGS, and their combinations across diseases, and compared it to the pseudo- R^2 for the model containing clinical covariates alone. PXS-alone models had an average Δ pseudo- R^2 of 0.07 compared to 0.009 for PSS and 0.007 for PGS (Figure 5F). The PXS-alone model for type 2 diabetes had the largest Δ pseudo- R^2 (0.08), and the PXS-alone model for lower GI polyps had the lowest (0.02). The PXS+PSS+PGS model (0.08) had the largest average Δ pseudo- R^2 among combined-score models.

We shared the proposed question set with a panel of clinicians to help determine the feasibility of the exposures included in the PXS and PSS. The clinicians assessed how collectable a construct is and the barriers to collection in a custom exposome survey module (Text S1). The goal was to better understand the work required to translate these findings into a clinical setting. Figure S9 outlines an example workflow for administering the question set to a patient and indicates how these scores could be used for better disease prognosis. Table S3 outlines the current diagnosis guidelines for each disease and discusses how an integrated PXS/PSS system can be used to support current clinical guidelines.

Discussion

Across diseases, PXS provided the largest improvement in classification performance beyond clinical factors, with the best performance observed for the combined PXS+PSS+PGS model. The findings are notable because, although the Human Genome Project and subsequent genomics efforts have transformed our understanding of disease biology, genetic information alone explains only a modest fraction of inter-individual variation for many common diseases²⁰. The exposome framework highlights that physical, behavioral, occupational, and social environments contribute substantially to disease susceptibility and often represent more proximal or modifiable targets for prevention and intervention⁴. The proposed minimum exposure question set can facilitate scalable integration of exposomic information into healthcare.

Improvements in classification performance were achieved with the inclusion of self-reported questionnaire data, suggesting that relatively simple exposure assessments can capture clinically relevant information that complements both genomic and social risk measures. One possible explanation for the strong performance of PXS is that questionnaire-derived exposures capture cumulative environmental and behavioral influences that may be more closely associated with

current disease status than inherited genetic variation. Because many common diseases arise from complex interactions between biological, behavioral, occupational, and social factors, exposure-based measures may capture risk information not reflected in genomic or clinical covariates alone. Moreover, unlike genomic testing, exposure questionnaires can be completed during routine clinical encounters and updated over time as exposures change, providing a practical and scalable mechanism for incorporating dynamic risk information into healthcare.

The key value of integrating PXS, PSS, and PGS is not only a modest increase in average classification performance but also the identification of largely non-overlapping high-score groups. When comparing different score types for the same disease, correlations between PGS and either PXS or PSS were weak on average (mean $r \approx 0.04$ and 0.03 , respectively), and there was minimal overlap among high-score groups (PGS and PXS 0.94%; PGS and PSS 1.0%; PXS and PSS 1.6%; triple overlap 0.18%). The weak correlations between score types suggest that genetic, environmental, and social risk capture distinct dimensions of disease susceptibility. Accordingly, reliance on any single domain will systematically miss high-risk groups. Although improvements in AUC across single scores were modest, integrated models substantially altered risk classification patterns, suggesting that exposure and social information may identify individuals who would not be recognized through genetic testing alone. These values should not be interpreted as outcome-based clinical utility estimates because the present analysis is cross-sectional and uses study-defined probability thresholds. However, they do suggest that considering standardized exposure and social-risk information in a clinical setting can change which patients are flagged for closer evaluation.

A key aspect of this study is the deliberate separation of intervenable physical exposures (PXS) from structural and social factors (PSS), reflecting distinct clinical and public-health intervention pathways. Whereas structural factors such as education level, income, and housing type are strongly shaped by upstream societal forces and are often not readily modifiable at the individual level^{21–23}, many physical and behavioral exposures can be altered through workplace protections or home/environmental remediation^{24–28}. Building separate scores improves interpretability and helps align risk stratification with feasible actions (e.g., targeted lifestyle counseling, occupational education on personal protective equipment, or trigger mitigation), in contrast to aggregating heterogeneous determinants into a single, difficult-to-translate composite. Thus, an “elevated risk” label may warrant a different response depending on whether risk is primarily exposure-mediated, structurally mediated, or inherited^{29,30}.

The asthma phenotype illustrates how exposure- and social-risk scores can inform tailored intervention strategies despite similar overall disease risk. For asthma, PXS and PSS are potentially clinically useful: a high asthma PXS can guide targeted assessment and mitigation of triggers such as household dampness or mold proxies, irritant exposures, and occupational triggers^{31–33}; a high asthma PSS can flag structural constraints that shape exposure burden and impede control, prompting interventions such as care navigation, housing-related referrals or remediation support, and improve medication access and adherence³⁴.

The 66-question panel is a PEGS-derived prototype designed to balance clinical feasibility with exposomic coverage rather than a finalized universal clinical instrument. Our proposed framework is intended to complement, not replace, a well-taken social history. Clinicians already ask about housing, occupation, diet, substance use, and other social factors, but the data are variably collected and structured and are often not retained in a form that supports quantitative, cross-disease risk assessment. The added value of PXS/PSS is thus standardization, computability, and reuse across encounters. While these domains are present in routine care, the proposed framework is intentionally broader than many existing clinical screening approaches that focus on a narrow set of social determinants and lifestyle variables. These scores incorporate residential and occupational characteristics that can act as proxies for multiple co-occurring exposures and reveal intervention opportunities not otherwise considered in routine care.

This study has several limitations relevant to clinical translation. The outcomes and exposures were self-reported, introducing the potential for misclassification, recall bias, reverse causality, and post-diagnosis behavior change. Accordingly, these findings should be interpreted as the cross-sectional classification of self-reported disease history rather than causal inference or prospective prediction. This is especially true as the study cohort is based in one state; further follow-up should be conducted in biobank data. Future studies should evaluate incident outcomes and incorporate data on diagnosis timing in analyses. Future work should also determine whether PXS and PSS provide value beyond established clinical risk assessment frameworks because the baseline model used for comparison primarily includes demographics and clinical covariates rather than disease-specific clinical risk models. Additionally, some individuals with a high PXS, PSS, or PGS were not classified as cases. This may be because sufficient time has not passed for a disease to progress after chronic exposure^{35–37} or due to exposure to protective environmental or genetic factors not captured in the PEGS data^{38–41}. Importantly, individuals with

a high PXS, PSS, or PGS not diagnosed with a disease are targets for screening and intervention, especially if they are younger than the average age of disease onset.

Although the proposed 66-question assessment retained most of the discrimination observed in disease-specific models, substantial disease-specific exposure heterogeneity suggests that additional questions may be required for broader applications. External validation, recalibration, and adaptation across populations and care settings will be necessary before implementation.

Additionally, PXS and PSS are intended to function as structured prompts for targeted follow-up rather than stand-alone diagnostic tools. Future implementation studies should evaluate how exposure and social-risk information is collected, displayed, and stored within electronic health records while balancing clinical utility with patient privacy and disclosure preferences.

In summary, we demonstrate that questionnaire-derived PXS and PSS provide information complementary to polygenic risk for common disease classification, and, critically, that these score types identify largely distinct high-risk subgroups. This independence is clinically relevant because it supports a unified, multi-domain framework in which elevated risk can be matched to the most appropriate next step: exposure mitigation, structural support, and/or additional clinical evaluation. By proposing a minimum exposure assessment framework, we provide a starting point for eventual integration of exposomic information into routine healthcare, pending external and prospective validation. These results support prospective evaluation of standardized exposure assessment as a complement to routine history-taking and phenotype-specific clinical care, rather than immediate replacement of existing workflows.

