# The architecture of exposome-phenome associations

Chirag J Patel[1*], John PA Ioannidis[2,3], Arjun K Manrai[1]

1 Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
2 Departments of Medicine, of Epidemiology and Population Health, and Biomedical Data Science, Stanford University, Stanford, CA, USA
3 Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA

*Correspondence to Chirag J Patel

Conflicts: none

Keywords: exposome, phenotypes, phenome, ExWAS, risk factors, atlas

# Abstract

Non-genetic exposures—including nutrients, lifestyle factors, consumables, and pollutants—substantially contribute to phenotypic variation. Most studies assess only a few exposures or phenotypes, yielding fragmented exposome-phenome relationships. Systematic approaches are needed to quantify how the exposome—the totality of environmental exposures—relates broadly to clinically relevant phenotypes. We developed a resource benchmarking the exposome's role using data from the National Health and Nutrition Examination Survey (NHANES), cataloging 619 exposures and 278 phenotypes, and systematically testing associations (Phenotype-exposure-wide association study [P-ExWAS]). Among ~119k associations, 5% (n=5,661) were Bonferroni significant, and 40% replicated across independent population samples. Single exposures explained modest variance (median $R^2$=0.5%; interquartile range [IQR]: 0.27–1.10%). Twenty simultaneous exposome factors increased median variance explained to 3.5% (IQR: 1.8–7.8%), comparable to 1M genetic variants. The exposome-phenome atlas is available at: http://apps.chiragjpgroup.org/pe_atlas/.

# Introduction

Phenotypic trait variation and disease is caused by both genetics and environmental exposures[1–3]. It is a challenge to quantify the role of the *exposome* [4,5], the totality of environmental exposures, including physical, chemical, psychosocial factors humans encounter[6]. Genomic research has moved on from studying single genetic factors and variants by leveraging novel assay technology and genome-wide approaches to identify genetic risk factors that are reproducibly associated with disease with high efficiency [7,89]. With the explosion of biobank sized samples, researchers now have refined estimates of the genetic architecture [8]. The exposome is a conceptual complement of the genome, but to-date it has been more difficult to tackle in the same massive fashion as the genome. The exposome consists of an array of environmental exposure measures from broad domains[4], such as biomarkers of exposure to extrinsic factors such as pollutants (e.g. heavy metals or hydrocarbons);  self-reported behavior and lifestyle, such as dietary intake, and smoking; biomarkers of behavior and lifestyle, such as biochemical assays of cotinine, nutrients, or Vitamin D; and/or assays of infectious disease agents.

Interrogating exposome-phenome relationships has been limited to studies that target a few candidate exposures and phenotypes. These candidate studies are presented selectively in millions of papers on claimed associations yielding fragmented and often biased snapshots of the exposome-phenotype maze [10]. While candidate approaches have been successful in identifying factors with large effects, such as smoking[11], these millions of papers to-date have not yielded robust associations [12] moreover, many reported results might be false positive [13]. For example, disciplines such as nutritional epidemiology have yielded numerous associations regarding single dietary factors and patterns in disease outcomes have been non-robust[14,15]. Analogous debates have been made in fields studying other domains of the exposome, such as environmental epidemiology [16,17]. Previous criteria to gauge causality, such as Bradford-Hill may not be readily applicable in new exposome epidemiology scenarios, e.g. if most of the true associations to be discovered have small effect sizes and not readily discernible biological plausibility[18], analogy, coherence and specificity, and there is no possibility to validate in experimental studies.

Nevertheless, currently large data-rich cohorts and biobanks are readily available and, furthermore, they contain biosamples and/or indicators of residence to link participant-level information with molecular and digital markers of the exposome at a large-scale[4,19]. Moreover, while causation may remain difficult to establish with certainty, the ability to explain the variance in some phenotype by considering diverse exposures that are not necessarily causative may have value on its own, e.g. for prognostic purposes.

There is a need for scalable resources to quantify exposure–phenotype associations, enabling cross-cohort comparison and improving future study design and power. Such a resource should comprehensively catalog associations, standardize analytical methods, assess replicability across independent samplings, and quantify the variance each exposure explains for individual and multiple phenotypes.

# Results

In brief, we developed a pipeline (Methods, Figure S1) to analyze data from participants of the National Health and Nutrition Examination Survey (NHANES)[20] in 10 serial cross-sectional surveys that were sampled in years 1999-2000, 2001-2002, 2003-2004, 2005-2006, 2007-2008, 2009-2010, 2011-2012, 2013-2014, 2015-2016, and 2017-2018 (Figure S1, Methods). We cataloged a total of 346 real-valued continuous phenotypes and 810 biomarkers or self-reported questionnaires responses that measure pollutant, dietary, infectious, or smoking-related exposures across all 10 surveys. See Supplementary Results for sample size and power. Figures S2-S4 describe the distribution of demographic characteristics (sex, age, ethnicity, education, income) for each association.

**Exposure-wide associations across the phenome**

We used survey-weighted regression to associate phenotypes with exposures under 9 different modeling scenarios that adjust for demographic and social attributes: the (1) main reported model, which consists of age, age$^2$, sex, income (household income index divided by the poverty level), ethnicity (5 groups) and education (3 groups: above high school, high school equivalent, and below high school), and survey year (e.g, 1999-2000, 2001-02, 2003-04, 2005-06, 2007-08, 2009-10, 2010-11, 2012-13, 2014-15, 2016-17 as a categorical) (2) base model, with no adjustments, (3) sex and survey year, (4) age and age$^2$ and survey year (5) sex, age, and age$^2$,

and survey year (6) ethnicity and survey year, (7) income, education, and survey year (8) age, age$^2$, sex, and ethnicity and survey year, and (9) age, age$^2$, sex, income, education, and survey year.

We scaled all continuous exposures and phenotypes by their SD (see Methods) and ran regression models to obtain standardized β-coefficients, p-values, and $R^2$ (Figures 1-4) Categorical exposures were compared to predefined reference groups. Significance was defined by a Bonferroni threshold ($\alpha \approx 4 \times 10^{-7}$) and the Benjamini–Yekutieli FDR.

The number of associations across all P-E associations that passed the Bonferroni threshold was 5,661 (5% of 119,521) (Figure 1A, blue shaded points; Figure S5). The p-value corresponding to a Benjamini-Yeukuteli correction of 5% was $5.2 \times 10^{-4}$ (Figure 1A, blue and red shaded points). The total number of associations that exceeded the FDR of 5% was 10,876 (13%). All summary statistics are in Data Table 1.

The total number of tests per phenotype (n=278) conducted was 16 to 654 (median of 406). The average percent of associations (per phenotype) that were Bonferroni significant was 5% (range of 0.02% to 20%). The most associations were found for serum bilirubin, waist circumference, and body mass index (20% of tests for these phenotypes were significant for over 640 total tests).

We observed a large range of associations by exposome or phenome category (Figure 1BC). For example, the anthropometric phenome category saw the highest number of associations (13% of phenotypes in this category had a Bonferroni significant association)(Figure 1B, Figure S5A). Of exposome variables, smoking and dietary/nutrient biomarkers were implicated in the most P-E associations, ~15% and 13% respectively (Figure 1C, Figure S5B).

**P-E associations with the lowest p-value replicate across cohorts**

We estimated the rate of “replication” or the number of times an association appeared in greater than 1 survey sampling. Across the atlas of associations, we found that an association (at a p-value threshold of 0.05) occurred in two surveys 5% of the time. To contrast, if a P-E association achieved FDR significance across all surveys (Figure 1) p-value significance was achieved in greater than 2 surveys at least 20% of the time (Figure S6). However, P-E replicated rates vary depending on the number of surveys available for a P-E association. Specifically, for FDR-significant P-E associations assessed in only 2 surveys were found in both surveys 39% of the time at p-value less than 0.05 (Figure S6). For associations that were at least FDR significant, the median $I^2$ was 0, 5, 0, 26, 6, 14, 14, 20, and 18% for associations in 2, 3, 4, 5, 6, 7, 8, 9, 10 surveys. See Supplementary Results for heterogeneity of association across survey samples (Figure S6-S7)

**Variance explained of the exposome**

We estimated the variance explained ($R^2$) attributable to the exposure variable (after subtracting the potential role of demographic attributes [Methods]) (Figure 2B-D). Demographic factors, including age, age$^2$, ethnicity, income, education, and sex explained a large range of overall phenotypic variation, from ~0 to 80% (Figure 2D, x-axis). In comparison, single exposures added a median of 0.14% (Figure 2A, D).

The median $R^2$ for all associations that were Bonferroni significant was 0.6% (IQR: 0.3 and 1%; 5th to 95th percentile 0.1% to 3.6%) (Figure 2BC). The median $R^2$ was 0.02% for non-significant associations. We observed a range of $R^2$ by exposure associations across domains of the phenome and exposome (Figure 2BC). For example, phenotypes in the inflammation category had exposures explain 3% of variance on average across all P-E associations (Figure 2B) that were Bonferroni significant. For exposures, pollutant factors explained on average 0-3% of variation across all phenotypes. Organochlorine exposures explained ~3% of variance on average across all phenotypes. On average, dietary biomarkers accounted on average 1% of variation; however, dietary factors measured through an interview explained on average 0.5% of variation in phenotype.

Second, we also estimated the $R^2$ due to the additive contribution of 20 exposures simultaneously. For phenotypes that had greater than 20 exposures associated at FDR-level of significance, we imputed exposure data where missing (see Methods). When considering 20 exposome factors simultaneously in 119 phenotypes, the median variance explained is 3.5% (interquartile range 1.8% to 7.9%), greater than the median $R^2$ for single exposures (Table S4). When considering all phenotypes with less than or equal to 20 exposures in the model, the median $R^2$ was 1.6% (interquartile range 0.7 and 3.5%) (Figure 1D, red points). The maximum multiple exposure $R^2$ estimated was 57% for triglycerides (Figure 1D, red points).

**A comprehensive atlas across the exposome and phenome**

Association sizes correspond to the amount of the change in 1 standard deviation (SD) of exposure with a change for a 1SD change for the log-transformed exposures (for continuous exposures) or versus the reference group (for categorical variables) ("adjusted beta", Figure 3). For associations between Bonferroni significant exposome factors and phenotypes (association sizes for a 1SD change in exposome factor), the 5th to 95th percentile range was -0.17 to 0.19 (0.03 to 0.24 in absolute values).

**Dense correlational web of the exposome**

Exposures exhibit a dense correlation web (Figure 4AB). The median correlation between exposures was 0.01 and the median absolute value correlation was 0.05 across all correlations. For exposure-exposure correlations that passed the Bonferroni threshold ($p < 0.05/201,265$), the median for Bonferroni-corrected correlations (alpha threshold of 2e-7) was 0.19 and the median absolute value correlation was 0.21 (Figure 4C); the interquartile range of Bonferroni-corrected correlations was 0.08 to 0.37 (Figure 4C) (0.11 to 0.38 for the absolute values). The 95th percentile reached a correlation of 0.69 (0.69 for the absolute values).

The dense correlational web is described for a sampling of exposures in a correlation globe (Figure 4AB, see methods). Figure 4A depicts 50 randomly selected correlations sampled from all 201,265 correlations. Figure 4B depicts exposures identified in their association with BMI and Hemoglobin A1C%. Correlations are only drawn between nodes (or exposures) whose absolute value of correlation was greater than 0.25; these

correlations are among the top 15% of the distribution of correlations (Figure 4C). Correlations for exposures associated with BMI and A1C% overall have larger correlations than a randomly selected subset (Figure, 4C, blue).

