



**HAL**  
open science

## **Decoding the exposome: data science methodologies and implications in exposome-wide association studies (ExWASs)**

Ming Kei Chung, John S House, Farida S Akhtari, Konstantinos C Makris, Michael A Langston, Khandaker Talat Islam, Philip Holmes, Marc Chadeau-Hyam, Alex I Smirnov, Xiuxia Du, et al.

### ► **To cite this version:**

Ming Kei Chung, John S House, Farida S Akhtari, Konstantinos C Makris, Michael A Langston, et al.. Decoding the exposome: data science methodologies and implications in exposome-wide association studies (ExWASs). *Exposome*, 2024, 4, <10.1093/exposome/osae001>. <hal-04662309>

**HAL Id: hal-04662309**

**<https://hal.science/hal-04662309v1>**

Submitted on 25 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.








L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

## 2022 NIEHS Catalytic Workshop Series on the Exposome

## Decoding the exposome: data science methodologies and implications in exposome-wide association studies (ExWASs)

Ming Kei Chung <sup>1,2,3</sup>, PhD, John S. House <sup>4</sup>, PhD, Farida S. Akhtari<sup>4</sup>, PhD, Konstantinos C. Makris <sup>5</sup>, PhD, Michael A. Langston<sup>6</sup>, PhD, Khandaker Talat Islam<sup>7</sup>, PhD, Philip Holmes<sup>8</sup>, PhD, Marc Chadeau-Hyam <sup>9</sup>, PhD, Alex I. Smirnov<sup>10</sup>, PhD, Xiuxia Du<sup>11</sup>, PhD, Anne E. Thessen <sup>12</sup>, PhD, Yuxia Cui<sup>13</sup>, PhD, Kai Zhang<sup>14</sup>, PhD, Arjun K. Manrai<sup>1</sup>, PhD, Alison Motsinger-Reif <sup>4,\*</sup>, PhD, Chirag J. Patel <sup>1,T,\*</sup>, PhD and Members of the Exposomics Consortium

<sup>1</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

<sup>2</sup>School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, China

<sup>3</sup>Institute of Environment, Energy and Sustainability, The Chinese University of Hong Kong, Hong Kong, China

<sup>4</sup>Biostatistics and Computational Biology Branch, Division of Intramural Research, National Institute of Environmental Health Sciences, Durham, NC, USA

<sup>5</sup>Cyprus International Institute for Environmental and Public Health, School of Health Sciences, Cyprus University of Technology, Limassol, Cyprus

<sup>6</sup>Department of Electrical Engineering and Computer Science, University of TN, Knoxville, TN, USA

<sup>7</sup>Department of Population and Public Health Sciences, Keck School of Medicine of the University of Southern CA, Los Angeles, CA, USA

<sup>8</sup>Department of Physics, Villanova University, Villanova, Philadelphia, USA

<sup>9</sup>Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

<sup>10</sup>Department of Chemistry, NC State University, Raleigh, NC, USA

<sup>11</sup>Department of Bioinformatics and Genomics, College of Computing and Informatics, University of NC at Charlotte, Charlotte, NC, USA

<sup>12</sup>Department of Biomedical Informatics, University of CO Anschutz Medical Campus, Aurora, CO, USA

<sup>13</sup>Exposure, Response, and Technology Branch, Division of Extramural Research and Training, National Institute of Environmental Health Sciences, Durham, NC, USA

<sup>14</sup>Department of Environmental Health Sciences, School of Public Health, University at Albany, State University of NY, Rensselaer, NY, USA

\*To whom correspondence should be addressed to: Email: chirag\_patel@hms.harvard.edu; Email: alison.motsinger-reif@nih.gov

For full consortium author list, please see: <https://www.exposomicsconsortium.org/view/EXPOSOME-2023-007>

## Abstract

This paper explores the exposome concept and its role in elucidating the interplay between environmental exposures and human health. We introduce two key concepts critical for exposomics research. Firstly, we discuss the joint impact of genetics and environment on phenotypes, emphasizing the variance attributable to shared and nonshared environmental factors, underscoring the complexity of quantifying the exposome's influence on health outcomes. Secondly, we introduce the importance of advanced data-driven methods in large cohort studies for exposomic measurements. Here, we introduce the exposome-wide association study (ExWAS), an approach designed for systematic discovery of relationships between phenotypes and various exposures, identifying significant associations while controlling for multiple comparisons. We advocate for the standardized use of the term "exposome-wide association study, ExWAS," to facilitate clear communication and literature retrieval in this field. The paper aims to guide future health researchers in understanding and evaluating exposomic studies. Our discussion extends to emerging topics, such as FAIR Data Principles, biobanked healthcare datasets, and the functional exposome, outlining the future directions in exposomic research. This abstract provides a succinct overview of our comprehensive approach to understanding the complex dynamics of the exposome and its significant implications for human health.

**Keywords:** exposome; Exposome-Wide Association Study (ExWAS); phenotype; data science; false discovery rate; epidemiology.

## Introduction

The *exposome* encompasses an individual's life-course environmental exposures.<sup>1,2</sup> The original context focused on studying the environment with objective and higher precision methodology such as using exposure biomarkers. As the concept spans across multiple disciplines in medicine, sciences, and public health, it was later elaborated by others from different perspectives.<sup>3-6</sup> Nevertheless, the ultimate goal remains the same: to quantitatively characterize the phenomenon of multiple

exposures in humans, and ultimately how the totality of human exposure influences phenotypic traits, and it is necessary for investigators to understand two fundamental concepts that can be used to guide research and development in exposomics.

### Concept 1: The contribution of genetics and the environment to phenotype

With few exceptions, most human diseases have numerous contributing factors, which can be broadly classified as genetic

Received: May 8, 2023. Revised: October 16, 2023. Accepted: November 20, 2023

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

variation and environmental exposure. The fundamental concept includes that the relationship between these entities can be conceptualized as:

$$\text{Phenotype (P)} = \text{Genotype (G)} + \text{Environmental Exposures (E)} + \text{Interactions (GxE)} \quad (1)$$