Methods

Study Cohort

The Personalized Environment and Genes Study (PEGS) is a large cohort of demographically diverse individuals in North Carolina. Unlike many studies, which focus on a single disease or environmental factor, PEGS gathers a wide range of information on genetics, diet, lifestyle, medications, diseases, and environmental exposures through questionnaire, genomic, and geographic data to help researchers understand how these factors collectively influence human health.

The 19,445 enrolled participants are demographically similar to North Carolina's population. About two-thirds are female (62.3%), and one-third are male (37.6%). The cohort is racially and ethnically diverse, with approximately two-thirds of participants identifying as White (63%), over a quarter as Black (27.6%), and 4.5% as another race. About 5% of participants identify as Hispanic (5.1%). The average age of participants is 50.2 years. This diversity allows researchers to study health disparities and ensure that findings are widely applicable to different populations. We used PEGS Data Freeze 3.1 in this analysis. Cohort demographics can be found in Table 1.

Exposome Information

PEGS participants complete three surveys comprising questions on health, medications, occupational and hobby exposures, lifestyle information, and social and demographic factors. The Health and Exposure Survey (N = 9,449) requests information on general demographics, disease status, family history, lifestyle factors, and social and occupational exposures; the External Exposome Survey (N= 3,618) collects data on household, occupational, and hobby exposures; and the Internal Exposome Survey (N=3,071) is focused on internal exposures such as diet, medication, physical activity, stress, sleep, and infections. Participants also provide biospecimens such as blood, urine, stool, saliva, nasal cells, hair, nail clippings, baby teeth, and household dust. Whole-genome sequencing (WGS) has been completed for 4,737 individuals. The data components available for the PEGS cohort can be found in Table S4. For the present study, we used data from individuals who completed all three surveys.

Phenotype Information

We determined case-control groups for phenotypes based on self-reported, ever-diagnosed disease status. These outcomes represent self-reported history of ever having been diagnosed by a clinician. Phenotypes were selected based on adequate case counts and span heterogeneous clinical contexts relevant to patient-reported exposure assessment. Specifically, the panel includes conditions that differ in diagnostic paradigm, including phenotypes commonly anchored by objective clinical measures (e.g., hypertension, high cholesterol, type 2 diabetes, iron-deficient anemia), conditions in which symptoms and clinical history remain central to diagnosis (e.g., asthma, migraine), and structural or anatomic findings with variable downstream clinical consequences (e.g., fibroids, ovarian cysts, lower GI polyps). Table S5 outlines the case and control information for each disease phenotype.

Whole-genome sequencing data

The Broad Institute performed WGS for 4,737 PEGS participants with the most complete survey data. The reads were aligned to the hg38 human reference assembly, and joint genotyping was done using the Genome Analysis Tool Kit (GATK). For quality control, we used several tools to perform consistency checks on self-reported and genotype-inferred sex and ancestry. To ensure the quality of the data, a rigorous filtering process was applied to the sequenced samples. We excluded seven participants whose samples may have been derived from a mixture of DNA from multiple individuals or with potential aneuploidies. To avoid confounding from familial relationships within the cohort, we removed 120 participants who were second-degree relatives or closer. Finally, we identified and removed three pairs of duplicate samples. We also analyzed the raw data with DeepVariant to ensure accuracy. The final genotypes were a consensus of identical results from both DeepVariant and GATK, along with the highest-quality variants identified by GATK. After this process, our final dataset included 4,607 participants with high-quality WGS data. Details of the WGS data and its quality control and filtering are available. To ensure that only high-confidence samples were used in downstream analyses, we performed stringent filtering to exclude participants with discrepancies between their self-reported and WGS-inferred race, those identified as outliers in the WGS data, and individuals with a degree of relatedness. We also removed participants lacking key demographic data such as age or sex.

PXS and PSS Development

We classified exposures from PEGS questionnaires as physical or social based on their mode of impact and potential intervention strategies. Physical exposures include occupational exposures such as diesel or gasoline exhaust and dust, hobby-related exposures such as glue or lead fishing weights, and stress and mood factors that may trigger biological responses. These exposures are typically linked to personal behaviors, occupations, or activities and may be mitigated through interventions such as protective equipment, counseling, or a change in occupation. In contrast, social exposures, including parental and participant education level, household income, and housing type, reflect broader structural and societal influences that are less amenable to individual-level change. Given the distinct intervention pathways, we deemed it essential to develop distinct scores for intervenable physical exposures and non-intervenable social exposures for developing targeted interventions and translating PXS and PSS into clinical practice.

We built PXS and PSS using the strategy outlined in Figure 1B. We used data from participants who completed all three PEGS surveys. Individual missing exposures were imputed using the missForest R package. First, we created a derivation dataset that included all PEGS participants without WGS (n = 432). This dataset was used to select features for creating the PXS and PSS. Second, PEGS participants with WGS (n = 2,344) were used as a validation dataset to enable score comparisons. This held-out dataset was randomly split into a training set (n = 1,758) and a test set (n = 586) to evaluate classification metrics.

We created disease-specific PXS and PSS scores using features from the PEGS questionnaires. We used three machine-learning tools for feature selection: LASSO regression, random forest, and XGBoost (Figure 1C). For LASSO, features with non-zero coefficients were retained. For random forest and XGBoost, the mean feature importance score was calculated, and all features with a feature importance score over the mean value were retained. We then used a feature selection voter to select the final features for a score; if two of the three machine-learning models retained a feature, it was included in the score. Features included in the final score were used as classifiers in a multiple logistic regression model, with their coefficients (betas) used as the weights that comprise the final score. We then calculated a final score (PXS or PSS) for each participant using a weighted sum of the scaled survey feature responses.

PGS Computation

PGS are available for 4,600 PEGS participants for 3,030 phenotypes. Our previous work describes in detail the computation of PGS in the PEGS cohort. Briefly, to calculate PGS for all PEGS participants with WGS data and ensure a high number of variants for analysis, genome-wide imputation was performed using the TOPMed imputation server. After quality control, there were over 2.5 million variants from 4,600 participants. PGS were calculated using models from the Polygenic Score Catalog, and the genetic data were harmonized to ensure consistency. To avoid ambiguity with certain types of genetic variants and differences between sexes, the analysis was restricted to variants on autosomes. PLINK 2.0 software was used to compute PGS for each participant. Palindromic variants were removed, and data were aligned to match the PGS models. Of 3,490 possible PGS models, 3,030 were computed for 4,600 participants, excluding those with less than 70% of the variants in the model.

For some phenotypes, the PGS Catalog contains multiple PGS. In these cases, all PGS with similar phenotype definitions as those used for PEGS were calculated for PEGS participants, and

the PGS with the highest AUC was chosen for comparison. For example, the PEGS definition of type 2 diabetes is not based on HbA1c, so PGS for HbA1c measures were not considered. If multiple PGS for type 2 diabetes remained after eliminating those based on HbA1c, the PGS with the highest AUC was retained for comparison.