**Demographic adjustment influences inferred association sizes**

Next, we computed the difference between adjusted associations subtracted from univariate estimates (Figure 3) to assess the impact of adjustment. The average difference between a minimally adjusted and a fully adjusted model and each scenario was 0.01 (Figure 5A), demonstrating some bias due to demographic adjustment. The standard deviation sizes were largest for the fully adjusted minus univariate model (SD: 0.1), fully adjusted minus ethnicity model (SD: 0.1). The SD were least for the fully adjusted minus the age, sex, and ethnicity models and the fully adjusted minus the age, sex, and income/education models (Figure 5A). See the Supplementary results for findings regarding specific exposures.

For some associations, the sign of association was opposite depending on the demographic correction scenario. Of the significant findings, 932 out of 5194 (15% of total significant Bonferroni identified pairs) exhibited a switch of coefficient sign between the univariate model (a model with no demographic or social factor adjustment) and the adjusted model. For example, BMI and blood cadmium, had positive associations (e.g,, for an increase in exposure, there is linear increase in BMI); however, when controlling or adjusting for factors in the "main" model, the association becomes stronger (e.g., the standard errors are reduced) and opposite in direction (Figure 5B).

**Consistency of associations across exposure categories and sample matrices**

Self-reported dietary nutrients and variables dominate nutrient exposure assessment in epidemiological studies. Overall, we found that 1,452 phenotypes had Bonferroni significant associations with self-reported questionnaire derived nutrient variables; however, among these phenotypes, they exhibited a median $R^2$ of only 0.2%.

Next, we hypothesized that a dilution of correlation size was due to measurement noise. Self-reported dietary nutrients were assessed on two days. The median correlation across 69 dietary nutrients recall in day 1 vs day 2 was 0.36 (IQR range of 0.28 and 0.43). We estimated the correlations of associations (e.g., beta-carotene-P association on day 1 vs. beta-carotene-P association on day 2) (Figure S9). Across all levels of significance, we observed a 0.84 correlation between day 1 self-report vs. day 2 (Figure S9).

Dietary biomarkers, on the other hand, had a larger median $R^2$ of 1% across the 1101 phenotypes with Bonferroni significance, 5 times larger than their self-reported counterparts. We also estimated the correlation between biomarkers and their self-reported counterparts (e.g., day1/day2 average of trans beta carotene vs. serum trans beta carotene) (Figure S10). The correlations between self-reported and biomarkers were smaller, and we observed a Pearson correlation of 0.52. For those blood nutrient variables that were Bonferroni significant, we observed a correlation of 0.60.

Blood and urine pollutant biomarkers reflect the biological relationship between exposure and excretion. We observed a positive and strong correlation between associations. For example, for blood vs. urinary biomarkers, we observed a 0.72 Pearson correlation. When only considering blood biomarkers that were Bonferroni significant, the P-E associations between blood vs. urine biomarkers was 0.78 for cadmium, 0.96 for cotinine, and 0.71 for mercury (Figure S11).

**Shared associational architecture between exposures**

For a pair of exposures, shared “associational architecture” measures the correlation, or the similarity of their correlations across all phenotypes (Figure S12AB). For example, the correlation between associations for blood trans- vs cis-beta-carotene was 0.98; in other words, the association coefficients across the phenome were very similar for those two exposures. The associational architecture between serum cotinine and 3-fluorene had 0.90 correlation.

Exposures from the same category (e.g., dietary interview, organochlorine, VOC, dietary biomarkers) tended to have similar phenotypic associations; the degree of shared associational architecture is larger within categories than across categories. For all correlations within exposure variables that were dietary biomarkers, the median absolute value of shared architecture was 0.2 (IQR 0.01 to 0.35). Similar shared associational architecture was observed within smoking biomarkers (0.2, IQR 0.08 to 0.35). The shared associational architecture between dietary biomarkers and self-reported dietary nutrients had a median absolute value correlation of 0.24. We examined the degree of shared associational architecture between phenotypes. The shared associational architecture between BMI and body weight was 0.98. BMI and cardiorespiratory fitness had opposite associational architecture (the sign of the correlations were opposite between BMI and fitness), a correlation of -0.83. Hemoglobin A1C% had an opposite architecture compared to HDL-Cholesterol (correlation of -0.54). See Supplementary Results for shared associations between blood and urine measured exposures.

**Comparison with variance explained due to genetics from genome wide association studies**

We compared the variance explained due to ~1M imputed genotypes from genome-wide association studies performed on 29 of the phenotypes[21] examined here (Figure S13, S14; Table S4). Across the 29 phenotypes, the median incremental $R^2$ due to genetics was 7.9 % (IQR 2.8 % to 9.3 %; maximum 21 %) and the median incremental $R^2$ due to exposome (20 exposome variables across 39 phenotypes) was 7.9% (IQR 3.1% to – 12%; maximum 57%). We found that the multiple exposome factors, when modeled simultaneously, had explained variance comparable to the entire genomic array across the phenotypes (Figure S13). Specifically, 55% (n=16) phenotypes had higher exposomic versus genetic $R^2$ (Figure S13). For example, the total genetic (1M common SNPs) and exposomic (20 factors) variance explained for BMI was similar, ~10% for both.

# Discussion

Here, we develop an atlas of exposures of the *exposome* and quantitative continuous phenotypes of the *phenome* and an approach to comprehensively associate each individual exposure and phenotype, called “phenotype-by-exposome-wide association study” (P-ExWAS). To benchmark and contextualize associations

for future exposomic investigation, we provide our catalog and exposure-phenotype associations as a browsable atlas. The primary findings in developing the atlas includes the “complexity” of associations. Specifically, we identified many thousands of associations that pass FDR criteria of even more stringent Bonferroni adjustment. The association sizes and variance explained for most phenotypes are modest and they varied across the phenome.

Our broad findings have several implications. First, having exposome intra- and inter-domain effect sizes can guide study design. P-ExWAS across the domains can give a sense of the prior landscape for investigators, guiding exploration before embarking on assessing various exposure and phenotype hypotheses. The current atlas may be used to replicate or synthesize findings that emerge from new biobank investigations or bolstering existing cohorts. Moreover, often estimates of sample sizes for cohorts are arbitrary and most exposome studies are underpowered[22]; now, our atlas can provide some ballpark guidance for an empirical estimate of power and sample size calculations for associating specific domains.

Second, potentially many more independent exposures will be required to explain a larger share of phenotypic variation. When broadening to 20 multiple exposures simultaneously, variance explained can be significantly larger but, on average, remains moderate. For most situations where risk stratification of complex diseases is desired, single markers may seldom suffice. Demographic context remains indispensable: age, sex, and socioeconomic status alone explained a larger share (up to 80% in some phenotypes). The low variance explained by exposome factors in current models may be due to measurement error (especially for survey-derived information) that dilutes effect sizes, suboptimal ability of the used measurement tools to capture the most essential exposures, and limits in the value of cross-sectional single measurements to represent variables that evolve over time.

Whereas single environmental exposures seldom exceed 1 % variance explained in NHANES, we observed from published data[21] that variance explained from genome-wide common variants for some phenotypes are comparable to what is achieved with exposomic variables. However, for the majority of routine clinical

phenotypes, genetics and demographics together still leave >50 % of variability unexplained — underscoring the potentially complementary, rather than competing, roles of genetics and exposure information[1]. Caveats of this analysis include that the genetic $R^2$ are from individuals of European ancestry. The incremental $R^2$ from the exposome may become larger as the number of measured exposome factors increases with newer assays [23]. Exposome information may eventually be integrated in predictive models for precision public health type of applications in the same way as genome-wide predictive tools are being considered for the same purpose.

Third, another implication of our data is that most exposures are not "specific" in their association. A mainstay to aid in the inference of causal associations, one of Bradford Hill's criteria is specificity of associations across phenotypes: it is difficult to untangle causality if an exposure is broadly correlated with many phenotypes simultaneously. Most exposure-phenotype associations would fail the Bradford Hill's specificity criterion as most exposures were associated with many phenotypes across domains [13,24]. Moreover, some exposures within the same category are highly correlated and may have very similar phenotypic associations (e.g., cis beta and trans beta carotene) making it difficult to tell if one or several of them are important.

Fourth, data processing and confounding control, such as how to transform variables (e.g., log transform; inverse-variance normalize; break into categories, such as quantiles), and what variables to adjust for to control confounding may influence findings and results[25–28]. Regarding adjustments, bias considerations need careful thought. Environmental exposure associations are sensitive to age, sex, and socioeconomic status. Here we are exploring the space of possible and realistic ways of both processing and inclusion of adjustment variables, and an outstanding question remains how exposures may explain modifiable differences between demographic groups. Critically, the causal structure between exposures and phenotypes are also unknown - while demographic factors such as income and occupation are strongly correlated with exposure levels[29,30] and geospatial routes of exposure [31], a list of what variables to adjust for potential confounders is elusive and residual confounding remains a threat to assessing causality. On the other hand, over-adjusting for colliders can cause an opposite problem. We show the degree to which exposure-phenotypes are influenced by an *a priori* list of factors. Nevertheless, given an association between a phenotype and an exposure the Phenome-

Exposome Atlas can also depict the degree to which other exposure factors also correlate. Whether these additional exposure factors may be confounders or colliders requires careful consideration in each case.

We speculate the size of associations will also be dependent on demographic structure. Specifically, we claim that certain exposome variables will have associations that are more pronounced in populations with different case mix and differences in exposome architecture across populations may help refine clinical risk estimates.

Fifth, we observed differences in consistency on the type of exposome measure. Notably, we found high concordance between blood and urine measures of heavy metals (e.g., lead and cadmium). By contrast, diet recalls taken during 2 times explained only 0.2% of variance (e.g., beta-carotene quantity consumed as estimated from a food frequency questionnaire), while their biomarkers (e.g., blood beta-carotene) captured greater than 1%. Biomarkers are "objective", but also may be impacted by endogenous processes, such as metabolism.

There are several limitations in our deployment, with many directions to evaluate next. First, as above, given the currently limited definition of the exposome and its measurements, we are guaranteed to capture only a limited subset of the exposome-phenome architecture. There are numerous biobanks/cohorts with similar phenotype measures, and to some extent, exposure measures such as *Human Health and Exposure Analysis Resource* (HHEAR: https://hhearprogram.org/data-center) (Table 1).

Second, there are a few key differences to note across different resources. Many existing epidemiological studies and biobanks have only relatively few candidate self-reported instruments of exposure. As we have shown, these may tend to have small association effects compared with objective measurements. Moreover, some resources collect "geographical" measured exposures. Geographical estimates, such as air pollution markers measure exposure for a collective of individuals who live in the same location. The advantages of this type of measure includes a more direct connection to the source of exposure; however, association sizes will be lower than individual-level exposures [32] [1].

Third, and critically, cohorts with longitudinal, repeated measurements may give a longitudinal view of how exposures change through time and how these changes lead to disease outcomes or phenotypes. Longitudinal disease outcomes may be captured in different ways and algorithms (e.g., via International Classification of Disease Code or combination of biomarker and self-reported information). Several investigations have commenced in biobank samples to broadly interrogate exposures in longitudinal outcomes [33,34] and the degree to which these phenotypes and self-reported/geospatial exposures can be integrated needs to be evaluated. These resources may have additional limitations such as missing data and convenience sampling (e.g. on a volunteer basis), which may bias and hinder generalization of exposure-phenotype relationships.