Environmental exposures are nongenetic factors that include radiation, chemicals, socioeconomic entities, and climatic manifestations.<sup>1</sup> In data science terms, the exposome attempts to relate all environmental exposures to estimate the total contribution to phenotype. It defines how exposures are classified, for example, what domains do each exposure belong to (heavy metals, geospatial air pollution, etc.), how to “name” them (based on chemical structure or origin), and the units of measure (concentration in tissue or other matrices). Researchers have been estimating the total contribution of both the genome and the exposome in human phenotype by writing down the equation above in terms of their variance components<sup>7-9</sup>:

$$\text{Var(P)} = \text{Var(G)} + \text{Var(E}_{\text{shared}}) + \text{Var(E}_{\text{nonshared}}) + \text{Var(Interactions[GxE])} + \text{Var(Interactions[GxG])} + \dots + \text{error} \quad (2)$$

where *var* represents the variance of the corresponding characteristic; *E<sub>shared</sub>* represents the shared environmental factors across members; *E<sub>nonshared</sub>* represents the nonshared environment across members. Under this model and assuming independence between factors, the *heritability* is the variation explained by genetics over the total variation in the disease of interest ( $\text{var(G)}/\text{var(P)}$ ). In the past, family-based (eg, twin) studies were used to estimate *heritability* while more recently, *genome-wide association studies* (GWAS) are used. In the language of data science, a GWAS can be thought of as a type of feature selection method, where the features are the genetic variants and the response variable is the phenotype (a trait or a disease of interest). The goal of GWAS is to identify the subset of genetic variants that is most strongly associated with the human trait of interest. Investigators test individual single nucleotide polymorphisms (SNPs) for association with the phenotype or disease, and then correcting for multiple comparisons to control for the risk of false positives (ie, associations that are observed by chance). The next step is to estimate the total variance explained ( $R^2$ ) of the identified genotypes. This quantity is known as  $\text{var(G from GWAS)}/\text{var(P)}$ . The greater the heritability of a phenotype, the more the variation in the trait can be explained by genetics. *Can the same be achieved for environmental exposures of the exposome?*

Critically, the variation of the exposome is written down as the contribution of the shared and nonshared environment. Examples of the shared environment include exposures that are shared in a household or postal Code, such as outdoor air pollution while nonshared exposures are those that individuals encounter specific to their own experiences. These quantities can be written down as:

$$c^2 = \text{Var(E}_{\text{shared}})/\text{Var(P)} \quad (3)$$

$$e^2 = \text{Var(E}_{\text{nonshared}})/\text{Var(P)} \quad (4)$$

and are analogous to the coefficient of determination, or the total variance explained in a model.

To ascertain both the total contribution of the environment, and to attribute specific factors of the environment to this contribution, it is essential to account for time-varying, repeated and mixture exposures in the analysis to explain differences in phenotype not currently explained by candidate environmental factors and genetic factors and to solve Equations (1) and (2). From twin-based and genome-wide investigations, the total contribution of genetics is anywhere from 30%–50%,<sup>9,10</sup> and shared exposome is 10%,<sup>9</sup> leaving a large amount to be described by the nonshared exposome.

## Concept 2: Enhancing exposomic measurement at an epidemiological scale through cohort studies

No single universal method and approach is known that can capture a representative landscape of the totality of exposures, and exposomics studies typically require a combination of multiple methods for such a purpose.<sup>11-13</sup> The increasing use of large and complex observational studies such as the National Health and Nutrition Examination Survey (NHANES) and the UK Biobank—with comprehensive measurement of both genes and exposures—are becoming increasingly prevalent in health studies. New analytical skills are essential to perform data-driven research to understand the contribution of genetics and exposures as well as complex gene and environment interactions to phenotypic outcomes to address Equations (1) and (2). Apart from basic statistical inference of the associations between exposures and diseases, disentangling and identifying important exposures and building predictive models are also becoming routine analytical procedures.

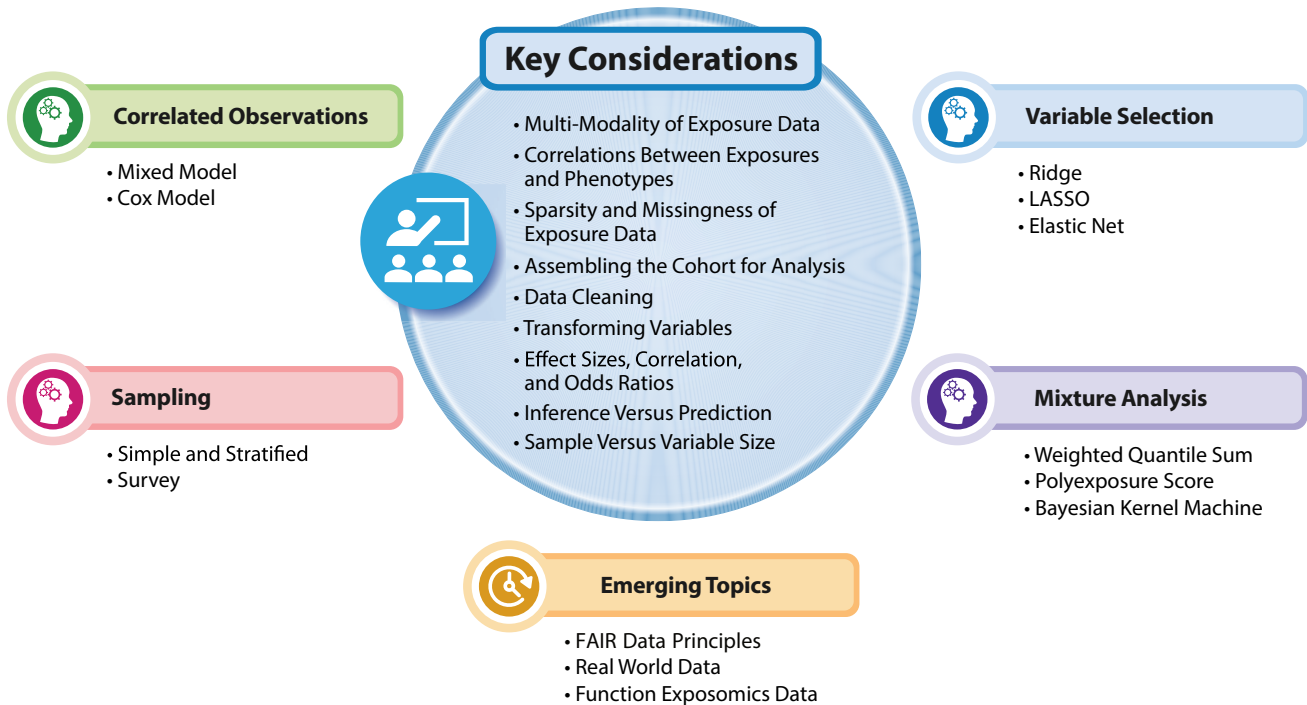
This essay aims to describe the above concepts from a data science perspective, providing a guide for the next generation health researchers to eventually examine and appraise exposomic studies (Figure 1). Finally, we share our view on potential topics that could have major influences concerning the development and practice of exposomics in the coming decades. We will discuss exposome-driven analysis as an extension of an *observational epidemiological study*, where the exposome and phenome are measured in human samples, in contrast to experimental studies where the investigator can assign individual subjects to different predefined exposure groups.