Score Analysis

To quantify the classification accuracy of PXS, PSS, and PGS for self-reported, ever-diagnosed disease status, we divided the validation dataset into training and testing sets. To compare PXS and PSS, we limited our analysis to participants with available WGS data who completed all three PEGS surveys. We then assessed the classification performance of the scores for each phenotype (i.e., asthma, high cholesterol, fibroids, hypertension, iron-deficient anemia, lower GI polyps, migraine, ovarian cysts, and type 2 diabetes) in the test dataset. Depending on phenotype, covariate adjustment included age (fibroids/ovarian cysts) or age+sex (all other phenotypes). We fit separate models for each score (PGS, PXS, PSS), pairwise combinations, the combined model including all three scores, and a covariate-only model. The model combinations were as follows: age+sex, PXS, PSS, PGS, PXS+PSS, PXS+PGS, PSS+PGS, PXS+PSS+PGS.

We quantified classification performance using the area under the receiver operating characteristic curve (AUC) computed on the test set from predicted probabilities output by each logistic regression model using the pROC R package⁴². We calculated ROC curves and AUC using an empirical ROC estimator. To quantify uncertainty, we constructed nonparametric bootstrap 95% confidence intervals for the test set AUC by resampling the test set with replacement (B = 1000 resamples; fixed seed). For each bootstrap sample, we recomputed AUC; we skipped resamples in which the outcome contained only one class. The 2.5th and 97.5th percentiles of the bootstrap distribution are reported as the 95% CI. We calculated AUC for adjusted (model score with age+sex or age alone) and unadjusted (score-only) models.

We then calculated McFadden's adjusted pseudo- R^2 on the test set for all adjusted and unadjusted models, as well as a model containing age+sex or age alone. McFadden's adjusted pseudo- R^2 measures the relative improvement of the log-likelihood of the fitted model over the null model, adjusting for model complexity. This represents a goodness-of-fit measurement of the model on the outcome. Importantly, this is not variance explained as a traditional R^2 value and is not rescaled from 0-1.

We examined the categorical NRI, comparing each model to a covariate-only base model. We computed NRI in the test set using probabilities for both the base and comparison models. Probability categories were defined by a sequence of cutpoints corresponding to the 50th to 95th percentiles (5% step) of the covariate-only predicted probability distribution in the test set. For each cutpoint, we estimated NRI using event/non-event reclassification tables derived from predicted probability categories under the base versus comparison models.

Calibration was assessed in the test set for both raw and recalibrated probabilities. We estimated calibration intercept and slope by fitting a logistic calibration model, regressing the observed outcome on the logit of determined risk:

$$\text{logit}\{\text{Pr}(Y=1)\}=\alpha+\beta\cdot\text{logit}(p),$$

where α is the calibration intercept, and β is the calibration slope. Overall accuracy was summarized using the Brier score, defined as the mean squared error between observed outcomes and predicted probabilities. For α , β , and the Brier score, we calculated bootstrap 95% confidence intervals by resampling the test set with replacement ($B = 1000$), excluding resamples with only one outcome class and taking percentile intervals. We also summarized calibration by comparing mean predicted probabilities and mean observed event rate in the test set on both the probability scale and the log-odds scale. To visualize calibration, we constructed decile calibration curves by grouping individuals into 10 bins based on predicted probabilities and plotting the mean predicted probability versus the observed event rate within each bin.

We then examined the associations between PXS, PSS, and PGS, their combinations, and respective diseases using the following regression model: Disease $\sim$ Score(s) + Age + Sex. We used a Benjamini-Hochberg false discovery rate of 10% to assess significance. We then calculated McFadden's adjusted pseudo- R^2 for all adjusted and unadjusted models, as well as a model containing age+sex or age alone.

For each disease, individuals were classified as high or low separately for PGS, PXS, and PSS scores using within-disease percentile thresholds. High scores were defined as values at or above the 90th percentile of the score distribution, and low scores as values at or below the 10th percentile. This approach enabled direct comparison of concordance and independence between PGS, PXS, and PSS classifications for each disease.

Minimum Exposure Question Set Development

To determine the minimum number of survey questions sufficient to accurately classify disease, we used the results from our disease-specific PXS/PSS feature selection. We examined the number of times a single questionnaire-based exposure was selected (from one to nine) across the nine diseases. We then constructed a single PXS or PSS to classify all diseases by including all questions that were selected at least once, at least twice, and incrementally up to the maximum number of times a single exposure was selected: seven for PXS and nine for PSS. For each PXS/PSS, we calculated the AUC, McFadden's pseudo R^2 , NRI, and calibration following the same procedure used for the disease-specific PXS and PSS. We then compared the AUC of each PXS or PSS to the AUC of PGS alone, calculating the difference between each value to assess how each number of exposures performed relative to PGS.

We then assessed the correlations between these exposures using the correlations calculated for the PEGS exposures⁴³. This allowed us to determine which exposures are both highly correlated and implicated in disease classification. This helped reduce the number of PEGS exposure questions for the proposed minimum exposure question set. To do so, we calculated the effective number of questions (Meff) from the total number of questions selected for at least one disease (M). We calculated the eigenvalues of the correlation matrix and employed the Kaiser rule by exposure category. The Kaiser rule retains the number of eigenvalues with a value greater than one within each category of exposures. We considered the final number of eigenvalues to be Meff.

We then used these results to determine a minimum exposure question set and test the classification performance of PXS and PSS built using these questions. We used the Kaiser rule to determine the number of effective features and thus the number of questions. We then selected the top N questions based on the number of times a question was selected across diseases. For example, if the Kaiser rule reported 10 effective features for occupational exposures, we selected the 10 questions assessing occupational exposures that were selected the most across diseases. This resulted in a single PXS created from 61 questions and a PSS created from five questions that can be used to classify multiple diseases. Using this final question set, we created a final, disease-agnostic PXS and PSS. With these PXS or PSS, we calculated the AUC, McFadden's pseudo R^2 , NRI, calibration, and associations for each disease as previously described.

To assess clinical implementation feasibility of the proposed minimum exposure question set, two clinicians independently reviewed the 66 candidate items using a structured rubric. Reviewers rated each item for likely capture pathway, respondent or staff burden, temporal stability, clinical actionability, privacy sensitivity, reviewer confidence, and pilot recommendation. Feasibility was rated on a four-level implementation scale: structured EHR likely, standard workflow collectable, proxy/linkage/custom workflow, or direct patient report essential. Because many PXS items represent behaviors, diet, home conditions, or occupational exposures that are not routinely captured in structured EHR fields, direct patient report was interpreted as a feasible capture pathway when paired with low burden and clear patient-facing wording. Agreement between reviewers was summarized using exact agreement and Cohen's kappa for categorical fields; ordinal fields were additionally reviewed for near-agreement. Invalid non-rubric entries were excluded field-wise from agreement calculations.

Ethics approval and participant consent

The present study used data from the Personalized Environment and Genes Study (PEGS), an ongoing cohort study of adults in North Carolina conducted by the National Institute of Environmental Health Sciences (NIEHS), a part of the National Institutes of Health (NIH). All PEGS participants provided written informed consent for participation, including the collection and use of questionnaire data, biospecimens, genetic and genomic data, geospatially derived exposure information, and related health information for research. PEGS participants also consented to broad sharing of their data with qualified researchers and data repositories for relevant research, subject to appropriate privacy protections, controlled-access review, and data use agreements. 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; ClinicalTrials.gov identifier NCT00341237). The current analyses used de-identified data from PEGS Data Freeze 3.1. No individual participant is identifiable in the reported results.