Our atlas will complement existing efforts to document exposure phenotype relationships. A recent serum-only exposome mapping in a Chinese cohort (You et al.) [35] , prioritized breadth and population representativeness. Whereas You et al. profiled 267 blood chemicals – some of which interrogated here – in 5,700 volunteers and modelled cross-sectional disease risk, our atlas interrogates 619 exposures measured in blood, urine, and questionnaires against 278 continuous phenotypes across ten nationally (US) representative NHANES waves (~68K participants), enabling replication, variance-partitioning and direct comparison with polygenic architecture. We further examine measurement concordance (e.g., blood vs urine metals, recalls vs biomarkers) and provide open code plus an interactive database, features absent from the single-wave serum study. Together, the two resources illustrate complementary strengths—analytical precision for targeted chemicals versus scalable, multi-matrix surveillance—and underscore the need to couple large, longitudinal cohorts with high-resolution assays in future exposomics research. Other resources such as Exposome Explorer [36], provide a view of the published literature, and serve chiefly to provide concentration ranges, analytic methods, and short-term reproducibility, important to refine measurement. We view our atlas can complement databases such as Exposome Explorer and provide standards by which to catalog associations, such as documenting association sizes, sample sizes, and model types [37].

Newer consortia, such as the US *Human Health and Analysis Resource*, aim to enhance existing cohort data with new measures of the molecular exposome[4,5], such as small molecules that are measured via high-resolution mass spectrometry and/or targeted mass spectrometry, mostly at a single timepoint. The "targeted" mass spectrometry measures are like those assessed in this study, such as biomarkers of pollutants measured in blood and urine. High-resolution mass spectrometry provides a larger view of the wide array of exposures present in blood. Our approaches should generalize as the number of exposome variables with available measurements increases, and the distributions estimated here may be used to gauge the utility of future associations. However, the NHANES participant data are cross-sectional, and associations detected may be "reverse causal". Next generation atlases may consider both the stability of exposures and their relationship to longitudinal outcomes. To properly evaluate the role of the exposome in precision public health [38,39] and/or bridge gaps in the field, resources such as the *Exposome Phenome* atlas and future enhancements can offer helpful guidance.

# Online Methods

We systematically associated environmental exposures and phenotypes and created a data resource ("*Phenome-Exposome Atla*s") that provide a benchmark of summary statistics between exposures and phenotypes, leveraging participant data from the Centers for Disease Control and Prevention (CDC) National Health and Nutrition Examination Survey (NHANES). Figure S1 describes the pipeline workflow.

We developed a R statistical package ("*nhanespewas*") to conduct all analyses. The features of the *nhanespewas* includes (a) cataloging of phenotypes and exposures to associate within the NHANES surveys, (b) R package to associate all of the exposures with phenotypes using a survey-weighted linear model[40] and providing a user-specified array of modeling assumptions (e.g., potential confounder and covariate choice), (c) aggregating pairwise associations across independent surveys (a meta-analysis) to output an overall estimate across surveys and assess replicability, and (d) providing a browsable database (*Phenome-Exposome Atlas*) of summary statistics that contain the overall association size or correlation across all survey years interrogated, the standard error of the association, p-value, and variance explained attributable to the exposure

(and after consideration of potential adjusting covariates). The code and package is located here: https://github.com/chiragjp/nhanespewas. An introduction to the package is found here: pe_quickstart.Rmd (https://github.com/chiragjp/nhanespewas/blob/main/pe_quickstart.Rmd)

*Attaining and cataloging participant data from the National Health and Nutrition Examination Survey*

First, we download all participant data and data dictionaries, encompassing 10 surveys (approximately 10,000 total participants per survey) (Figure S1A). Participant data in NHANES are divided into different tables in 5 components (demographics, diet, laboratory, questionnaire, and examination). We downloaded a total of 1360 tables and stored all data in a sqlite database. See the code located in the path *download/download_nhanesa.R.*

We next categorized each variable downloaded into being a “phenotype” or an “exposure”. We defined exposures as exposures if they are (a) exogenous in origin (e.g., pollutants) (b) biomarkers of exposure (e.g., cotinine),(c) reflective of lifestyle or behavior (e.g., self-reported food intake). While many of these variables do exhibit heritability (e.g., have G-E correlation), it is secondary as the origin of the factor is still external.

We processed some phenotypic variables and exposure variables. Blood pressure was a phenotypic variable whose measurement was repeated multiple times: we took the average of the measurements. Physical activity questionnaire information was collected in a series of questionnaire items; we processed these measurements to be in total metabolic equivalent hours (see websites: [41–43]). We processed multiple smoking variables (see ref. 26) to be a variable that denotes past, current, or never smoking. We next tabulated the sample size for each exposure and phenotype variable pair, used in downstream pipelines. Our cataloging process is in the directory *select* and *pe_sample_size.R*. A tutorial on the catalog is found here: pe_catalog.Rmd (https://github.com/chiragjp/nhanespewas/blob/main/pe_catalog.Rmd).

To ease interpretation, we categorized each cataloged phenotype or exposure variable into different categories. The categories of phenotypic variables (n=278) that included anthropometric (n=11, e.g., height),

aging (n=2, e.g., telomere length), blood parameters (n=19, e.g., basophils number) , bone (n=3, total calcium), dexa (n=136, e.g. lean fat), electrolyte (n=5, e.g., bicarbonate), fitness (n=3, e.g., Vo2max), hormone (n=17, e.g., iodine), inflammation (n=4, e.g., C-reactive protein), iron (n=8, e.g., iron), kidney (n=8, e.g., urinary creatinine), lipids (n=7, e.g., Apolipoprotein B), metabolic (n=6, e.g., glucose), microbiome (n=16, e.g., observed OTU mean), nutritional status (n=13, e.g., folate), and PSA (n=4, e.g., prostate specific antigen ratio).

The categories of exposome variables (total n= 619) included volatile organic compounds (VOCs, n=64, e.g., nitromethane), amine/amide (n=5, e.g., acrylamide), diakyl (n=9, e.g., dimethylphosphate), dietary biomarkers (n=49, e.g., vitamin B12), nutrients from dietary interview (n=178, e.g., alpha-carotene), heavy metals (e.g., blood lead), hydrocarbons (e.g., 1-napthol), infection (n=31, e.g., urinary chlamydia), organochlorine (n=65, e.g., PCB199), organophosphate (n=5, e.g., Acephate), phenols (n=13, e.g., e.g. urinary tert phenol), phthalates (n=19, e.g., mono(carboxynoyl) phthalate), polyfluorinated compounds (n=15, e.g., 2-(N-Methyl-perfluorooctane sulfonamido) acetic acid), priority pesticides (n=19, e.g., 2,4,5-trichlorophenol (ug/L)), pyrethoid (n=1, e.g., oxypyrimidine), smoking behavior (n=9, e.g., are you a current, ever, or never smoker?), smoking biomarker (n=12, e.g., cotinine), supplements (n=83, e.g., number of supplements reported).

Example exposome and phenotype variables are in Table S1-S2; the entire list is in Data Table 2-3.

*Model assumptions and data processing*

The core functions of the package we now describe are available in the *quantpe.R* and *petable.R* (Figure S1A). We model the relationship between an exposure and phenotype pair in the population with the following linear model:

Pp = intercept + bi*Ei + error, where Ei is an individual environmental factor or exposure of the exposome, the intercept is the average value of the phenotype when the exposure value is 0, and the b term denotes the association between the exposure and the phenotype. The lowercase p and i indexes a specific phenotype exposure respectively. We assume the error is also normally distributed with zero mean and standard deviation

of 1. We modify this core model in a number of ways to consider multiple adjustment scenarios and transformation of the Pp and Ei variables.

We transform the exposure and phenotype variables to ensure the comparability of estimates (e.g., bi) across all pairwise P-E associations (Figure S1B). Our pipeline inputs phenotypes, such as weight, body mass index, glucose, are quantitative and continuous valued variables. We transformed the phenotype in three possible ways. In the first, all phenotype variables are "scaled" (mean subtracted and divided by the survey-weighted weighted standard deviation). In the second, we divided the continuous exposures into groups by their percentile (e.g., reference group: [0,25) percentile, [25, 50), [50, 75), [75, 100)). Our pipeline also allows for rank based inverse normal transform, commonly used in genome-wide association studies [44].

We transform the exposures (Ei) in the following ways:

(a) for continuous and blood or urine biomarkers of exposure, we log transform (base 10) and scale (mean subtracted and divided by the survey-weighted weighted standard deviation). To account for zero values before log10 transformation, we add a small constant equal to the smallest non-zero value.

(b) for categorical exposure variables, we chose the largest category as the reference group, and analyzed the categorical variable in the same model

(c) for ordinal variables, we kept as ordinals and analyzed as a numbered rank (e.g., 1-3).

*A pipeline to conduct survey-weighted associations*

The NHANES is a complex and multi-stage cross-sectional survey designed by the National Center for Health Statistics; the multi-stage survey design provides US-wide generalizability for phenotype-exposure associations. Because of the complex study design, unique analytics approaches must be employed to ensure accurate association mapping between exposure and phenotype (versus approaches that assume a simple random sample). NHANES is sampled hierarchically: the first stage of sampling consists of regions of the country, or primary sampling unit (PSU) that are selected proportional to the average age, sex group prevalence, race and income. The second stage of sampling includes segments and households. Segments

within the PSUs that are consistent with census blocks. The third stage of sampling includes individual participants. Each participant is also a weighted sample, such that it is proportional to the number of people in the population that that person represents by age, race, income, and non-response rate.

Our pipeline iterated through all combinations of phenotypes and exposures in the NHANES:

```
1 [surveys, phenotypes, exposures] = select_phenotype_and_exposure_pairs(phenotypes, exposures)
2         For each Pp [phenotypes]:
3                 For each Ei [exposures]:
                          surveys = get_surveys_for_pheno_expo(Pp, Ei)
5                         data = get_expo_pheno_tables(Pp, Ei, surveys)
6                         weighted_data = figure_out_weight_across_surveys(data)
7                         weighted_data = transform_exposure_phenotype(weighted_data)
8                         weighted_data = create_survey_design(weighted_data)
9                         model = run_model(Pp, Ei, covariates, weighted_data)
10                        stats = summary_stats(model)
```

The NHANES participant data is organized into multiple tables based on properties of the measures that include the type of measure (e.g., dietary indicator or biomarker of exposure) on a subsample of the overall sample (FIgure S1A). Specifically, demographic variables, such as age, sex, race/ethnicity, and income are measured on the overall and entire sample; however, some assays of the exposome and phenome are measured on yet a subsample. A tutorial on the ExWAS approach in R is seen here: https://github.com/chiragjp/nhanespewas/blob/main/exwas_tour.Rmd

First, in line 1, we select pairs of phenotypes and exposures that have at least 500 measured participants. We further only select pairs that are present in more than one survey sampling. In lines 5-10 (Figure S1 B-D), we developed functions that query and merge the three or more data tables that contain the (1) phenotype and (2) exposure variables and (3) demographic variables and calculate the subsample weights to create a new NHANES-specific data table structure (see functions: *get_expo_pheno_tables*, *figure_out_weight* in the package). Next, we needed to combine data across multiple survey samplings and compute the new sample weights. For each survey, we calculated the subsample weights are selected by the smallest subsample being analyzed. For example, to estimate an association between fasting glucose (P) and urinary phthalate (E) biomarker, the data table that contains the smallest sample size will be the sample weight chosen. Last, we averaged the survey weights in accordance with the guidance provided by National Center for Health Statistics [45].