## Data science and exposome research

To establish the exposomic concept as a research paradigm, commonly investigated environmental factors in human observational investigation can be classified into three domains—general external, specific external, and internal,<sup>1</sup> or alternatively, four categories that include ecosystems, physical/chemical, lifestyle, and social.<sup>4</sup> These schemas encourage the adoption of a cross-disciplinary perspective on mixture of exposures when answering a broad research question on the role of the exposome in health outcomes. Specifically, the shared and nonshared environment contributions introduced in Concept 1 consist of both general external and specific internal factors. On the other hand, the internal exposome can be thought of as phenotypic changes that are induced when exposed to the external exposome,<sup>14</sup> therefore, enabling more in depth analysis of the relationships between exposures and outcomes such as “mediation analysis.”<sup>15</sup>

To effectively communicate exposome research, it is essential to convey key data science concepts that complement introductory public health courses and are extensible to research areas more specific to traditional disciplines such as air pollution,

## Exposome-Wide Association Study (ExWAS)



**Figure 1.** Key considerations for Exposome-Wide Association Studies.

chemical mixtures, and climate change. We begin with the *exposome-wide association study* (ExWAS), as it is instrumental in how to estimate the quantities and factors introduced in Concept 1 and 2.

### Overview of exposome-wide association studies (ExWASs)

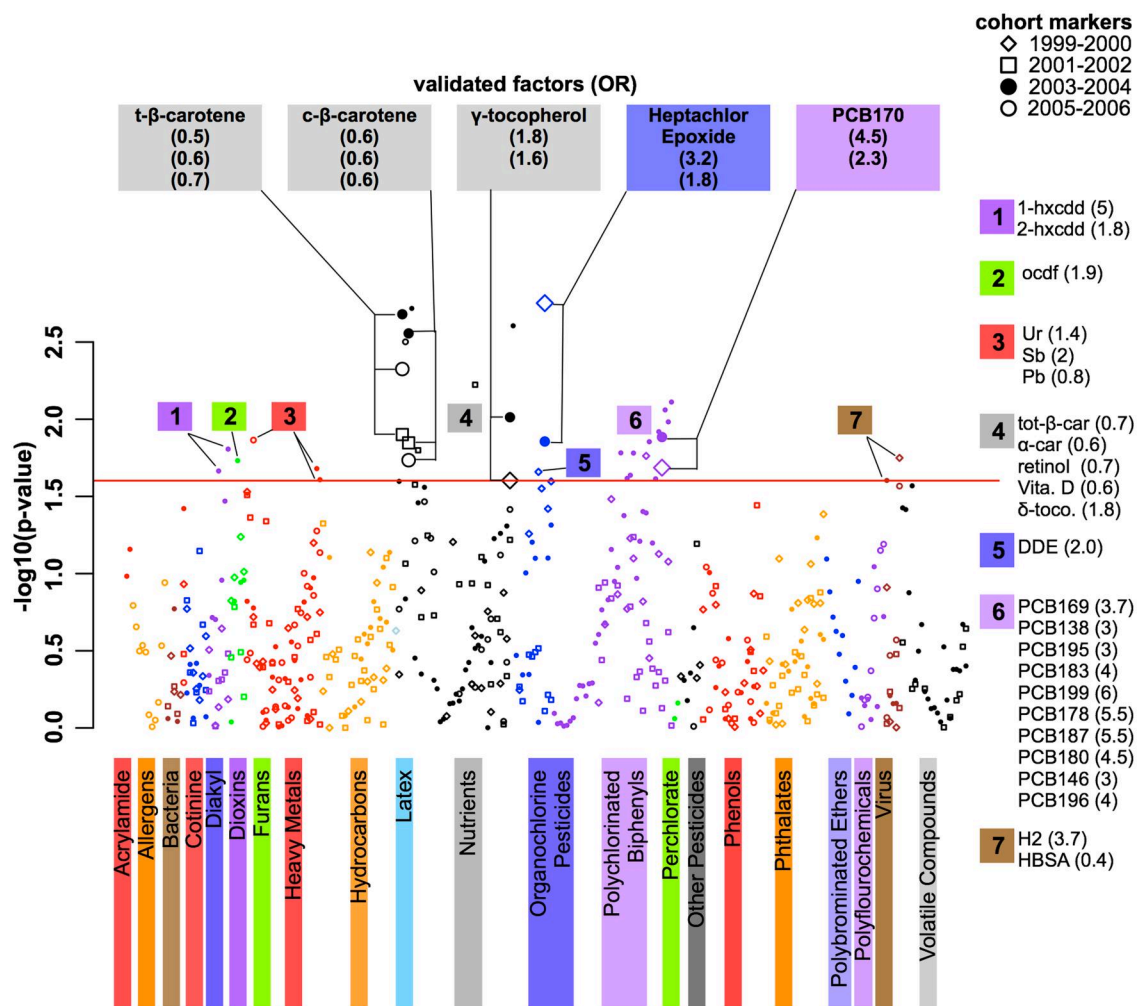
ExWAS is a data-driven analytical approach for conducting large-scale exploratory studies in exposomics, inspired by the GWAS paradigm in human genetics. It is robust as it works across different study designs, including but not limited to: cross-sectional, cohort, longitudinal, and (nested) case-control investigations. It is also highly interpretable as it can be driven by a variety of methods based on regression and other techniques. Fundamentally, ExWAS attempts to systematically model all the pairwise relationships between a single phenotype and multiple exposures, with a goal to identify statistically significant associations while controlling for the effects of multiple comparisons. For example, ExWAS was used to study the association of 266 environmental factors with type 2 diabetes and both risk (eg, heptachlor epoxide) and protective factors (eg,  $\beta$ -carotenes) were identified.<sup>16</sup>

In a typical ExWAS, the aim is to identify analytically important exposure-outcome pairs across all measured exposures. The choice of regression method, whether it is a basic linear regression or its extension, depends largely on the study design.<sup>17-19</sup> Statistical significance, or signal to noise, is estimated through the p value associated with the beta coefficients of the corresponding predictors. In a traditional hypothesis-driven study, the

threshold for type I error rate is set to 5%. Simply put, it means that if a null hypothesis is repeated 100 times, then five of them could be statistically significant, ie, false positives. The phenomenon of inflated significant findings by conducting many statistical tests is a version of data dredging.<sup>20</sup> In ExWAS, spurious associations due to multiple comparisons are controlled, for example, using false discovery rate (FDR),<sup>21</sup> which is an expected ratio of false positive to total positive findings in a study. Similarly, a Manhattan plot showing  $-\log_{10}$  p values enables quick visual inspection of all the associations (Figure 2).