Data Availability

Individual-level PEGS data cannot be deposited in an unrestricted public repository because the cohort includes human participant questionnaire, health, biospecimen-derived, genetic, genomic, and geospatial information. PEGS data are shared through a controlled-access process described in detail in the PEGS Scientific Data descriptor⁴⁴. Briefly, PEGS maintains both a non-PII/de-identified dataset and a more restricted PII/geolocation-containing dataset, governed by

regulatory and compliant Data Use Agreements. The non-PII version includes individual-level survey, exposure, health, omics-derived, polygenic-score, metadata, data-dictionary, and supporting documentation resources. The restricted PII/geolocation version includes genomic data, residential geolocation details and related information needed to support additional geospatial linkages, when justified by the proposed research.

Qualified academic, governmental, nonprofit, and commercial researchers may apply for access through the NIEHS PEGS Access System at <https://www.niehs.nih.gov/research/atniehs/labs/crb/studies/pegs>. Applications specify the requested dataset version, requested data elements, proposed analyses, need for PII/geolocation data and external geospatial linkage, applicable ethics approvals, and data-security controls. Requests are reviewed by the PEGS Executive Leadership Committee for consistency with participant consent, NIH IRB approvals, scientific validity, necessity of requested data elements, privacy and re-identification risk, and adequacy of data protections.

Code availability

Analysis code used to generate the results, tables, and figures in this study is available at https://github.com/dtlloyd96/PEGS_Clinical_Exposure_Question_Set. The repository includes a README, software/package versions, and instructions for reproducing the analysis after obtaining approved access to PEGS Data Freeze 3.1.

Figures

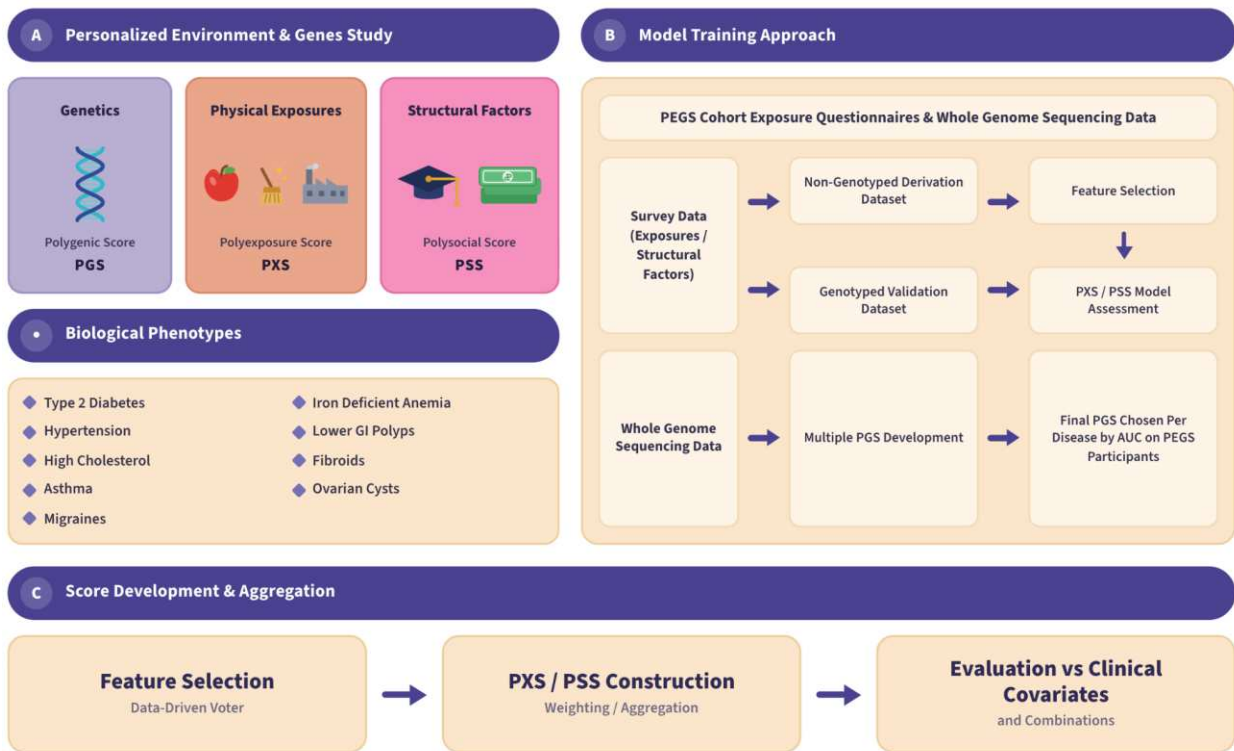


Figure 1: Overview of Personalized Environment and Genes study and score development

(A) Personalized Environment and Genes Study (PEGS) information collected with questionnaires and the phenotypes used for analysis. (B) Process of developing the PXS, PSS, and PGS. (C) Development of individual PXS and PSS by disease.