**Executing the pipeline**

We cataloged a total of 322 real-valued phenotypes and 859 biomarkers or self-reported questionnaires responses that measure environmental, dietary, or behavioral exposures across all 10 surveys. Each exposure-phenotype pair can be assessed in multiple surveys (1-10), and/or, have multiple categorical levels (e.g., smoking is a categorical variable with 2 levels and a reference group). Across 322 phenotypes, 859 exposures, 10 surveys, and multiple variable levels, the total number of possible associations were 563,627.

In our pipeline to maximize power and replication, we filtered for phenotype-exposures pairs that had at least 500-600 participants across at least 2 or more surveys (maximally a phenotype and exposure would appear in all 10 surveys), resulting in 278 unique phenotypes and 619 unique exposures. We calculated 119,643 associations between a phenotype (e.g., body mass index) and exposure (e.g., beta-carotene) pair. The total number of associations in 2, 3, 4, 5, 6, 7, 8, 9, 10 independent surveys was 34,206, 21,451, 34,048, 9,011, 8,412, 1,702, 9,741, 999, and 1,455. We used survey-weighted regression to associate phenotypes with exposures while adjusting for age, age$^2$, sex, race/ethnicity, income, survey year, and education. We also performed associations with no adjustments (univariate) in each survey separately.

We associated *P* (total = 278) with every *E* (total = 651) in which overlapping samples were available across the surveys (total surveys = 10; 1999-2000, 2001-2002, 2003-2004, 2005-2006, 2007-2008, 2009-2010, 2011-2012, 2013-2014, 2015-2016, 2017-2018). An association could occur in 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 surveys. We used survey-weighted regression to accommodate complex survey design, including PSU, stratum, and subsample weights as input and implemented in ref[46] (Figure S1C-D). We considered two models to account for demographic-based confounding or stratification, a base model (as above) and a “demographic adjusted” model, which consists of covariates age, age$^2$, race/ethnicity (Non-Hispanic White [reference], Non-Hispanic Black, Mexican American, Other Hispanic, and Other Ethnicity), household-to-income poverty ratio (ranging from 1-5), education (less than ninth grade, less than high school, high school graduate [reference], attended college, college graduate), and survey as a categorical variable. The summary statistics consists of the

association sizes (beta coefficients), standard errors, p values, the sample size, surveys and $R^2$ for the entire model.

We estimated the Bonferroni family-wise error rate and the Benjamini-Yekutieli false discovery rate (FDR)[47] across all tests and primarily report findings that exceed either the family-wise error rate (Bonferroni). We also secondarily report the or false discovery rate at thresholds less than 5%.

*Random-effects meta-analysis across survey sampling*

We also estimated concordance of P-E associations across independent survey samplings (Figure S1 DE). First, we calculated the variation of the P-E associations across independent survey samples. For example, if a P-E association was measured in 3 surveys, we estimated the association separately for each survey and calculated the variation of the P-E correlation across the three surveys. We used a "unrestricted weighted least squares" meta-analytic approach [48] to estimate cross-survey standard error, and heterogeneity estimates (degree to which variation on association sizes can be explained across surveys, e.g. $I^2$).

Separate and independent samplings of US NHANES by year (e.g., 1999-2000 or 2003-2004) provide an opportunity to estimate "replication" rates, or the times that a P-E association has lower nominal significance level in greater than one survey sample. An association can occur in 2-10 surveys. For each scenario, we counted the number of times a P-E association was achieved with p-value less than 0.05; for example, for a P-E association that can be assessed in 5 surveys, that P-E association can be significant up to 5 times.

*Exposure-Exposure Correlation Globes*

We measured the partial Pearson correlation between 619 exposures with pair-wise complete data, a total of 200k correlations. The partial correlation is the correlation between exposures, adjusted by age, $age^2$, income to poverty threshold, education, and ethnicity. To visualize an "exposome globe" as an example, we queried the correlation database for correlations among randomly selected exposures and exposures associated with

Body Mass Index or Glycated Hemoglobin. We filtered to only visualize correlations that were greater than 0.25. We visualized the globe using the *igraph* R package [49].

*Shared associations between exposures and phenotypes*

We defined the "shared" associational architecture between pairs of exposures (e.g., phthalate and heavy metal) or phenotypes (e.g., body mass index and glucose) by the similarity of their phenotype-exposure associations. Specifically, for a given exposure, there is an array of associations estimated for *X* number of phenotypes (e.g., a row in the Atlas, Figure 3). We estimate the correlation distance between a pair of exposures. For example, if a pair of exposures (e.g., E1 and E2), have the same association for each n phenotypes (e.g., P1… Pn), their association architecture is equal and the correlation will be 1.

*Estimating the Aggregate Contributions of Genetic and Exposomic Factors to Phenotypes*

To compare the variance explained by exposomic and genetic factors across 29 phenotypes, we constructed and evaluated multivariate models using a three-step workflow.

First, for exposome-based modeling, we identified exposures significantly associated (FDR-adjusted p-value < 0.05) with each phenotype using univariate regressions. Among these, we selected up to the top 20 exposures based on their univariate $R^2$ contributions. We multiple imputed missing data using Multiple Imputation with Chained Equations (MICE) [50], generating 10 imputed datasets via predictive mean matching. For each phenotype, we fit two linear regression models: a baseline model incorporating demographic covariates (age, sex, ethnicity, education), and an expanded model that additionally included the selected exposures. The incremental $R^2$ attributable to exposures was calculated as the difference in variance explained between these two models.

Second, for genetic modeling, we leveraged previously published genome-wide association study (GWAS) summary statistics derived from approximately one million genetic variants (UK Biobank) across 125 quantitative phenotypes of anthropometry and biomarkers. 29 of these 125 phenotypes (23%) overlapped with those queried in this investigation. Genetic contributions ($R^2$) to each phenotype were obtained from

quantitative trait GWAS models that included age, sex, genotyping array, and principal components as covariates. The genetic $R^2$ was defined as the incremental variance explained by the genetic model beyond these baseline covariates. The median genetic $R^2$ across all 125 phenotypes queried as reported in ref was 7% (interquartile range of .4% and 9%; max $R^2$ of 36%).

We directly compared exposome-based and genetic-based $R^2$ values across the 29 overlapping phenotypes (Table S4). We selected phenotypes based on the availability of matched phenotypes data from NHANES and the UK Biobank. This allowed assessment of the relative explanatory power of exposomic versus genetic factors for each phenotype, critically informing the conclusions about the comparative utility of exposomic data relative to genetic predictors.

*Concordance of associations with different types of measurements of the same exposure*

Some variables are assessed at different times (part of a 2-day interview) or sample matrices (e.g., in both urine and blood). We examined the concordance of phenotype-exposure associations across these samplings (e.g., day 1 exposure vs. day 2, or blood vs. urine measure). For self-reported nutrients, we compared the phenotype-exposure associations for day 1 vs. day 2 Number of dietary supplements reported, Calcium, Carbohydrate, Copper, Folic acid, Folate, DFE, Iron, Energy, Lycopene, Lutein + zeaxanthin, Magnesium, Niacin, Phosphorus, Potassium, Selenium, Thiamin (Vitamin B1), Vitamin B12, Riboflavin (Vitamin B2), Vitamin B6, Vitamin C,Vitamin D (D2 + D3), Vitamin K, Zinc, Total fat, Iodine, and Total sugars. We also compared the phenotype-exposure associations between self-reported measures and their biomarker counterparts, including alpha carotene, vitamin b12, cis beta-carotene, cryptoxanthin, retinol, trans-beta-carotene, and vitamin D. Last, we compared phenotype-exposure associations between blood-based and urine based indicators, including mercury, cadmium, and cotinine.

*Browsable Atlas of Associations*

All results from the PE-WAS can be visualized, downloaded, and browsed online (Figure S1EF: http://apps.chiragjpgroup.org/pe_atlas/). Specifically, users may query an *ExWAS* for a specific phenotype and examine phenotypic specific associations, $R^2$, and exposome correlation globes. The package and data

resource can be found here: https://github.com/chiragjp/nhanespewas. The full cohort database can be found here at doi: 10.6084/m9.figshare.29182196. The full summary statistics can be found at doi: 10.6084/m9.figshare.29186171

## Data Availability

The full cohort database can be found here at doi: 10.6084/m9.figshare.29182196. The full summary statistics can be found at doi: 10.6084/m9.figshare.29186171. NHANES data are publicly available: https://wwwn.cdc.gov/nchs/nhanes/default.aspx. All results from the PE-WAS can be visualized, downloaded, and browsed online: http://apps.chiragjpgroup.org/pe_atlas/

## Code Availability

All code is available as a R analytics package developed under an MIT license here: https://github.com/chiragjp/nhanespewas. Analytic code was written under R version 4.4.2 (R versions supported include >= 4.0. The libraries required for the package  include DBI,RSQLite, dplyr, tidyr, survey, stats, tibble, broom, rlang, tidyselect, logger, purrr, stringr, ggplot.

## Acknowledgements

We thank Gary Miller for their review of early findings. This study was funded through NIH grants, including NIEHS R01ES032470 (CJP, JPAI, AKM), NIDDK R01DK137993 (CJP, AKM), NIEHS U24ES036819 (CJP).

## Author information

CJP, AKM, and JPAI conceived of and wrote the initial draft of the paper. CJP obtained the data and analyzed the data.

# Ethics Declaration

Competing interests: the authors declare no competing interests. This study was deemed “not human subjects” research by Harvard Institutional Review Board (IRB): IRB24-1004.

# Figure Legends

**Figure 1. Associational architecture of the exposome on the phenome.** A.) Upper left panel shows the -log10(p-value) vs. the exposure type. B.) number of significant P-E associations per phenotype category (total E associations shown in text above bar) , C.) number of significant P-E associations per exposome category (total phenotype associations shown in text). Red: associations below Bonferroni (4e-7), Green: associations below FDR (5e-4) but greater than Bonferroni, Blue: associations greater than FDR.

**Figure 2. Variance explained by the exposome across phenotype groups.** A.) R-squared ($R^2$) for exposure across exposure categories and phenome categories. B.) Cumulative distribution of $R^2$. The median $R^2$ for each phenotypic category is annotated. C.) Cumulative distribution of $R^2$ across 119k P-E associations. The median $R^2$ for each exposome category is annotated. Colors for A-C: Red: associations below Bonferroni ($4x10^{-7}$), Green: associations below FDR ($5x10^{-4}$) but greater than Bonferroni, Blue: associations greater than FDR. D.) $R^2$ attributable to the exposure versus only to Demographics (age, $age^2$, ethnicity, income, education, and sex). Red: $R^2$ attributable to multiple simultaneous exposures (up to 10). Demo.: Demographics (age, $age^2$, ethnicity, income, education, and sex)

**Figure 3. Phenome-Exposome Atlas.** 278 phenotypes across 18 categories are depicted column wise, 625 exposures across 18 categories are depicted row wise. Each entry in the matrix is the linear association relationship (“Adjusted Beta”) between exposure and phenotype. Grey denotes association that could not be estimated due to pair-wise missingness or total sample size less than 500.