Since ExWAS is a discovery-based approach, confirmation of statistically significant results is essential. In the simplest form, samples in a study could be split into two parts—one for discovery and one for validation, whereby the investigator seeks concordance of association statistics (eg, association size or beta-coefficient) in more than one sample of the study, such as a held-out sample, an independent survey cycle, or entirely new cohort.<sup>13,16</sup> Replicability can be assessed through an independent set of data.<sup>22</sup> To assist in conducting and interpreting results from ExWAS, we have also tabulated key resources, from identifying datasets to locating R statistical packages for analysis, in Table 1.

It can be useful to teach ExWAS via GWAS, which is a hypothesis-free approach, like ExWAS, to identify genetic factors associated with outcomes<sup>23</sup> (eg, G-P correlations), or high-throughput *inference*. They both use regression methods to identify factors associated with outcome. They both use similar summary statistics (eg, odds ratios, correlations) to convey the relationship between the G-P or E-P. Finally, the exposome or genome as a whole can be related to phenotype via *predictive*



**Figure 2.** A Manhattan plot illustrating the findings of an ExWAS for type 2 diabetes. The X-axis represents the exposures, while the Y-axis shows the corresponding probability values. Each point in the plot signifies the association test for a single exposure. The red horizontal line indicates the threshold for statistical significance. Reproduced from Patel et al.,<sup>16</sup> used under Creative Commons Attribution License.

approaches, where the summary statistics include total variance explained or area under a receiver operating characteristic curve (AUC). Connection between aggregate summary statistics, AUC, and attributable fraction in genetics research are possible. In genetics research, the aggregate risk, or “architecture” (the total spectrum of disease risk along the genome) are described in terms of “frequency of genetic variant” and effect size (eg, odds ratio), so investigators can visualize the risk of disease relative to how frequent a risk factor is.

How would one articulate the “architecture” of exposome-phenotype associations? As of this writing, the architecture of the exposome is articulated in terms of effect size versus signal to noise, or p value of single associations. Furthermore, genetic variants and exposures differ, in its dynamism and modality<sup>24</sup>: environmental exposures can change over time and space, which could depend on many factors including an individual’s behavior and lifestyle. Therefore, the “architecture” as ever-changing through the life course. Genetic variants, in contrast, are static, as they are inherited at conception and remain relatively constant throughout an individual’s life. Their architecture may change in the presence of environmental factors. In GWAS, genetic variants can be measured with high accuracy using well-established genotyping technologies,<sup>25</sup> whereas the technologies to measure the chemical exposome, such as organic chemicals

(DDT, PCBs, and PBDEs et cetera) and metals (Pb, Cd, and Cr et cetera), are generally involved using targeted methods with mass spectrometry. Until recently, targeted and untargeted mass spectrometry have been evaluated for measuring the exposome.

### Study designs for exposomics research

There are a number of study designs to elucidate the role of an exposure to a phenotype (Table 2). In region-wide or nationwide scale biobanks, subsamples can be extracted based on these basic designs to answer the research question. Since 2010, more than 60 studies applying the ExWAS approach have been published. These studies were designed to identify exposures associated with chronic diseases such as childhood obesity,<sup>26</sup> dementia,<sup>17</sup> coronary heart disease,<sup>27</sup> and autism.<sup>28,29</sup> It is also used to study exposures related to other outcomes such as mental well-being,<sup>18</sup> depression,<sup>30</sup> coffee consumption,<sup>31</sup> COVID-19,<sup>32,33</sup> and child behavior<sup>34</sup> (Table 3). Across these studies, single cohort, integrated cohorts, surveys, and biobank samples were used and sample sizes were between ~1000 to ~50000. Typically, the number of external exposures under investigations was between 50 to 200 (in 1 case >900), and exposures were often classified into different categories to aid interpretation. Some studies assessed all the available environmental factors in the

**Table 1.** Epidemiological data science resources for ExWAS analyses and interpretation

Type	Name	Description
Epidemiological data source Data portals	The Trans-Omics for Precision Medicine (TOPMed): <a href="https://topmed.nhlbi.nih.gov/">https://topmed.nhlbi.nih.gov/</a>	The TOPMed program consists of ~180k participants from more than 85 different studies. It consists of ancestrally and ethnically diverse sets of participants, focusing on phenotypes of heart, lung, blood, and sleep disorders with multi-omic data such as whole-genome sequencing (WGS) data and other omics (eg, transcriptomics, epigenomics, metabolomics and proteomics) data integrated with molecular, behavioral, imaging, environmental, and clinical data.
	Database of Genotypes and Phenotypes (dbGaP): <a href="https://www.ncbi.nlm.nih.gov/gap/">https://www.ncbi.nlm.nih.gov/gap/</a>	dbGAP is a NIH-maintained database that archives the data and results from genotype-phenotype studies in humans. It contains both open access and controlled access data. Data is organized as studies with phenotype data and genotype data such as SNP assays, methylation data, CNVs, genomic sequencing, exome data, expression arrays, RNA-Seq data, etc.
	Environmental influences on Child Health Outcomes: <a href="https://echochildren.org">https://echochildren.org</a>	The “ECHO Cohort” integrates numerous child cohorts and cover 5 outcomes, including obesity, pre-, peri-, and postnatal outcomes, upper/lower respiratory disease, wellness, and neurodevelopment totalling over 16k children.
	Human Health Exposure Analysis Resource (HHEAR): <a href="https://hhearprogram.org/">https://hhearprogram.org/</a>	The HHEAR is a centralized network of exposure analysis services and expertise available to eligible researchers who want to include or broaden exposure analysis to their human health studies. The HHEAR Data Center maintains a repository for HHEAR data including epidemiologic, biomarker and environmental exposure data, and associated data science tools.
Cohorts	The All of Us Research Program: <a href="https://allofus.nih.gov/">https://allofus.nih.gov/</a>	The All of Us Research Program aims to collect health data from >1 million participants in the USA with a focus on involving previously underrepresented populations. The cohort consists of ~429 k participants with electronic health records, survey data, genomic data, labs and physical measurements and biospecimens. The surveys include questions on overall health, lifestyle, medical history, and social determinants of health.
	The UK Biobank: <a href="https://www.ukbiobank.ac.uk/">https://www.ukbiobank.ac.uk/</a>	The UK Biobank is a large-scale biomedical database consisting of ~500 k participants from the UK with genetic, health and survey data. It is a longitudinal study with health and lifestyle survey data, physical measurements, biospecimens, imaging, electronic health records, biomarkers, wearables and multi-omic data (genotyping, whole genome sequencing and whole exome sequencing).
	The Million Veteran Program (MVP): <a href="https://www.research.va.gov/mvp/">https://www.research.va.gov/mvp/</a>	The MVP investigates the roles of genetics, lifestyle, exposures and military experiences on the health and wellness of Veterans in the USA. The MVP cohort consists of ~930 k participants with electronic health records, self-reported surveys and genotype data. The surveys comprise of information on health, lifestyle, military experiences and exposure, medical history and diet.
	The Nurses' Health Study (NHS): <a href="https://nurseshealthstudy.org/">https://nurseshealthstudy.org/</a>	The NHS consists of three prospective cohorts with ~275 k nurses, primarily female, with questionnaire data and biospecimens. Questionnaires are administered biennially and include questions on health, medical history, lifestyle, diet, behavior, environment and nursing occupational exposures. Biospecimens such as blood, urine, buccal DNA and toenail samples are available for a subset of participants.
	The Health Professionals Follow-up Study (HFPS): <a href="https://www.hsph.harvard.edu/hfps/">https://www.hsph.harvard.edu/hfps/</a>	The HFPS is an all-male study designed to be the complement to the primarily female Nurses Health Study. This study is comprised of ~22 k males in health professions such as dentists, pharmacists, optometrists, podiatrists, osteopaths, veterinarians, etc. Questionnaires are administered biennially and include questions about diseases such as cancer, heart disease and other vascular diseases, and health-related topics like smoking, physical activity, lifestyle, diet and medications.