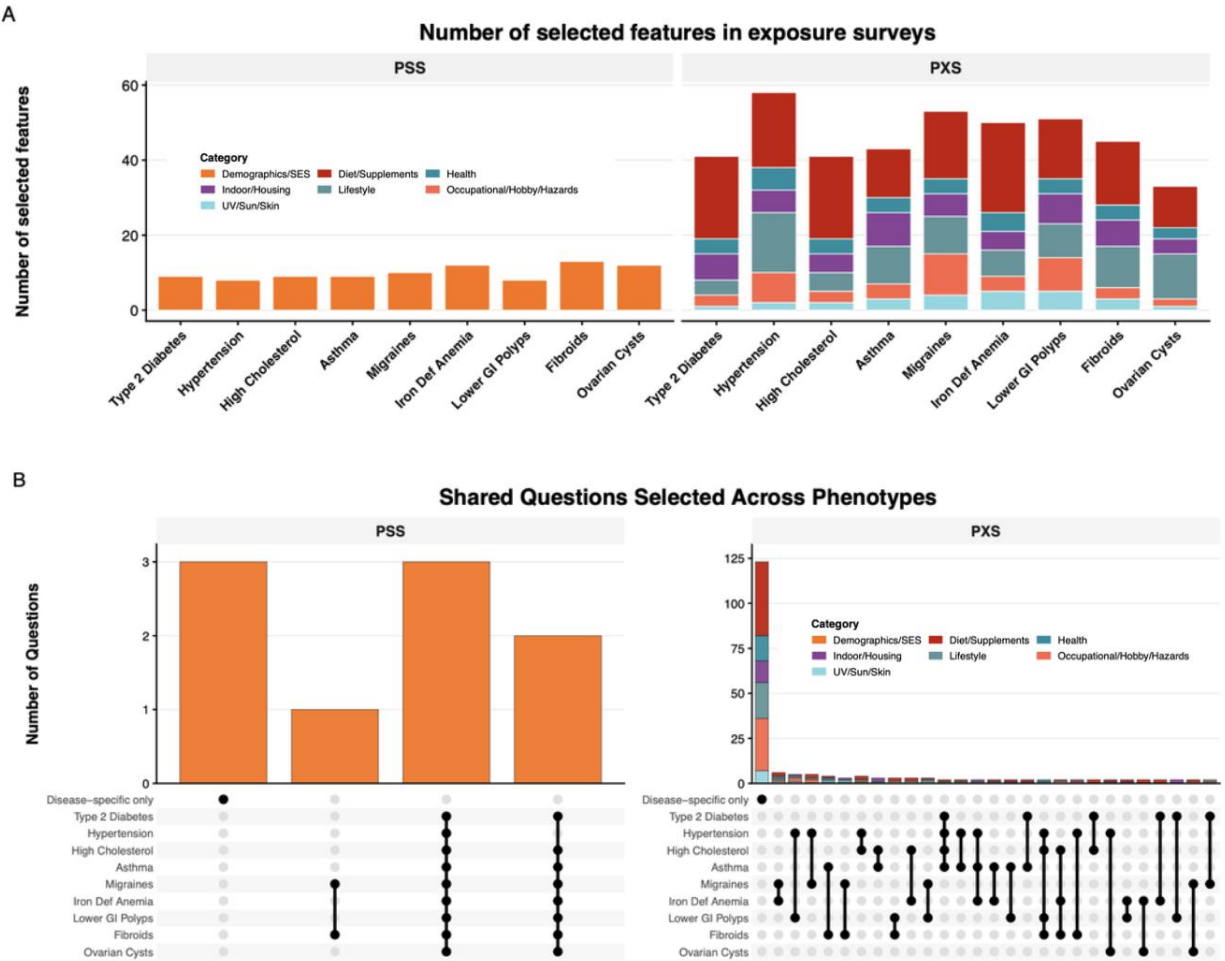
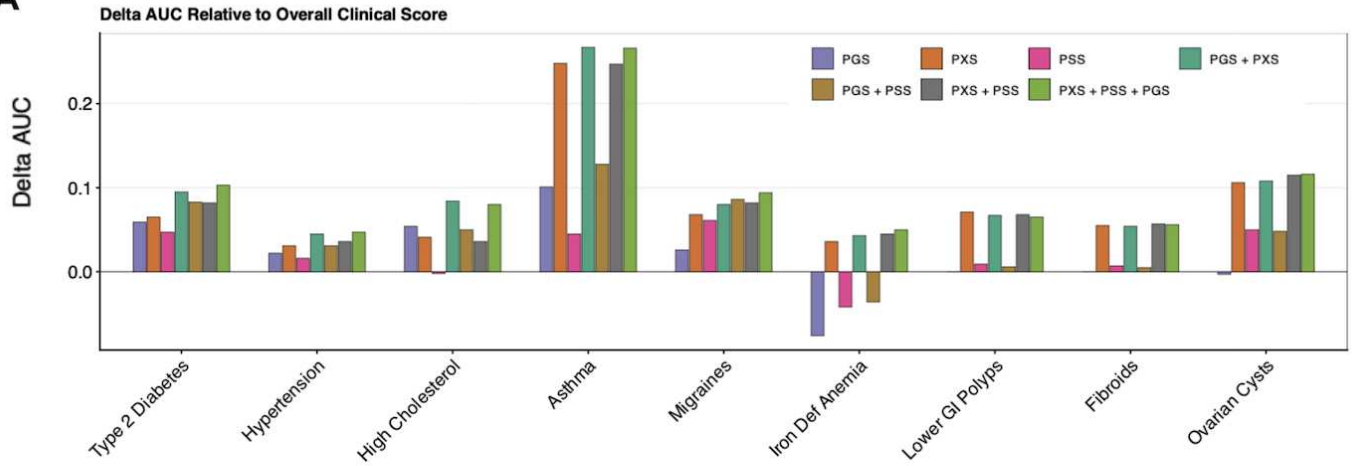


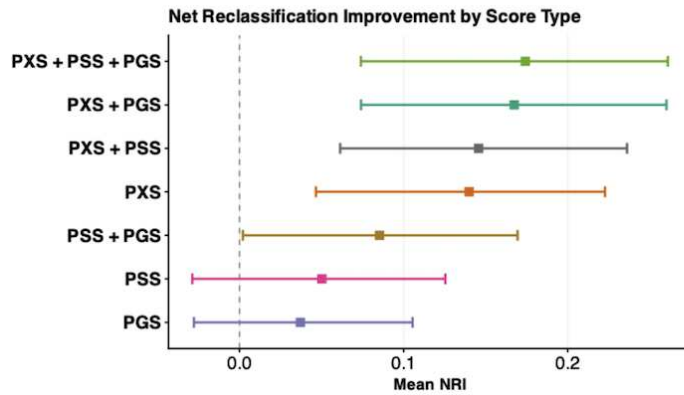
Figure 2: PXS Features

(A) Number of features chosen for the PXS. The Y-axis is the number of selected features by disease for each score in the data-driven voter, faceted by the developed score (PSS or PXS). The stacked bar colors show the proportion of exposures that fit into more specific exposure categories. (B) UpSet plots showing the overlap in exposures chosen for the PXS and PSS by disease outcome. The disease-specific bar notates exposures that were selected for PXS or PSS for a single disease.