**Figure 4. Exposome-exposome correlational globe and distribution of exposure-exposure correlations across the exposome.** A.) An example Exposome Globe depicts exposure factors and their correlations chosen randomly (thresholded for absolute values of correlations above 0.25). B.) Exposome Correlation Globe for exposures associated with Hemoglobin A1C or BMI. Node colors include: pollutants (red), infection (yellow), nutrients (red). C.) Distribution of exposure-exposure correlations. Grey denotes randomly selected correlations. Blue line denotes exposome correlations for exposures associated with BMI or Hemoglobin A1C. Black denotes correlations that achieved Bonferroni significance. Lines depict correlations at -0.25 and 0.25.

**Figure 5. Exposure-phenotype associations across model specifications (adjustment choice).** A.) We chose Bonferroni significant associations from a fully specified model with Age, Sex, Ethnicity, Income, Education, and Survey Year as covariates and compared those to “univariate” (e.g., only consider the exposure), age+exposure, sex + exposure, age+sex+exposure, age+sex+ethnicity, income + education + exposure, age + sex + income _ education + exposure, age + sex + ethnicity. Distributions show the difference between P-E associations from a fully specified model to those with other configurations. B.) Association for exposures associated with Body Mass Index under different scenarios. Cadmium shows both positive and negative association under different covariate scenarios.

# Tables

| Cohort/Repository Name | NHANES (USA) | UK Biobank (UK) | AllOfUs (USA) | PEGS (USA) | HHEAR (USA) |
|---|---|---|---|---|---|
| **Design & Sampling** | Surveillance/Cross-Sectional; children and adults | Longitudinal/cross-sectional; older adults | Longitudinal/case-control; children and adults | Cross-sectional case-control; adults | Consoritium of existing cohorts; mostly children |
| **Sample Size** | ~200k | ~500k | ~1M | ~20k | ~35k |
| **Exposure Types Measured** | Blood, urine (targeted exposome), questionnaire; medication | questionnaire, activity sensor, geospatial; medication | activity sensor; geospatial | blood, urine; activity sensor; geospatial; questionnaire; medication | blood, urine (targeted and untargeted) |
| **# of Exposure Variables** | ~500 | ~130 | TBD | ~1000s | varies per cohort: up to ~200s |
| **Phenotypes / Outcomes** | Physiological, self-report, mortality; 'omics, imaging | Physiological, self-report, claims codes linked from electronic record | Physiological, claims codes linked from electronic record | Self-report | Adjudicated outcomes per cohort |
| **Biospecimens** | By application | By application | By application | | Yes |
| **Geospatial Linkage** | Limited, by application | Yes | Yes | Yes | Limited |
| **Longitudinal Design** | No, aside from mortality outcomes | For disease outcomes, limited exposure | For disease outcomes | No | Per cohort basis only |
| **Data Access** | Open (and limited for Medicare linkage) | By application | By application | By application | Open |
| **Limitations** | Cross-sectional; some data missing not at random prevents massive analysis | Limited exposure information and not repeated systematically; convenience cohort | Limited exposure information and not repeated systematically; convenience/colunte er cohort | Lack of longitdunal outcomes | Lack of longtidunal outcomes and exposures in repository |
| **Strengths** | Surveillance, representative of non-institutionalized population | Massive; linkages with omics; best suited for genetics | Massive; linkages with omics; best suited for genetics | Breadth of both environmental and genomic data; active recruitment | Multiple cohort settings to evaluate replication |

**Table 1.** Examples of cohorts (National Health and Nutrition Examination Survey [NHANES], UK Biobank, US AllOfUs, Human Health Exposure Analytics Resource [HHEAR], Personalized Genes and Environment Study [PEGS]) for exposome-wide study.

# Supplementary Results

### Cohort Characteristics, Sample Size, and Power

Each exposure-phenotype pair can be assessed in multiple surveys (1-10), and/or have multiple categorical levels (e.g., smoking is a categorical variable with 2 levels and a reference group).

We considered only phenotype-exposure pairs that had at least 500 participants across 2 or more surveys (maximally a phenotype and exposure would appear in all 10 surveys) resulting in 278 unique phenotypes and 619 unique exposures with complete cases across all possible demographic covariates. The total number of associations possible to assess in 2, 3, 4, 5, 6, 7, 8, 9, and 10 independent surveys was 34,206, 21,451, 34,048, 9,011, 8,412, 1,702, 9,741, 999, and 1,455, respectively. To maximize power, for each association, we combined across 2 or more surveys, combining the sample weights as advised by the National Center for Health Statistics (see Methods).

The range of the sample sizes per association included 608 to 68,495 individuals. The median sample size was 7,464 and the interquartile range of 4,220 and 15,316. For a sample size of 608, the lowest $R^2$ that can be identified was 0.05 with a power of 80% and a p-value threshold of $1x10^{-6}$. For the same power and p-value threshold, for sample size of 4,200, the lowest $R^2$ that can be identified was 0.007; for a sample size of 7,464, the lowest $R^2$ that can be identified was 0.004; for a sample size of 15,316, the lowest $R^2$ that could be identified was 0.002; and for the maximum sample size (68,315), the lowest $R^2$ that could be identified was 0.001.

We estimated the demographic characteristics for each pairwise association, including average age, sex prevalence, ethnicity percentage, education percentage, and average income of the population sample for each association (Figures S2-S4). Overall, the median prevalence of females was 51.2% (IQR: 50.6, 51.5%). The median age was 40 (interquartile range [IQR]: 34, 42) and the median income-to-poverty ratio was 2.9 (IQR: 2.8, 2.9). The median prevalence of the Non-Hispanic White population was 65% (IQR: 62, 69%), Non-Hispanic Black was 11.8% (IQR: 11.5, 12.2%), Mexican Americans was 9.1% (IQR: 8.5, 10.2%), Other Ethnicity was 7.7% (IQR: 5.9-9.3%), and Other Hispanic was 5.9% (IQR: 4.8, 6.9%). There was a median prevalence of 24%, 26%, 21%, 14%, and 16% for individuals who were college graduates, attended college

(but did not finish), high school graduates, attended high school, and less than high school, respectively. The demographic characteristics changed slightly per survey year (Figure S2-4).

**Exposure-wide associations across the phenome**

We used survey-weighted regression to associate phenotypes with exposures under 9 different modeling scenarios that adjust for demographic and social attributes: the (1) main reported model, which consists of age, age$^2$, sex, income (household income index divided by the poverty level), ethnicity (5 groups) and education (3 groups: above high school, high school equivalent, and below high school), and survey year (e.g, 1999-2000, 2001-02, 2003-04, 2005-06, 2007-08, 2009-10, 2010-11, 2012-13, 2014-15, 2016-17 as a categorical) (2) base model, with no adjustments, (3) sex and survey year, (4) age and age$^2$ and survey year (5) sex, age, and age$^2$, and survey year (6) ethnicity and survey year, (7) income, education, and survey year (8) age, age$^2$, sex, and ethnicity and survey year, and (9) age, age$^2$, sex, income, education, and survey year.

We report multiple estimates of association between exposures and phenotypes as displayed in Figures 1-4. First, we "scaled" (see Methods: divided the variable by the standard deviation) both the continuous exposures and phenotypes. From this regression, we attain a linear association between the exposure and phenotype (also known as beta coefficient). We also estimate the p-value of the association (beta coefficient) and variance explained (or $R^2$) that the exposure explains on the phenotype. For categorical variables, we estimated effect size in reference to a priori chosen reference group. The Bonferroni threshold across all associations was 0.05/119,643 (~$4 \times 10^{-7}$). We also estimate the Benjamini-Yekutieli false discovery rate (FDR).

We estimated a lambda exposome inflation factor of 2.53 for all the associations[20]. P-values deviated from the uniform distribution (Figure S5).

**Heterogeneity of associations across survey samples**

We also estimated the $I^2$, or the proportion of variation of P-E associations across survey samples, estimated using an unrestricted weighted least squares [21](Figure S7). For associations that were at least FDR significant,

the median $I^2$ was 0, 5, 0, 26, 6, 14, 14, 20, and 18% for associations in 2, 3, 4, 5, 6, 7, 8, 9, 10 surveys. The highest $I^2$ was observed in associations involving organophosphate exposure compounds ($I^2$, of 93% Figure S7).

**Demographics influences inferred associations**

We also estimated the standard deviation of univariate adjustments for associations binned in different categories. The largest standard deviation between fully adjusted and univariate estimates came from phenotype-exposure associations of infection (SD: 0.19), smoking behavior (0.14), organochlorine (0.10), and smoking biomarker (0.09) (Figure S8).

**Shared associational architecture between exposures measured in urine and blood**

The shared associational architecture between markers measured in urine and blood, such as urinary and blood cadmium was 0.58. The shared associational architecture between urinary and blood lead was 0.71. The shared association between urinary and blood cotinine was 0.9. However, there are exceptions: there are cross-category correlations, such as between supplements and the dietary interview or between dietary biomarkers and dietary interview.

# Supplementary Figure and Table Legends

**Figure S1. Methods flow for creation of the PE-Atlas.** A.) Merging datasets to combine exposures and phenotypes. Phenotypes and exposures may be transformed (right panel). B.) Reweighting the data to reflect appropriate sample weighting across 2-10 survey samplings. C.) Iterating through co-variable adjustments. A different combination of co-variables are chosen from the left hand side. Energy and urine creatinine are added for dietary and urinary biomarker associations, respectively. Survey year is added to all associations. D.) Survey weighted linear regression across 2-10 survey samplings. $I^2$ is calculated to estimate heterogeneity. E.) Summary statistics are collected, including association size, p-values, and $R^2$ for each P-E association. F.) A web app atlas is available for download and browsing the atlas.

**Figure S2.** Distribution of income, percent female, and age for each P-E association per survey and variable. Blue diamond denotes the median; the red points denote the 10th and 90th percentiles; black points are the 25th and 75th percentiles. Variables include PIR: Household Income to Poverty Ratio [ranges from 1-5]; FE - female percent. Letters A-J denote survey period (e.g., A: 1999-2000, B: 2001-2002, etc).

**Figure S3.** Distribution of education for each P-E association per survey and variable. Blue diamond denotes the median; the red points denote the 10th and 90th percentiles. Variables include PIR: Household Income to Poverty Ratio [ranges from 1-5]. Ed:9-11- education achieved between 9-11; ed:c - attended college; ed:cgrad

- graduated college; ed:hsgrad - graduated high school; ed:<9 - less than 9th grade. Letters A-J denote survey period (e.g., A: 1999-2000, B: 2001-2002, etc).

**Figure S4.** Distribution of Ethnicity for each P-E association. M - Mexican; B - Non-Hispanic Black; W - Non-Hispanic White; OE - Other Ethnicity. Letters A-J denote survey period (e.g., A: 1999-2000, B: 2001-2002, etc).

**Figure S5.** P-value distribution for PE associations, binned by exposome categories (A) and phenotype categories (B).

**Figure S6.** Number of surveys that achieved a p-value less than 0.05 for a P-E association examined in 2, 3, 4, 5, 6, 7, 8, 9, 10 surveys.