(continued)

Table 1. (continued)

Type	Name	Description
	The National Health and Nutrition Examination Survey (NHANES): <a href="https://www.cdc.gov/nchs/nhanes/index.htm">https://www.cdc.gov/nchs/nhanes/index.htm</a>	The National Health and Nutrition Examination Survey (NHANES) is a vital program conducted by the CDC in the USA, designed to assess the health and nutritional status of the US population through interviews and physical examinations. NHANES uses a complex, multi-stage probability design to ensure its sample is representative of the US civilian noninstitutionalized population. NHANES plays a crucial role in exposomic sciences by providing extensive data on environmental exposures, such as various toxins, and contributing to biomonitoring efforts. The data gathered are pivotal for epidemiological studies exploring the relationship between environmental factors and health outcomes.
	The Personalized Environment and Genes Study (PEGS): <a href="https://www.niehs.nih.gov/research/clinical/studies/pegs/">https://www.niehs.nih.gov/research/clinical/studies/pegs/</a>	PEGS integrates genetic and environmental data for ~10 k racially and ethnically diverse participants and includes multi-dimensional data consisting of phenotypic data, genomic data, and extensive questionnaire-based and geospatial estimates of exposome-wide environmental exposures. The surveys include questions on health, lifestyle, medical history and various exposures such as residential and occupational environmental exposures, medication use, physical activity, stress, sleep, diet and reproductive history.
	Human Early Life Exposome (HELIX) project: <a href="https://helixomics.isglobal.org/">https://helixomics.isglobal.org/</a>	The HELIX project is a resource of multi-omics and exposome data for 1301 mother-child pairs from six European cohorts. The ExWAS (Exposome-wide association analyses) catalog can be used to query and download findings from the HELIX ExWAS. Summarized results for other omic analyses are also available for download.
<b>Location-based exposure sources</b>	The Center for Air, Climate, and Energy Solutions (CACES): <a href="https://www.caces.us">https://www.caces.us</a>	The CACES land-use regression (LUR) models provide estimates of outdoor concentrations for multiple pollutants by census tract. The CACES reduced complexity models (RCMs) estimate the impact of the emissions of multiple pollutants on human health.
	NASA Earthdata Collection: <a href="https://www.earthdata.nasa.gov/">https://www.earthdata.nasa.gov/</a>	The Earthdata collection includes measurements of the Earth's atmosphere, land, ocean, and cryosphere from a variety of sources, including sensor data from satellites and aircraft platforms, <i>in situ</i> measurements, field campaigns, and model estimates. These measurements can aid in the understanding of climate change, extreme weather patterns, hazards and disasters, air quality and water resources levels.
	CDC/ATSDR Social Vulnerability Index (SVI): <a href="https://www.atsdr.cdc.gov/placeandhealth/svi/">https://www.atsdr.cdc.gov/placeandhealth/svi/</a>	The SVI uses US Census data to calculate the social vulnerability at the census tract level (subdivisions of counties for which the Census collects statistical data). Each census tract receives an SVI rank based on 16 social factors which are also grouped into four related themes—socioeconomic status, household characteristics, racial and ethnic minority status, and housing type and transportation.
	ATSDR Environmental Justice Index: <a href="https://www.atsdr.cdc.gov/placeandhealth/eji/">https://www.atsdr.cdc.gov/placeandhealth/eji/</a>	The EJI ranks the overall effects of environmental injustice on health for each census tract. It ranks each census tract on 36 environmental, social, and health factors and groups them into ten domains and three overarching modules—the environmental burden, social vulnerability and health vulnerability modules.
<b>Statistical analysis</b>	rexposome: <a href="https://www.bioconductor.org/packages/release/bioc/html/rexposome.html">https://www.bioconductor.org/packages/release/bioc/html/rexposome.html</a>	An R package for the analysis of exposome data. Offers a set of functions to incorporate exposome data into the R framework and a series of tools to analyze exposome data.
	omicRexposome: <a href="https://bioconductor.org/packages/release/bioc/html/omicRexposome.html">https://bioconductor.org/packages/release/bioc/html/omicRexposome.html</a>	omicRexposome uses MultiDataSet for coordinated data management, rexposome for defining exposome data, and limma for association testing to facilitate the study of associations between exposures and omic data.
	MR-Base: <a href="https://www.mrbase.org/">https://www.mrbase.org/</a>	Platform for Mendelian Randomization using published GWAS summary statistics

(continued)

Table 1. (continued)