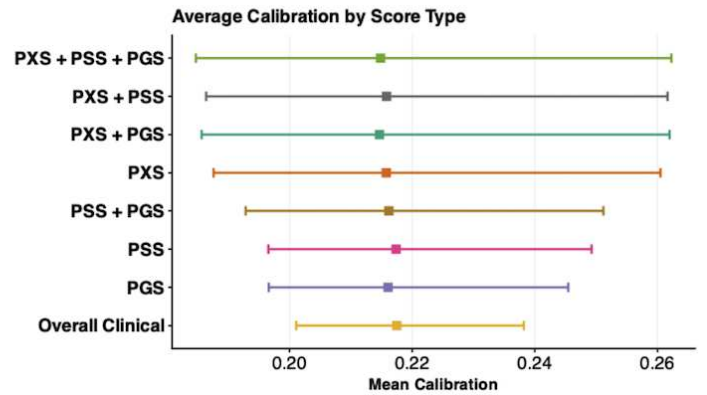
A



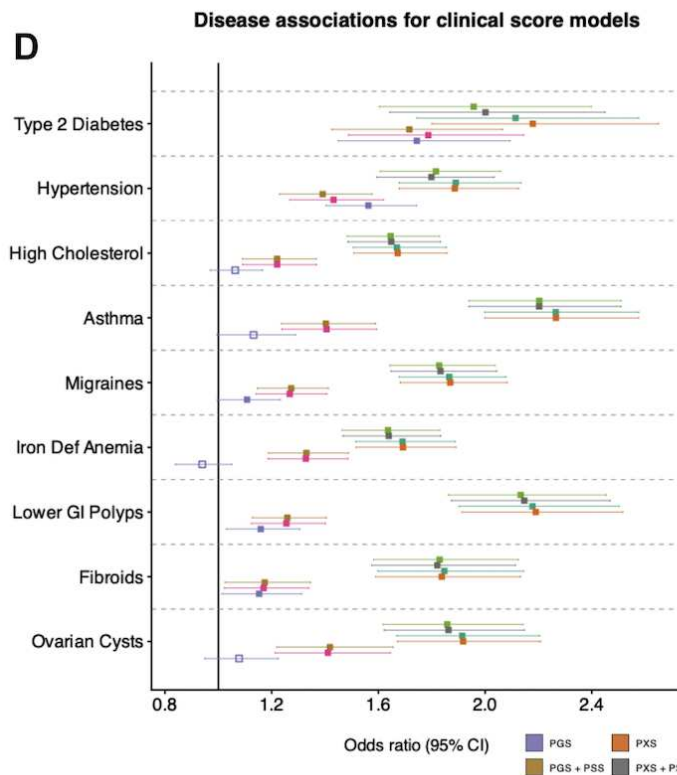
B



C



D



E

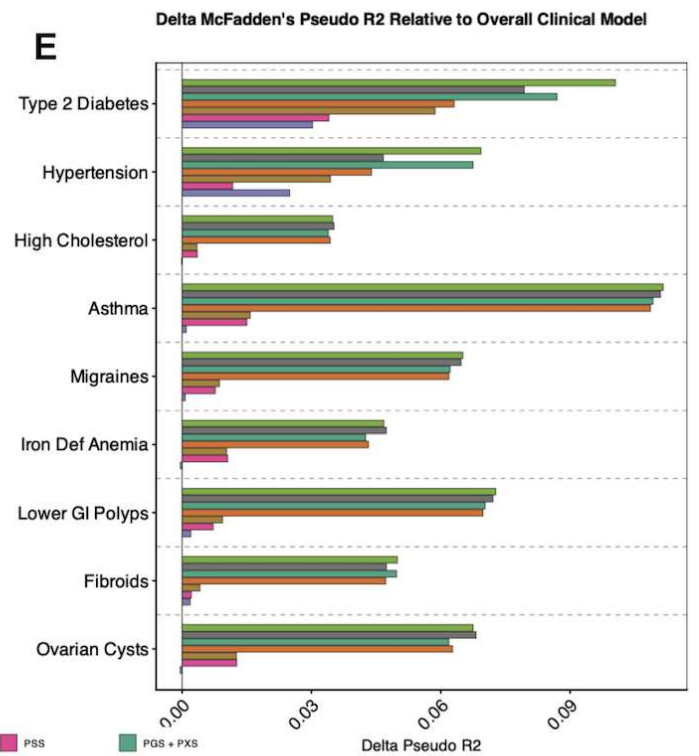


Figure 3: Disease Classification and Associations

(A) Comparison in Δ AUC between a covariate (age/sex) model and PGS, PXS, PSS, and their combinations. The X-axis is the disease, and the Y-axis is the Δ AUC over/under age+sex or age alone on the test set. Each bar is colored by score or score combination. (B) The average NRI over age+sex or age alone across predicted probabilities and phenotypes colored by score. The X-axis is the mean NRI, and the Y-axis is the score, ranked by average NRI. (C) Calibration by model type, pooled across predicted probabilities and disease. The X-axis is the mean calibration, and the Y-axis is the score. (D) The associations between each score and their combinations and diseases in the full validation set. The color of the odds ratio points indicates significance, with a solid point being statistically significant after adjusting for age and/or sex. The X-axis is the odds ratio and 95% confidence interval. Odds ratios per 1 SD increase in score, adjusted for age and/or sex. In multi-term models, the PXS or PSS (PSS+PGS) coefficient is plotted. (E) The Δ McFadden's pseudo R^2 over age+sex or age alone by outcome on the entire validation set. The X-axis is the Δ pseudo R^2 , and the Y-axis is the disease outcome.

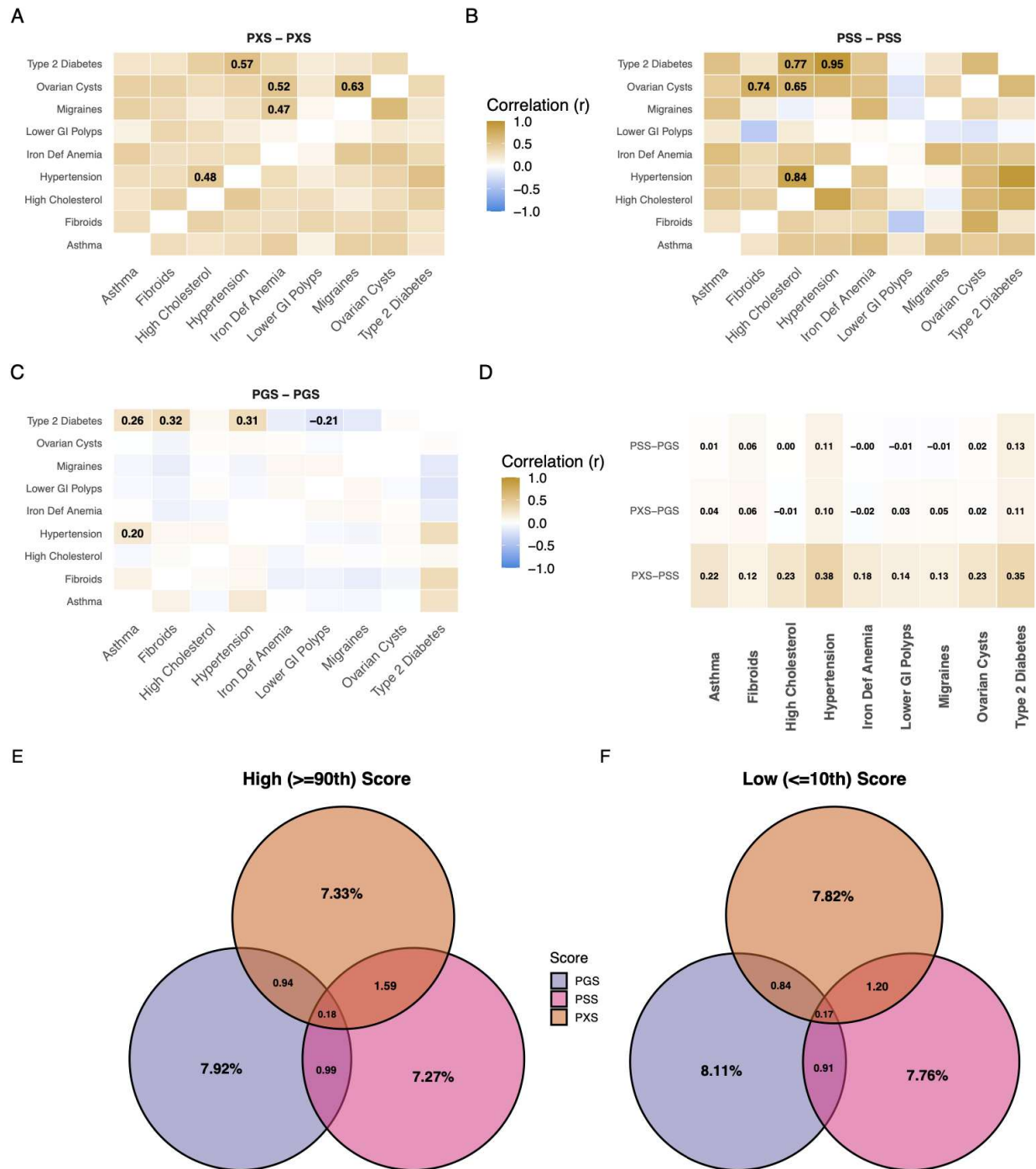


Figure 4: Score Populations

(A) Pearson correlations between PXS calculated across diseases. (B) Pearson correlations between PSS calculated across diseases. (C) Pearson correlations between PGS calculated across diseases. (D) Pearson correlations between PXS, PSS, and PGS for a single disease. The X-axis is the disease of interest, and the Y-axis is the score combination to which each correlation corresponds. (E) Population overlap among individuals with a score in the top 10%, pooled across diseases. Circles indicate the percentage of individuals in the top 10% in each

699 overlap category. All labeled numbers are percentages. (F) Population overlap among
700 individuals with a score in the bottom 10%, pooled across diseases. Circles represent the
701 percentage of individuals in the bottom 10% in each overlap category. All labeled numbers are
702 percentages.