**Figure S7.** $I^2$ of P-E associations for exposure categories.

**Figure S8.** Distribution of fully adjusted minus minimally adjusted models across different exposure categories.

**Figure S9**. Concordance of PE associations: Day 1 nutrient vs. Day 2.

**Figure S10.** Correlation of PE associations for blood vs. self-reported marker of alpha carotene (ACAR), vitamin b12 (B12), cis beta-carotene (CBCAR), cryptoxanthin (CRYP), retinol (RET), trans-beta-carotene (TBCAR), and vitamin D (VD).

**Figure S11.** Correlation of PE associations for blood vs. urine biomarker of cadmium (CD), cotinine (COT), and mercury (HG).

**Figure S12. Shared exposomic and phenotypic architecture. A.)** Shared exposomic architecture for exposures within each domain. (e.g., the concordance of beta coefficients across phenotypes). Pairs of exposures that demonstrate more common associations are seen in red; pairs of exposures that demonstrate least common associations are seen in blue). B.) Shared phenotypic architecture for exposures within each domain. (e.g., the concordance of beta coefficients across phenotypes). Pairs of phenotypes that demonstrate more common associations are seen in red; pairs of phenotypes that demonstrate least common associations are seen in blue).

**Figure S13.** Comparing Exposomic $R^2$ vs. Genomic $R^2$ across 29 quantitative physiological phenotypes. We examined up to 20 exposome factors simultaneously for each phenotype and compared it to the total $R^2$ of common genetic variants (from Tanigawa et al, PLOS Genetics 2022).

**Figure S14.** R-squared of genetics for phenotypes from Tanigawa et al (PLOS Genetics 2022). Y axis denotes the R2 due to all of the genotypes (m=1M) and covariates, X axis denotes covariate only model. Median $R^2$ due to genetics was 8%; 25th percentile was 1.7%, 75th percentile was 9% and max was 36%.

Table S1. Exposure categories. N: Number of variables in category.
Table S2. Phenotype categories. N: Number of variables in category.
Table S3. $R^2$ for each phenotype and the number of exposures in the model
Table S4. $R^2$ for genetics and exposome for 29 phenotypes.

Data Table 1. Summary statistics for all 100k E-P associations. The column evarname is the NHANES variable name for the exposure, and pvarname is the NHANES variable name for the phenotype. The field

estimate.uwls reports the unrestricted weighted least-squares (uwls) meta-analytic beta coefficient pooled across all contributing survey waves, and std.error.uwls is its corresponding standard error. The significance of the pooled coefficient is given by p.value.uwls, while statistic.uwls shows the test statistic (Z or t) associated with that meta-analysis. Heterogeneity metrics include Q.uwls, which is Cochran's Q statistic, h.squared.uwls, the $H^2$ heterogeneity ratio, and i.squared.uwls, the $I^2$ percentage of total variation attributed to heterogeneity; k.uwls denotes the number of independent NHANES survey waves (k) used. To assess consistency across cycles, model_concordance counts how many survey waves achieved a nominal p-value < 0.05, and model_concordance_1_2 counts how many waves had p < 0.05 with an effect sign matching the overall estimate. The fields model_number and total_n indicate, respectively, which numbered adjustment scenario (covariate set) was used in the regression and the total sample size summed over all included waves. Two columns, rsq_adjusted_base_diff and rsq_adjusted_diff, report differences in $R^2$: the former is the change in explained variance between the fully adjusted and unadjusted (base) models, while the latter is the change between the fully adjusted model and the next less-adjusted model in sequence. For multiple testing correction, pval_BY gives the Benjamini–Yekutieli false discovery rate–adjusted p-value across all tests, and pvalue_bonferroni gives the Bonferroni-corrected p-value. The categorical field sig_levels provides a symbolic significance indicator based on those corrected p-values. Descriptive labels for each phenotype include pvardesc and its broad pcategory and psubcategory, and similarly evardesc, ecategory, and esubcategory describe each exposure. In addition, the fields pnewsubcategory and enewsubcategory represent any further refined subcategory assignments made programmatically. Finally, the super-category of each exposure is denoted by super_ecat_number, a numeric code, and super_ecat_name, its corresponding text label.

Data Table 2. Exposome Catalog. Evarname is the NHANES variable identifier, super_ecat_name is the high level category name, enewsubcategory is the category of the variable, evardesc is the description of the variable.

Data Table 3. Phenotype Catalog. Pvarname is the NHANES variable identifier, pcategory is the phenotype category, pvardesc is a description of the variable.

# References


1. Lakhani, C. M. *et al.* Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nat. Genet.* **51**, 327–334 (2019).
2. Peters, A., Nawrot, T. S. & Baccarelli, A. A. Hallmarks of environmental insults. *Cell* **0**, (2021).
3. Schwartz, D. & Collins, F. MEDICINE: Environmental Biology and Human Disease. *Science* **316**, 695–696 (2007).
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. Wild, C. P. The exposome: from concept to utility. *Int. J. Epidemiol.* **41**, 24–32 (2012).
6. Miller, G. W. Integrating exposomics into biomedicine. *Science* **388**, 356–358 (2025).
7. Ioannidis, J. P. A., Loy, E. Y., Poulton, R. & Chia, K. S. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci. Transl. Med.* **1**, 7ps8 (2009).
8. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
9. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* (2022) doi:10.1038/s41586-022-05275-y.
10. Suchak, T. *et al.* Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database. *PLoS Biol.* **23**, e3003152 (2025).
11. Hill, A. B. The Environment and disease: association or causation? *Proc. R. Soc. Med.* **58**, 295–300 (1965).
12. Ioannidis, J. P. A., Tarone, R. & McLaughlin, J. K. The False-positive to False-negative Ratio in Epidemiologic Studies. *Epidemiology* **22**, 450–456 (2011).
13. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2**, e124 (2005).
14. Patel, C. J. & Ioannidis, J. P. A. Studying the elusive environment in large scale. *JAMA* **311**, 2173–2174 (2014).
15. Ioannidis, J. P. A. The Challenge of Reforming Nutritional Epidemiologic Research. *JAMA* (2018)

doi:10.1001/jama.2018.11025.

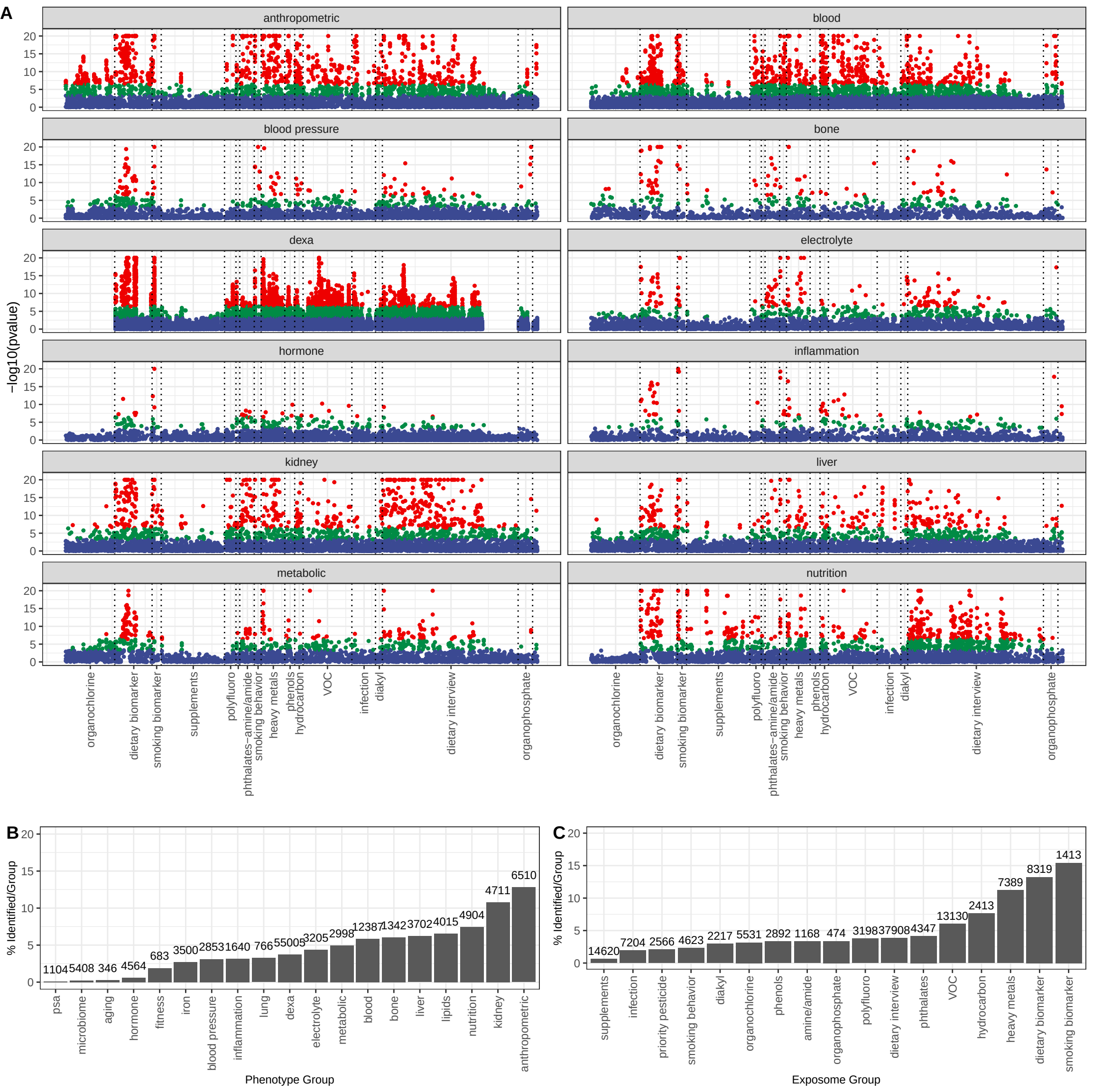
16. Blair, A. *et al.* Epidemiology, Public Health, and the Rhetoric of False Positives. *Environ. Health Perspect.* **117**, (2009).
17. Boffetta, P. *et al.* False-Positive Results in Cancer Epidemiology: A Plea for Epistemological Modesty. *J. Natl. Cancer Inst.* **100**, 988–995 (2008).
18. Ioannidis, J. P. A. Exposure-wide epidemiology: revisiting Bradford Hill. *Stat. Med.* **35**, 1749–1762 (2016).
19. Patel, C. J. *et al.* Opportunities and Challenges for Environmental Exposure Assessment in Population-Based Studies. *Cancer Epidemiol. Biomarkers Prev.* **26**, 1370–1380 (2017).
20. NHANES Questionnaires, Datasets, and Related Documentation. https://wwwn.cdc.gov/nchs/nhanes/Default.aspx.
21. Tanigawa, Y. *et al.* Significant sparse polygenic risk scores across 813 traits in UK Biobank. *PLoS Genet.* **18**, e1010105 (2022).
22. Chung, M. K., Buck Louis, G. M., Kannan, K. & Patel, C. J. Exposome-wide association study of semen quality: Systematic discovery of endocrine disrupting chemical biomarkers in fertility require large sample sizes. *Environ. Int.* (2018) doi:10.1016/j.envint.2018.11.037.
23. Chung, M. K. *et al.* Utilizing a Biology-Driven Approach to Map the Exposome in Health and Disease: An Essential Investment to Drive the Next Generation of Environmental Discovery. *Environ. Health Perspect.* **129**, 85001 (2021).
24. Ioannidis, J. P. A. Molecular bias. *Eur. J. Epidemiol.* **20**, 739–745 (2005).
25. Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P. A. & Boulesteix, A.-L. Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *Int. J. Epidemiol.* (2020) doi:10.1093/ije/dyaa164.
26. Patel, C. J., Burford, B. & Ioannidis, J. P. A. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J. Clin. Epidemiol.* **68**, 1046–1058 (2015).
27. Silberzahn, R. *et al.* Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science* **1**, 337–356 (2018).
28. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nat Hum Behav* **4**, 1208–