Type	Name	Description
<b>Chemical Information &amp; Interpretation</b>		
	Exposome-Explorer: <a href="http://exposome-explorer.iarc.fr/">http://exposome-explorer.iarc.fr/</a>	The Exposome-Explorer is a database of biomarkers of exposure to environmental risk factors for diseases. It contains information on known biomarkers of exposure to dietary factors, pollutants, and contaminants measured in population studies.
	The Blood Exposome Database: <a href="https://bloodexposome.org/">https://bloodexposome.org/</a>	Collated chemical lists from metabolomics, systems biology, environmental epidemiology, occupation, toxicology and nutrition curated from automated text mining from PubMed and PubChem databases.
	The Toxic Exposome Database (T3DB): <a href="http://www.t3db.ca/">http://www.t3db.ca/</a>	The database currently houses 3678 toxins described by 41 602 synonyms, including pollutants, pesticides, drugs, and food toxins, which are linked to 2,073 corresponding toxin target records. Altogether there are 42 374 toxin, toxin target associations. Each toxin record (ToxCard) contains over 90 data fields and holds information such as chemical properties and descriptors, toxicity values, molecular and cellular interactions, and medical information.
	CompTox Chemicals Dashboard: <a href="https://comptox.epa.gov/dashboard/">https://comptox.epa.gov/dashboard/</a>	The Dashboard contains chemistry, toxicity and exposure information for over one million chemicals, with over 420 chemical lists based on structure or category. It also enables access to the information in ExpoCast and ToxCast. Notably, one can also get access to EPA's Distributed Structure-Searchable Toxicity (DSSTox) Database, which contains accurate mapping of bioassay and physiochemical data on chemical substances to their chemical structure.
	PubChem: <a href="https://pubchem.ncbi.nlm.nih.gov/">https://pubchem.ncbi.nlm.nih.gov/</a>	Open chemistry database of the National Institutes of Health (NIH) since 2004. Small and large molecules with data on structure, identifiers, physiochemical properties, biological activity, as well as health, safety and toxicity data. Currently, it has over 115M compounds. Contributed to by academics, government agencies, chemical vendors and journal publishers.
	Chemical Entities of Biological Interest (ChEBI): <a href="https://www.ebi.ac.uk/chebi/">https://www.ebi.ac.uk/chebi/</a>	ChEBI is a dictionary of molecular entities. It focuses primarily on small chemical compounds that intervene in the biological processes of living organisms. Currently, it has over 60 000 annotated compounds.
	Tox21: <a href="https://ntp.niehs.nih.gov/whatwestudy/tox21/">https://ntp.niehs.nih.gov/whatwestudy/tox21/</a>	Testing of commercial chemicals and pesticides, food additives and chemical compounds in hundreds of cellular based assays and transcriptomic assays with dose-response characterization.

datasets, while others only selected a subset of the exposures, for instance, dietary exposures and/or other modifying factors.

## Characteristics of ExWAS

### Multi-modality of exposure data

One of the key characteristics of exposomics data is “multi-modality” of measurement, meaning that it encompasses multiple types and contexts of information.<sup>35,36</sup> Examples of these measurements include light and temperature (sub-molecular), biomarkers of chemicals (molecular), dietary intake and physical activity (lifestyle), income and education (socioeconomic status).<sup>37,38</sup> Large nationwide studies of the external exposures usually require integration through ZIP Code Tabulation Areas (areal representations of postal ZIP Codes), while the analysis of the internal exposome in the context of precision medicine involves multi-omics data such as genomics, transcriptomics, proteomics, and metabolomics.<sup>39</sup>

How large is the exposome? The “dimensionality,” or how many exposome variables are included in an ExWAS has analytic implications, such as signal-to-noise and false positive rates. Currently, the largest chemical database has over 275 million substances<sup>40</sup> but only make up a tiny fraction of the theoretical range (millions of billions).<sup>4</sup> Exposure information is commonly obtained through

geospatial modeling, laboratory measurement, questionnaire, or administrative records. Nevertheless, from a data analytical perspective, these diverse factors are viewed as one of the following types: categorical variables with nominal and ordinal subtypes, or numeric variables with interval and ratio subtypes. Raw data captured in numeric format can be encoded as it is, or recoded into categorical variables. For instance, one may record the number of cigarettes smoked per week and recode it into a variable with three categories: heavy, medium, and light smokers. While the best encoding choice depends on the context of the study, it is generally recommended to record data in the native format to avoid loss of information and biases.<sup>41-43</sup> The types of outcome variables can affect the choice of regression model. A logistic regression model and a linear regression model are used to analyze binary and continuous outcomes respectively. Similarly, the types of predictor variables can affect the interpretation of the beta coefficients—whether it is an increase in outcome per unit change of a continuous predictor, or a change in outcome relative to the reference group of a categorical variable.

### Correlations between exposures and phenotypes

Environmental exposures are known to be densely correlated.<sup>24,44,45</sup> Correlation patterns depend highly on the context of

**Table 2.** Common epidemiological study designs and their advantages and disadvantages for exposomic studies

Study design	Description	Advantages	Disadvantages
Cross sectional	For this design, data on the exposome and health outcomes are collected at a single point in time. It can provide a snapshot of the relationship between exposomic factors and health outcomes in a specific population.	Suitable for routine data collection and able to estimate population features such as prevalence of a disease.	Reverse causality: exposome factors coming before the outcome. Confounding: the observed association between an environmental exposure (the exposome factor) and a health outcome is distorted by the presence of another variable. In exposomic research, confounding can be particularly challenging due to the complex and multifaceted nature of environmental exposures.
Case control	It involves comparing the exposure history of individuals with a specific disease or health outcome (cases) to those without the outcome (controls). Cases are enrolled first and controls with similar demographic and other key characteristics as the cases are collected in the same population.	Relative simple and inexpensive to collect samples and able to conduct analysis to identify exposures associated with the disease. It is an efficient design for studying rare diseases.	Confounding by unknown factors (see above).
Cohort	In cohort studies, a group of individuals (cohort) is followed over time to assess the relationship between exposures and health outcomes. These studies can be prospective (following individuals forward in time) or retrospective (using existing data to follow individuals backward in time).	Particularly useful for studying the effects of long-term and multiple exposures, as well as investigating the role of critical periods and windows of susceptibility in life-course epidemiology.	Can be time-consuming, expensive, and may be affected by attribution bias. Confounding also remains an issue
Nested case-control	This design is a hybrid of cohort and case-control designs, where cases and controls are identified from within an existing cohort study.	Has the advantages of case-control design and lower cost of exposure measurement due to a reduced sample size.	May induce inefficiency when matching cases. When multiple outcomes are investigated, a new set of controls is required for each disease.

the analysis. Chemicals released to the environment from a single source or generated from the same biochemical processes (eg, diesel combustion) are correlated and often detected as a cluster, and have been used as the footprint to identify sources of exposures.<sup>46,47</sup> Organic chemicals tend to have higher correlation than water-soluble compounds due to their lipophilicity and accumulation in organisms. Correlation of exposures are also higher between unit members in a shared environment, and this correlation increases further with longer duration of residence.<sup>48,49</sup> One of the methods to intuitively visualize the correlation structure is correlation globe,<sup>50</sup> which can be further developed to show differences between and within sex groups<sup>51</sup> (Figure 3). In longitudinal studies, the within-person correlation of repeated measurements of the same exposures is generally higher than the between-person correlations; however, other factors, such as the solubility and exposure trends of the chemicals, could play an important role for this observation.<sup>52,53</sup> The consequence is that it is difficult to identify the true contributor(s) for a given outcome in a statistical model. It also causes instability to model parameters and their precision (standard errors) from tiny changes in the input data due to multicollinearity.<sup>54</sup> In ExWAS and exposome research, it is essential to check model assumptions, potentially across multiple exposures. Further,

correlation decreases the effective sample size and statistical power of an analysis.