703

704

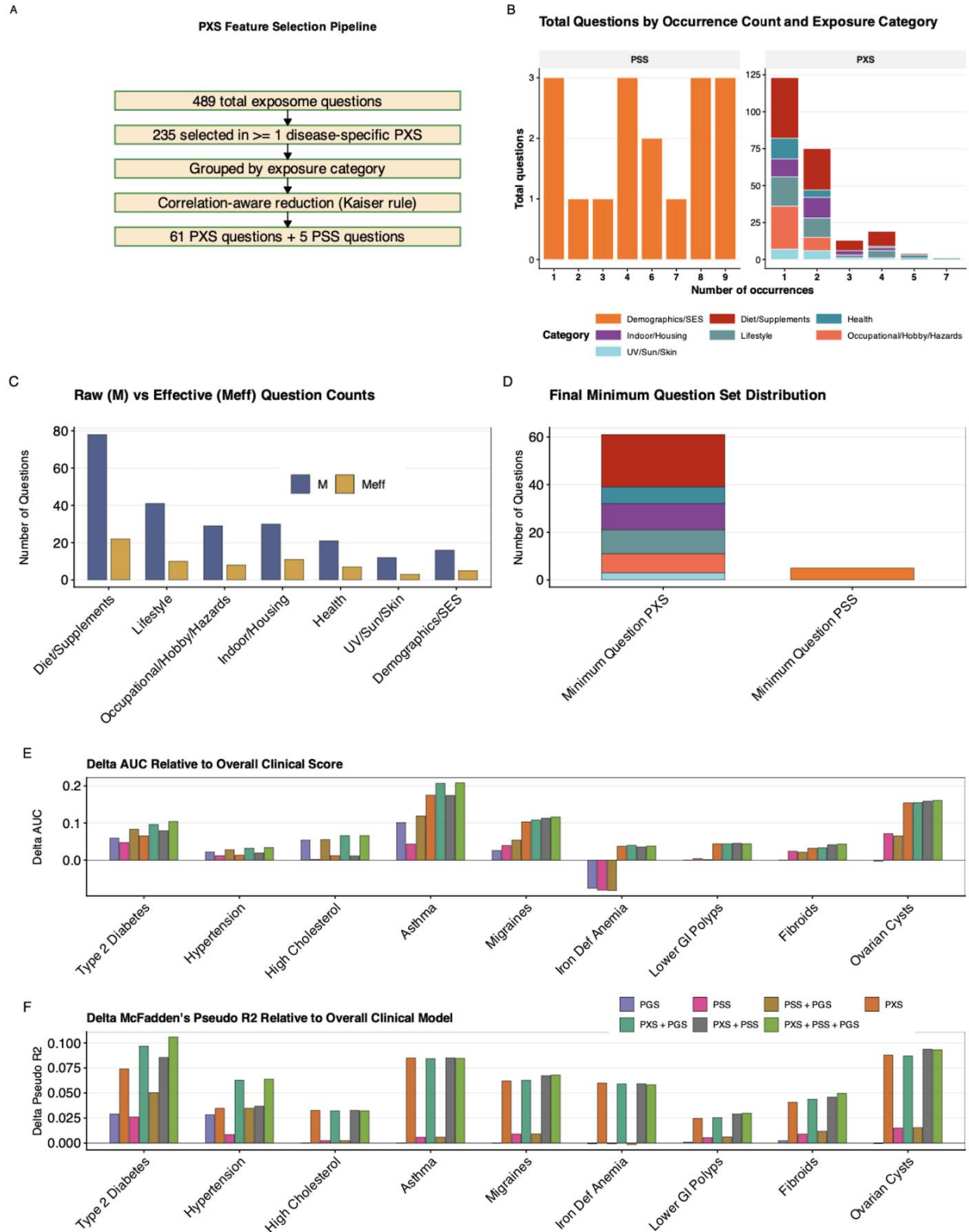


Figure 5: Minimum Exposure Question Set

(A) Feature selection pipeline for the development of the clinical exposure questionnaire framework. (B) The number of questions included in the PSS or PXS for N disease, colored by exposure category. The X-axis is the number of occurrences in the disease-specific scores, and the Y-axis is the number of questionnaire-based exposures in each bin. (C) Effective features calculated using the Kaiser rule compared to the original number of features. The X-axis is the exposure category, and the Y-axis is the number of questions in the raw (M) vs effective (Meff) feature categories. (D) Final number of questions by exposure category. The X-axis is the score category, and the Y-axis is the number of exposures in the minimum exposure question set. The bar is colored by the proportion of questions that fit into each exposure category. (E) Comparison in ΔAUC between a covariate (age/sex) model and PGS, PXS, PSS, and their combinations. The X-axis is the disease, and the Y-axis is the ΔAUC over/under age+sex or sex alone on the test set. Each bar is colored by score or score combination. (F) The $\Delta McFadden's$ pseudo R^2 over age+sex or age alone by outcome on the entire validation set. The Y-axis is the Δ pseudo R^2 , and the X-axis is the disease outcome.

Tables

Table 1: PEGS Demographics

Demographic Variable	Health and Exposure Survey	External Exposome Survey	Internal Exposome Survey
Number of participants	9426	3579	3034
Sex, N (%)			
Female	6317 (67)	2250 (70)	1927 (70)
Male	3103 (33)	986 (30)	843 (30)
Income, N (%)			
>\$80,000/year	2747 (30)		
<=\$80,000/year	6348 (70)		
Race, N (%)			
White	6711 (73)	2703 (84)	2334 (85)
Black	2055 (22)	338 (11)	272 (10)
Other	412 (5)	161 (5)	136 (5)
Ethnicity, N (%)			
Non-Hispanic/Non-Latino	8881 (98)	3435 (97)	2920 (93)
Hispanic/Latino	371 (2)	118 (3)	93 (7)
Age (years), mean (SD)	50.17 (15.96)	50.23 (14.34)	50.28 (14.34)
Body mass index (BMI), mean (SD)	28.65 (6.82)	28.02 (6.70)	27.91 (6.67)

References

1. Buck Louis, G. M., Smarr, M. M. & Patel, C. J. The exposome research paradigm: An opportunity to understand the environmental basis for human health and disease. *Curr. Environ. Health Rep.* **4**, 89–98 (2017).
2. Miller, G. W. & Banbury Exposomics Consortium. Integrating exposomics into biomedicine. *Science* **388**, 356–358 (2025).
3. Wild, C. The exposome: from concept to utility. *Int. J. Epidemiol.* **41**, 24–32 (2012).
4. Vermeulen, R., Schymanski, E. L., Barabási, A.-L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **367**, 392–396 (2020).
5. Figueroa, J. F., Frakt, A. B. & Jha, A. K. Addressing social determinants of health: Time for a polysocial risk score: Time for a polysocial risk score. *JAMA* **323**, 1553–1554 (2020).
6. Javed, Z. *et al.* Development and validation of a polysocial risk score for atherosclerotic cardiovascular disease. *Am. J. Prev. Cardiol.* **8**, 100251 (2021).
7. Zhao, Y. *et al.* Associations of polysocial risk score, lifestyle and genetic factors with incident type 2 diabetes: a prospective cohort study. *Diabetologia* **65**, 2056–2065 (2022).
8. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).
9. Akhtari, F. *et al.* Questionnaire-based polyexposure assessment outperforms polygenic scores for classification of type 2 diabetes in a multiancestry cohort. *Diabetes Care* **46**, 929–937 (2022).
10. He, Y. *et al.* Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of type 2 diabetes. *Diabetes Care* **44**, 935–943 (2021).
11. Javed, Z., Kundi, H., Chang, R., Titus, A. & Arshad, H. Polysocial risk scores: Implications for cardiovascular disease risk assessment and management. *Curr. Atheroscler. Rep.* **25**,

793 1059–1068 (2023).