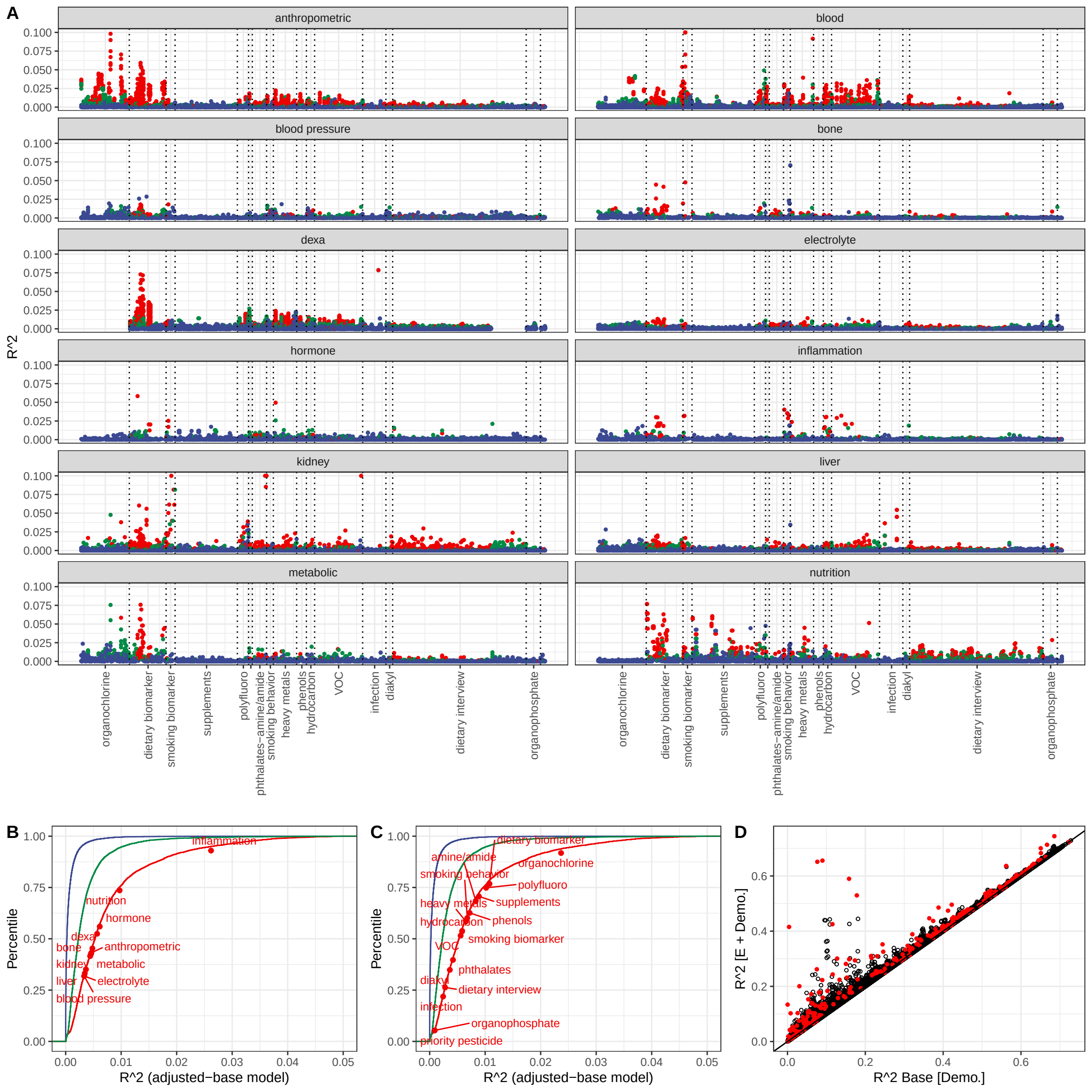
1214 (2020).

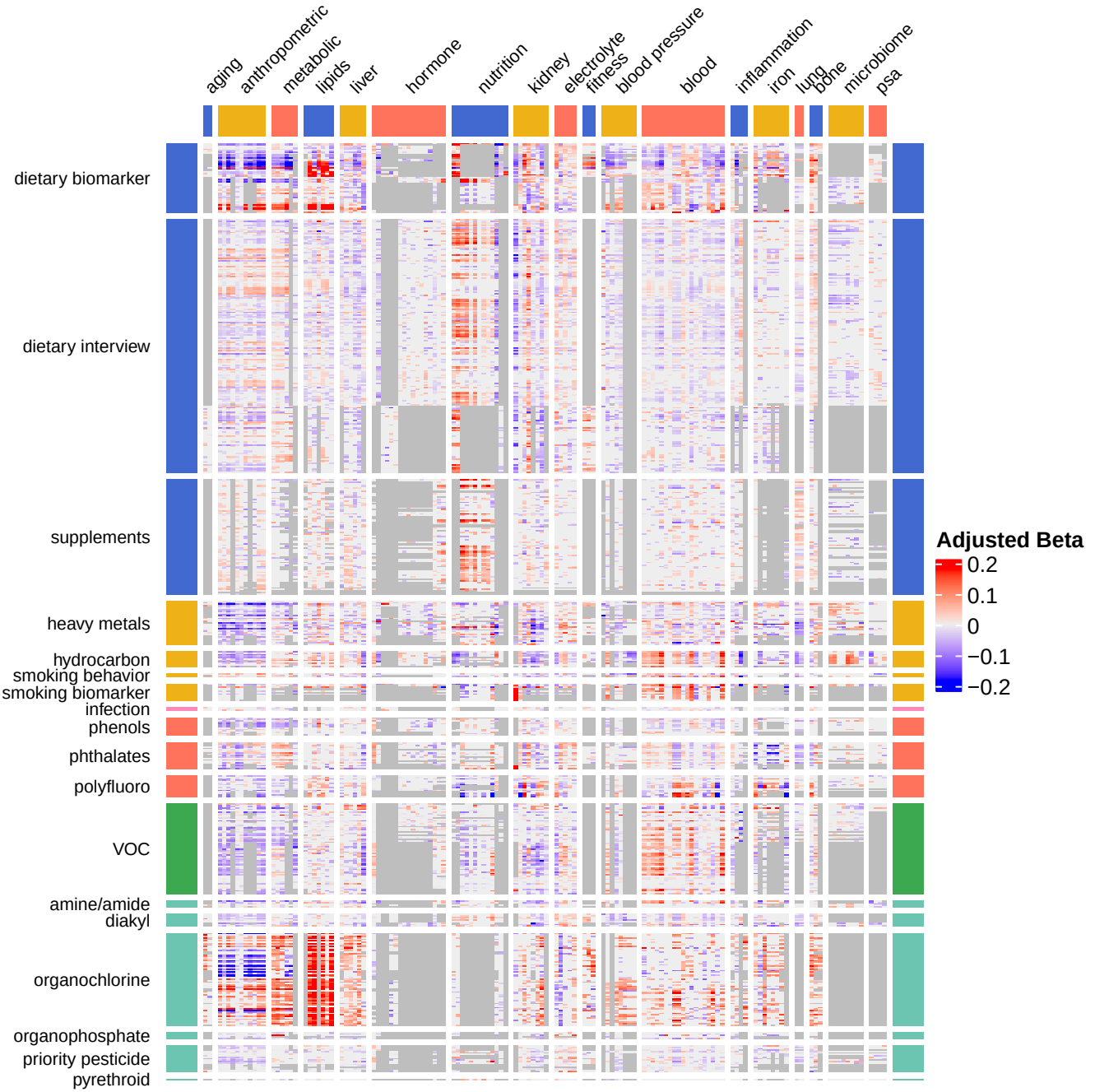
29. Patel, C. J., Ioannidis, J. P. A., Cullen, M. R. & Rehkopf, D. H. Systematic Assessment of the Correlations of Household Income With Infectious, Biochemical, Physiological, and Environmental Factors in the United States, 1999–2006. *Am. J. Epidemiol.* **181**, 171–179 (2015).
30. Nguyen, V. K., Colacino, J., Patel, C. J., Sartor, M. & Jolliet, O. Identification of occupations susceptible to high exposure and risk associated with multiple toxicants in an observational study: National Health and Nutrition Examination Survey 1999-2014. *Exposome* **2**, osac004 (2022).
31. Jbaily, A. *et al.* Air pollution exposure disparities across US population and income groups. *Nature* **601**, 228–233 (2022).
32. Di, Q. *et al.* Air Pollution and Mortality in the Medicare Population. *N. Engl. J. Med.* **376**, 2513–2522 (2017).
33. Argentieri, M. A. *et al.* Integrating the environmental and genetic architectures of aging and mortality. *Nat. Med.* 1–10 (2025).
34. 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).
35. You, L. *et al.* An exposome atlas of serum reveals the risk of chronic diseases in the Chinese population. *Nat. Commun.* **15**, 2268 (2024).
36. Neveu, V. *et al.* Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Res.* **45**, D979–D984 (2017).
37. Manrai, A. K. *et al.* Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health. *Annu. Rev. Public Health* **38**, 279–294 (2017).
38. Khoury, M. J., Iademarco, M. F. & Riley, W. T. Precision Public Health for the Era of Precision Medicine. *Am. J. Prev. Med.* **50**, 398–401 (2016).
39. Baccarelli, A., Dolinoy, D. C. & Walker, C. L. A precision environmental health approach to prevention of human disease. *Nat. Commun.* **14**, 2449 (2023).
40. Lumley, T. *Survey: Analysis of Complex Survey Samples*. (2014).
41. NHANES 2003-2004: Physical Activity - Individual Activities Data Documentation, Codebook, and