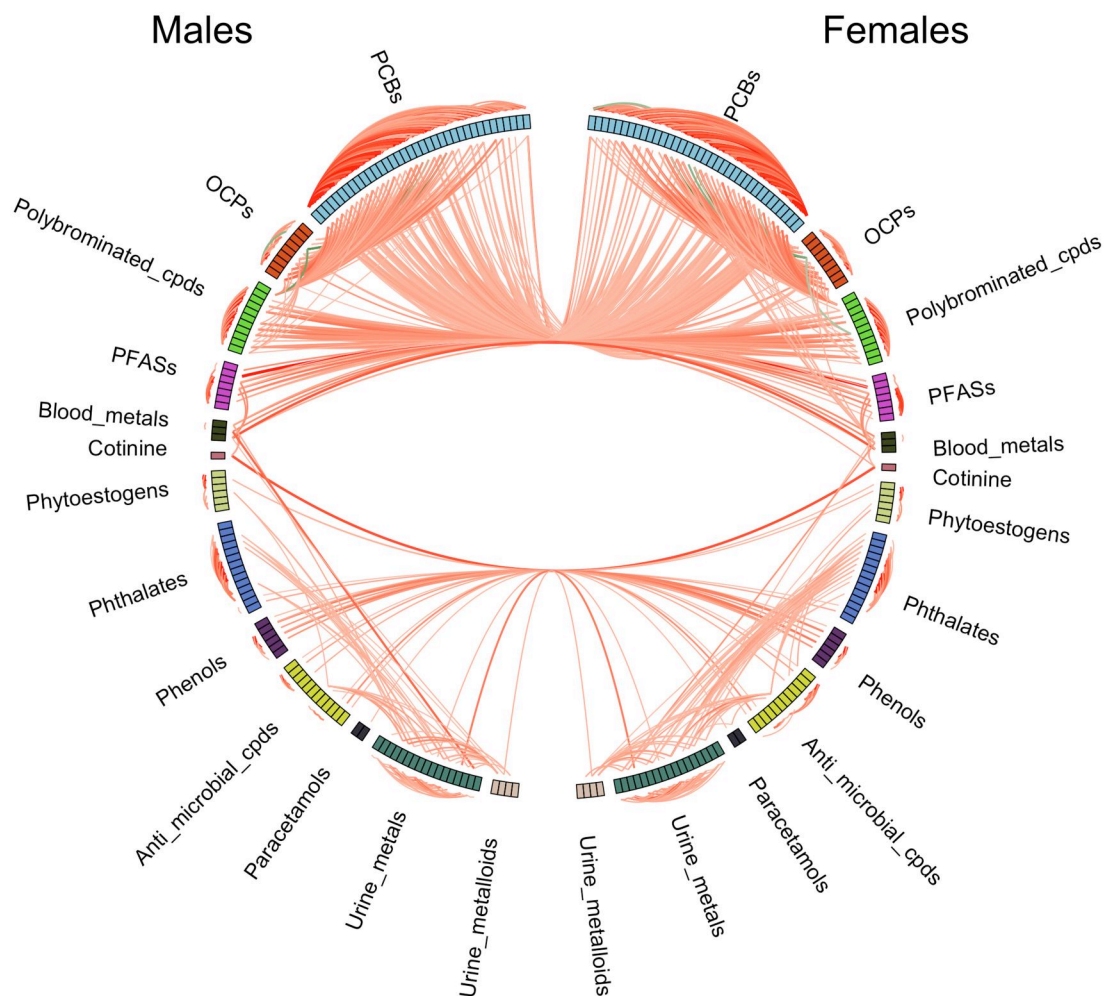
### ***The list of potential confounding variables is elusive and may need to be considered in a domain-specific manner***

As above, the exposome is densely correlated. A related issue, confounding, is common in exposomic observational studies. Confounding in the context of the exposome refers to a situation where the observed association between an environmental exposure (the exposome factor) and a health outcome is distorted by the presence of another variable, which is related to both the exposure and the outcome but is not an intermediate step in the causal pathway.

Potential confounders can influence both the exposure and the outcome variables and cause spurious associations in analysis, which can be controlled by including them in a regression model, or by stratifying among the hypothetical confounding variables. Nonetheless, a database of confounders is not available for ExWAS associations. There are several reasons for this, possibly. First, the phenomenon of confounding may differ from domain-to-domain, requiring analytic specifications to change for each association. Second, since the exposome is time-varying, there could be many types of sources of confounding, many of

Table 3. Key elements of selected exposome-wide association studies

Study	Data source	Study design	Sample size	Modality of exposures	Number of exposures	Phenotype	Model	Summary Statistic	Multiplicity Control Method
Zhang et al.	UK Biobank	Cohort	Over 500 000	7 categories of factors, eg. lifestyle, medical history, and socioeconomic status	210 modifiable factors	Dementia	Cox proportional hazard regression	Hazard ratio	Bonferroni
Shah et al.	Coronary Artery Risk Development in Young Adults (CARDIA)	Cohort	5 115	17 Food and beverage groups, 2 nutrient groups, and metabolome	19 dietary factors and metabolome	Cardiometabolic-cardiovascular disease	Linear regression	Beta coefficient	FDR
van de Weijer et al.	Geoscience and Health Cohort Consortium and the Netherlands Twin Register	Longitudinal national register	21 926	Multiple domains including social, physical, and demographic variables	133 variables	Well-being	Generalized estimating equations	Beta coefficient	Bonferroni
Saberi Hosnijeh et al.	European Prospective Investigation into Nutrition and Cancer cohort (EPIC)	Cohort	475 426	Various anthropometry measures and lifestyle factors	84 variables	B-cell lymphoma	Cox proportional hazard regression	Hazard ratio	FDR
Juivez et al.	A multi-centric birth cohort study in 6 European countries	Cohort	1298 mother-child pairs	19 categories of factors such as metals, built environment, and traffic	209 variables	Child cognitive function	Linear regression	Beta coefficient	Bonferroni
Jedynak et al.	5 different European cohorts	Cohort	708 mother-child pairs	Environmental chemicals	47 variables	Child behavior	Negative binomial regression Logistic regression	Incidence rate ratio Odds ratio	Bonferroni FDR
Uche et al.	NHANES 1999–2016	Multiple cross-sectional surveys	50 048	All available and useable environmental factors such as metals, allergens, toxins, allergens, and nutrients	more than 200 variables	Obesity	Logistic regression	Odds ratio	FDR
Milanlouei et al.	Nurses' Health Study	Cohort	62 811	Dietary factors	257 nutrients and 117 foods	Coronary heart disease	Cox proportional hazard regression	Hazard ratio	FDR
Granum et al.	A multi-centric birth cohort study in 6 European countries	Cohort	1270 mother-child pairs	18 exposure groups during pregnancy and at the subcohort follow-up, such as built environment, air pollution, traffic, and road traffic noise	197 variables	Allergy related outcomes in childhood	Logistic regression	Odds ratio	Bonferroni
Elhadad et al.	NHANES III (1988–1994)	Multiple cross-sectional surveys	17 752	Metabolites, nutrients, and lifestyle factors	245 variables	Coffee consumption	Linear regression	Beta coefficient	FDR
Choi et al.	UK Biobank	Cohort	over 100 000	Modifiable factors such as behavioral, social, and environmental	106 variables	Depression	Logistic regression	Odds ratio	Bonferroni