794 12. Li, Y. *et al.* Polysocial and polygenic risk scores and all-cause dementia, Alzheimer's
795 disease, and vascular dementia. *J. Gerontol. A Biol. Sci. Med. Sci.* **79**, glad262 (2024).

796 13. Bunce, A. *et al.* Patient-reported social risks and clinician decision making: Results of a
797 clinician survey in primary care community health centers. *Ann. Fam. Med.* **21**, 143–150
798 (2023).

799 14. Chen, M., Tan, X. & Padman, R. Social determinants of health in electronic health records
800 and their impact on analysis and risk prediction: A systematic review. *J. Am. Med. Inform.*
801 *Assoc.* **27**, 1764–1773 (2020).

802 15. Kepper, M. M. *et al.* The adoption of social determinants of health documentation in clinical
803 settings. *Health Serv. Res.* **58**, 67–77 (2023).

804 16. Descatha, A., Fadel, M., Sembajwe, G., Peters, S. & Evanoff, B. A. Job-exposure matrix: A
805 useful tool for incorporating workplace exposure data into population health research and
806 practice. *Front. Epidemiol.* **2**, 857316 (2022).

807 17. Oladeji, O. *et al.* Air pollutant patterns and human health risk following the East Palestine,
808 Ohio, train derailment. *Environ. Sci. Technol. Lett.* **10**, 680–685 (2023).

809 18. Krah, P. L. *et al.* Advances in comprehensive exposure assessment: Opportunities for the
810 US military. *J. Occup. Environ. Med.* **61**, S5–S14 (2019).

811 19. Woskie, S. R. *et al.* Burn pit exposure assessment to support a cohort study of US veterans
812 of the wars in Iraq and Afghanistan. *J. Occup. Environ. Med.* **65**, 449–457 (2023).

813 20. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human
814 populations. *Nat. Commun.* **10**, 3328 (2019).

815 21. Smeeding, T. M. Multiple barriers to economic opportunity for the 'truly' disadvantaged and
816 vulnerable. *RSF* **2**, 98–122 (2016).

817 22. Corak, M. Income inequality, equality of opportunity, and intergenerational mobility. *J. Econ.*
818 *Perspect.* **27**, 79–102 (2013).

- 819 23. Torche, F. Analyses of intergenerational mobility: An interdisciplinary review. *Ann. Am.*
820 *Acad. Pol. Soc. Sci.* **657**, 37–62 (2015).
- 821 24. Bendall, C. L. *et al.* Central obesity and the Mediterranean diet: A systematic review of
822 intervention trials. *Crit. Rev. Food Sci. Nutr.* **58**, 3070–3084 (2018).
- 823 25. Yeon, J. H. & Shin, Y. S. Effects of education on the use of personal protective equipment
824 for reduction of contamination: A randomized trial. *SAGE Open Nurs.* **6**,
825 2377960820940621 (2020).
- 826 26. ASGE Technology Committee *et al.* Minimizing occupational hazards in endoscopy:
827 personal protective equipment, radiation safety, and ergonomics. *Gastrointest. Endosc.* **72**,
828 227–235 (2010).
- 829 27. Morgan, W. J. *et al.* Results of a home-based environmental intervention among urban
830 children with asthma. *N. Engl. J. Med.* **351**, 1068–1080 (2004).
- 831 28. Laumbach, R. J. & Cromar, K. R. Personal interventions to reduce exposure to outdoor air
832 pollution. *Annu. Rev. Public Health* **43**, 293–309 (2022).
- 833 29. van Daalen, K. R. *et al.* Risk estimation for the primary prevention of cardiovascular
834 disease: considerations for appropriate risk prediction model selection. *Lancet Glob. Health*
835 **12**, e1343–e1358 (2024).
- 836 30. Norland, K., Schaid, D. J., Naderian, M., Na, J. & Kullo, I. J. Associations of self-reported
837 race, social determinants of health, and polygenic risk with coronary heart disease. *J. Am.*
838 *Coll. Cardiol.* **84**, 2157–2166 (2024).
- 839 31. O'Connor, G. T. *et al.* Early-life home environment and risk of asthma among inner-city
840 children. *J. Allergy Clin. Immunol.* **141**, 1468–1475 (2018).
- 841 32. Ahmad, M. F. *et al.* Pesticides impacts on human health and the environment with their
842 mechanisms of action and possible countermeasures. *Heliyon* **10**, e29128 (2024).
- 843 33. Eaves, L. A., Keil, A. P., Rager, J. E., George, A. & Fry, R. C. Analysis of the novel
844 NCWELL database highlights two decades of co-occurrence of toxic metals in North

845 Carolina private well water: Public health and environmental justice implications. *Sci. Total*
846 *Environ.* **812**, 151479 (2022).

847 34. Manuel, J. I. Racial/ethnic and gender disparities in health care use and access. *Health*
848 *Serv. Res.* **53**, 1407–1429 (2018).

849 35. Luo, H. *et al.* Long-term exposure to ambient air pollution is a risk factor for trajectory of
850 cardiometabolic multimorbidity: A prospective study in the UK Biobank. *EBioMedicine* **84**,
851 104282 (2022).

852 36. Gangemi, S. *et al.* Occupational exposure to pesticides as a possible risk factor for the
853 development of chronic diseases in humans (Review). *Mol. Med. Rep.* **14**, 4475–4488
854 (2016).

855 37. Kivimäki, M. & Steptoe, A. Effects of stress on the development and progression of
856 cardiovascular disease. *Nat. Rev. Cardiol.* **15**, 215–229 (2018).

857 38. Mentis, A.-F. A., Dardiotis, E., Efthymiou, V. & Chrousos, G. P. Non-genetic risk and
858 protective factors and biomarkers for neurological disorders: a meta-umbrella systematic
859 review of umbrella reviews. *BMC Med.* **19**, 6 (2021).

860 39. Yu, M. G. *et al.* Protective factors and the pathogenesis of complications in diabetes.
861 *Endocr. Rev.* **45**, 227–252 (2024).

862 40. von Mutius, E. & Smits, H. H. Primary prevention of asthma: from risk and protective factors
863 to targeted strategies for prevention. *Lancet* **396**, 854–866 (2020).

864 41. Yeh, P.-K., An, Y.-C., Hung, K.-S. & Yang, F.-C. Influences of genetic and environmental
865 factors on chronic migraine: A narrative review. *Curr. Pain Headache Rep.* **28**, 169–180
866 (2024).

867 42. <https://cran.r-pr...ROC/pROC.pdf>.

868 43. Lloyd, D. *et al.* Interactive data sharing for multiple questionnaire-based exposome-wide
869 association studies and exposome correlations in the Personalized Environment and
870 Genes Study. *Exposome* **4**, osae003 (2024).

871 44. Akhtari, F. S. et al. Personalized Environment and Genes Study (PEGS) Dataset-a
872 resource for genomic, exposomic, and geospatial data. *Sci Data* 13, 787
873 (2026). <https://doi.org/10.1038/s41597-026-07011-x>

874

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PEGSPXSSupplement07062026.pdf](#)
- [RSMotsingerReif.pdf](#)