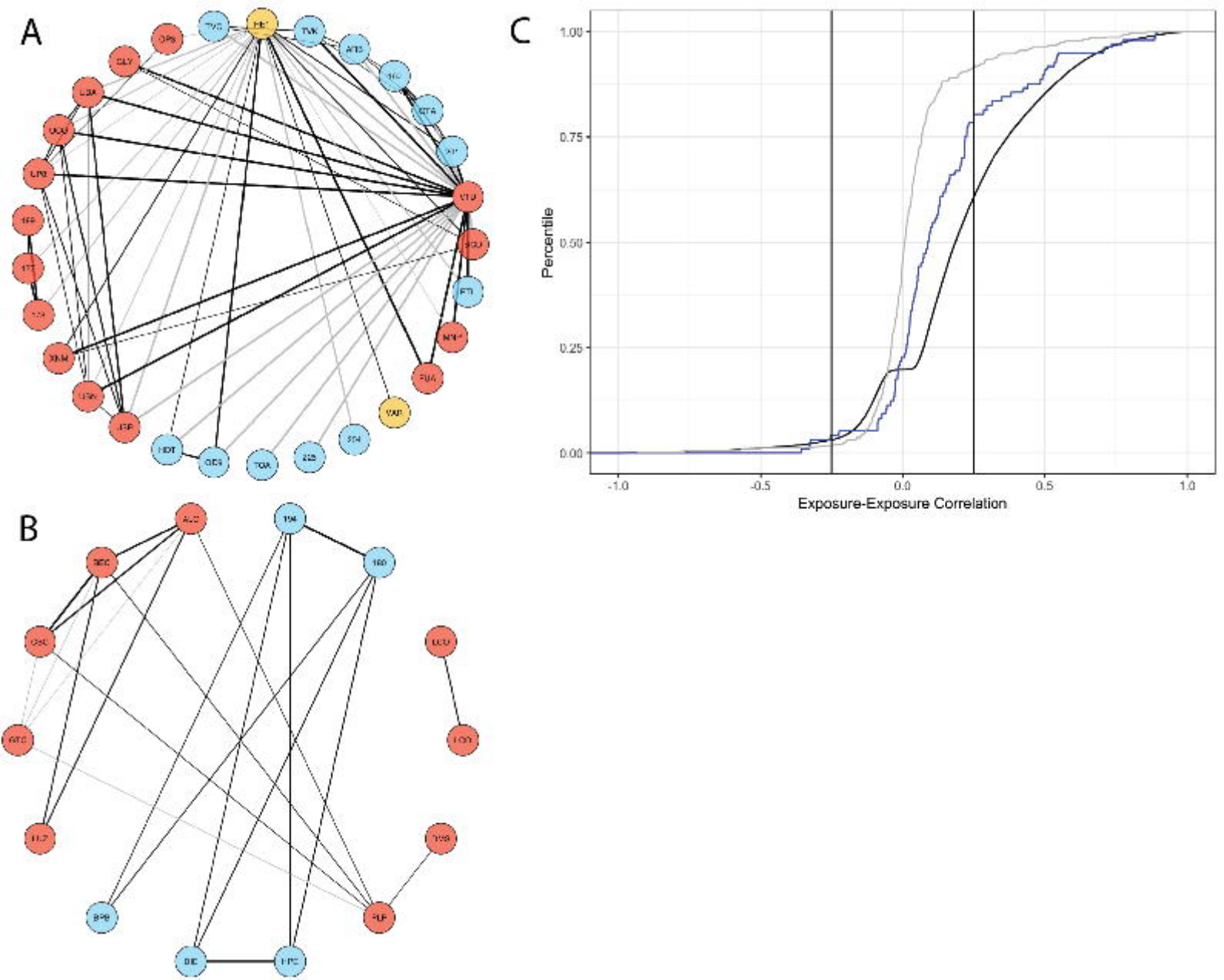
Frequencies. https://wwwn.cdc.gov/Nchs/Nhanes/2003-2004/PAQIAF_C.htm.

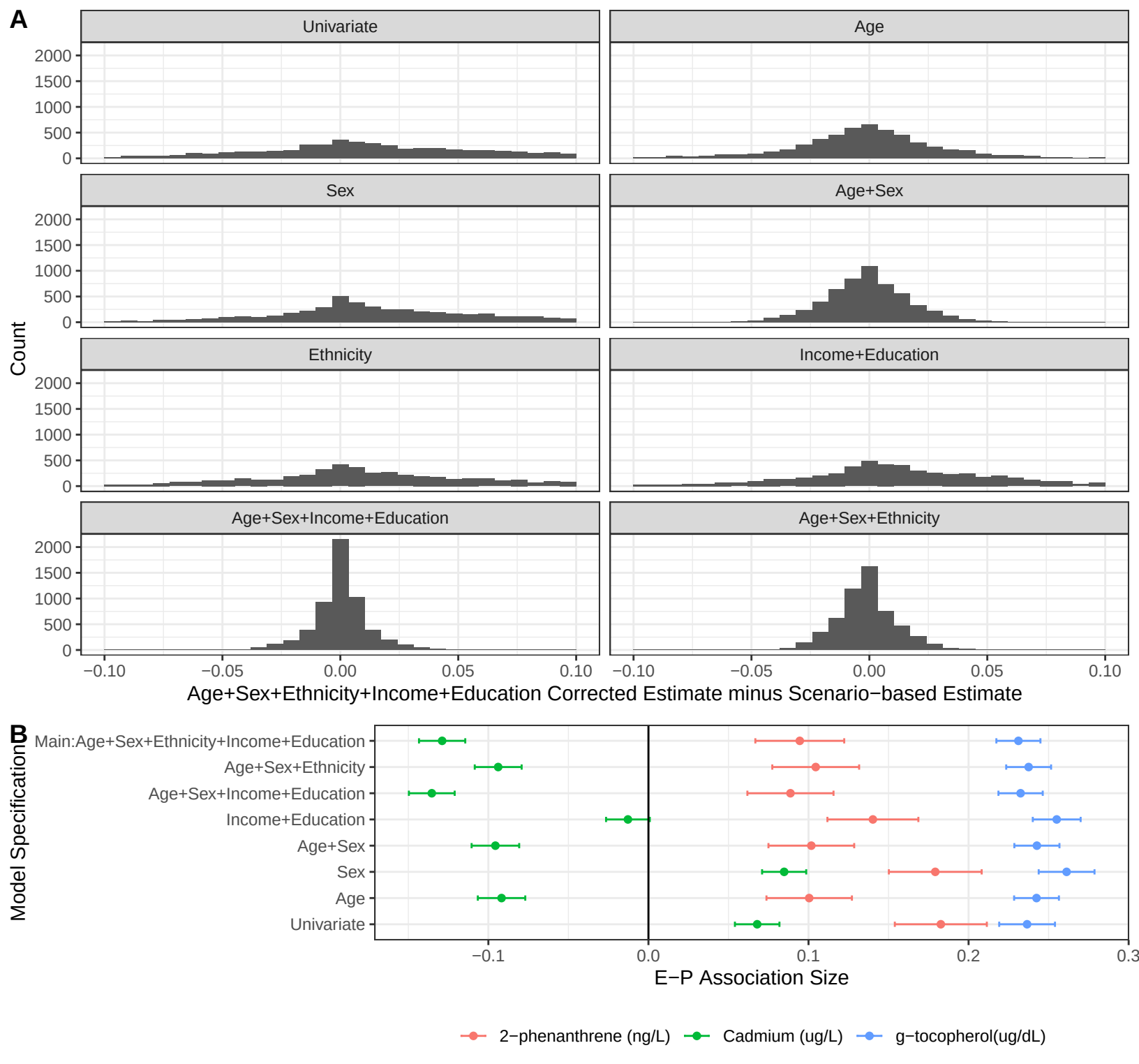
42. NHANES 2003-2004: Physical Activity Data Documentation, Codebook, and Frequencies. https://wwwn.cdc.gov/Nchs/Nhanes/2003-2004/PAQ_C.htm.
43. PAQ_J. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/PAQ_J.htm.
44. McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76**, 1262–1272 (2020).
45. NHANES Weighting. https://wwwn.cdc.gov/nchs/nhanes/tutorials/weighting.aspx.
46. Lumley, T. *Complex Surveys: A Guide to Analysis Using R*. (Wiley, Hoboken, 2010).
47. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 1165–1188 (2001).
48. Stanley, T. D. *et al.* Unrestricted weighted least squares represent medical research better than random effects in 67,308 Cochrane meta-analyses. *J. Clin. Epidemiol.* **157**, 53–58 (2023).
49. Csárdi, G. *et al. Igraph for R: R Interface of the Igraph Library for Graph Theory and Network Analysis*. (Zenodo, 2025). doi:10.5281/ZENODO.7682609.
50. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, Articles* **45**, 1–67 (2011).


A
anthropometric
blood
blood pressure
bone
dexa
electrolyte
hormone
inflammation
kidney
liver
metabolic
nutrition
−log10(pvalue)
organochlorine
dietary biomarker
smoking biomarker
supplements
polyfluoro
phthalates−amine/amide
smoking behavior
heavy metals
phenols
hydrocarbon
VOC
infection
diakyl
dietary interview
organophosphate
B
% Identified/Group
psa 1104
microbiome 5408
aging 346
hormone 4564
fitness 683
iron 3500
blood pressure 2853
inflammation 1640
lung 766
dexa 55005
electrolyte 3205
metabolic 2998
blood 12387
bone 1342
liver 3702
lipids 4015
nutrition 4904
kidney 4711
anthropometric 6510
Phenotype Group
C
% Identified/Group
supplements 14620
infection 7204
priority pesticide 2566
smoking behavior 4623
diakyl 2217
organochlorine 5531
phenols 2892
amine/amide 1168
organophosphate 474
polyfluoro 31983
dietary interview 7908
phthalates 4347
VOC 13130
hydrocarbon 2413
heavy metals 7389
dietary biomarker 8319
smoking biomarker 1413
Exposome Group


A
anthropometric
blood
blood pressure
bone
dexa
electrolyte
hormone
inflammation
kidney
liver
metabolic
nutrition
R^2
0.100
0.075
0.050
0.025
0.000
organochlorine
dietary biomarker
smoking biomarker
supplements
polyfluoro
phthalates−amine/amide
smoking behavior
heavy metals
phenols
hydrocarbon
VOC
infection
diakyl
dietary interview
organophosphate
B
Percentile
R^2 (adjusted−base model)
inflammation
nutrition
hormone
dexa
bone
anthropometric
kidney
metabolic
liver
electrolyte
blood pressure
C
dietary biomarker
amine/amide
organochlorine
smoking behavior
polyfluoro
heavy metals
supplements
hydrocarbon
phenols
smoking biomarker
VOC
phthalates
diakyl
dietary interview
infection
organophosphate
priority pesticide
D
R^2 [E + Demo.]
R^2 Base [Demo.]


aging
anthropometric
metabolic
lipids
liver
hormone
nutrition
kidney
electrolyte
fitness
blood pressure
blood
inflammation
iron
lung
bone
microbiome
psa
dietary biomarker
dietary interview
supplements
heavy metals
hydrocarbon
smoking behavior
smoking biomarker
infection
phenols
phthalates
polyfluoro
VOC
amine/amide
diakyl
organochlorine
organophosphate
priority pesticide
pyrethroid
Adjusted Beta
0.2
0.1
0
−0.1
−0.2


A
B
C
Percentile
1.00
0.75
0.50
0.25
0.00
-1.0
-0.5
0.0
0.5
1.0
Exposure-Exposure Correlation


A
Univariate
Age
Sex
Age+Sex
Ethnicity
Income+Education
Age+Sex+Income+Education
Age+Sex+Ethnicity
Count
2000
1500
1000
500
0
−0.10
−0.05
0.00
0.05
0.10
Age+Sex+Ethnicity+Income+Education Corrected Estimate minus Scenario−based Estimate
B
Model Specification
Main:Age+Sex+Ethnicity+Income+Education
Age+Sex+Ethnicity
Age+Sex+Income+Education
Income+Education
Age+Sex
Sex
Age
Univariate
−0.1
0.0
0.1
0.2
0.3
E−P Association Size
2−phenanthrene (ng/L)
Cadmium (ug/L)
g−tocopherol(ug/dL)