**Figure 3.** A correlation globe showing the associations among chemical biomarkers for females, males, and couples. The right half of the globe represents female biomarkers, while the left half represents male biomarkers. Only correlations greater than 0.25 or smaller than  $-0.25$  are displayed as connections. A red line signifies a positive correlation, whereas a dark green line represents a negative correlation. Both color intensity and line width correspond to the magnitude of the correlation. Reproduced from Chung et al.,<sup>51</sup> used under CC-BY-NC-ND 4.0 license.

which have not been identified. The exposome includes a vast array of factors across domains. This contrasts with GWAS, where “one” central confounder has been identified, known as population stratification, which describes how genetic variant frequency relates to ancestry. Untangling ancestry versus variant specific effects is achieved by stratifying GWAS analysis by ancestral groups, or accounting for ancestry in the regression model, and this adjustment does not need to change per genetic factor.

We give an example: suppose a ExWAS discovers a correlation between a certain environmental chemical and increased rates of a health condition. However, if individuals with the exposure also share a common lifestyle factor (like smoking), which independently increases the risk of the health condition, smoking becomes a confounding variable. It’s challenging to determine whether the health condition is due to the chemical, the smoking, or a combination of both. To partially address confounding in exposomic studies, comprehensive data collection is vital. This includes detailed information on a wide range of potential environmental exposures, as well as other demographic, genetic, and lifestyle factors that could influence health outcomes.

#### **Sparsity and missingness of exposure data**

The size of the chemical space (possible chemical species in the environment) is increasing.<sup>55</sup> However, individuals are typically

exposed to a small subset of this universe. If we measure the chemicals in human fluid or environmental samples and tabulate the results, low dose exposures will be the majority and many of the values in the data table for a specific chemical will be “missing” from a subset of samples, and have no values.<sup>56-58</sup> The sparsity could be caused by a lack of exposure or concentration that is too low to be detected, that is left censored data. For chronic diseases, it is often assumed that the individual effects of many exposures are marginal and impacts are attributed to the collective actions of a mixture<sup>57,59,60</sup>; however, deployment of modern machine learning techniques to analyze mixtures is impossible with non-random patterns of missingness. Left-censoring is one of the examples of missing not at random and is characterized by missing values that are below the limitation of detection of the measurement. An imputation method, Quantile Regression Imputation of Left-Censored data (QRILC), has been developed to impute unknown values.<sup>61</sup> The method works by sampling randomly from a truncated distribution of values predicted via quantile regression.

#### **Data processing**

Careful data preprocessing is essential given multi-modality is critical in an exposome study. Pre-processing refers to the various addition, removal, and transformation actions taken to make

raw data ready for statistical analysis and influences the interpretation and robustness of ExWAS outputs.

### Assembling the cohort for analysis

Data for a study are collected as specified by research protocols. Additional filtering steps to select eligible subjects into an analysis are common for large, general purpose observational studies or repurposing real-world data (eg, administrative healthcare records). If the analysis requires the integration of multiple independent datasets, variable harmonization is needed to increase comparability and interpretability of results.<sup>62-64</sup> Approach for harmonization generally involves assessing data dictionaries to identify common variables that have different recording formats. Then new variables are created through standardizing the measurement units and redefining levels in categorical variables. For instance, in one study, “ethnicity” is a categorical variable with levels “white,” “black,” “Hispanic,” and “others.” While in another study, the same variable contains “white,” “black,” “Hispanic,” “Asian,” and “others.” For compatibility, some investigators may merge “asian” into “others” for the new “ethnicity” variable and document how inconsistency is handled in the data dictionary.

### Data cleaning

When the information is gathered into a structured, tabular format, it is further processed to facilitate downstream statistical analysis. Variables (columns) with a high percentage of missing values, eg, >90%, could be removed to enhance reliability of analysis. For extreme values, data could be trimmed to remove a small percentage of subjects (rows) or replaced with a highest manually decided value for the variable, eg, age 99 can be substituted for age 137. However, trimming should be performed with caution to control for the level of biases introduced to the data.<sup>60,65</sup> Datasets with missing values are common across different fields in exposomics and missingness is caused by various reasons, from dropout of subjects in longitudinal analysis to beyond the reliable signal detection range in chemical measurement. Furthermore, when integrating data from different modalities (eg, different assays), often some assays will be measured and not others for a subset of the participants. Different imputation methods are available and the choice is typically based on the investigator’s knowledge about the missing mechanism (eg, missing completely at random, missing at random, censored), types of analysis and data, and imputation performance. In summary, data cleaning steps may include procedures such as handling missing values, detecting and managing outliers, removing duplicates, and binning variables. The aim is to enhance the quality of data, ensuring both internal validity (plausible values and ranges) and external validity (comparable units). This, in turn, enhances the interpretability of downstream statistical analyses.

### Transforming variables

Often, data are transformed prior to analysis to adhere to model assumptions and to enhance result interpretation. It is particularly important in exposome analyses where different exposures will have different units of exposure and differing prevalence of exposure. In a linear model, predictor ( $X$  variable) can be log-transformed to reduce the influence of extremely large values without trimming or substitution, whereas the same transformation is applied to log-normally distributed outcome ( $Y$  variable) as a simple way to fulfill the normality assumption of errors. A “fudge factor,” for example, +1, is added to zero value when log transformation is required.<sup>29</sup> For other non-normally distributed

variables, Box-Cox transformation<sup>66,67</sup> or an inverse normalized function<sup>68</sup> are options. Since multimodal data could have variables with different units and a large range of absolute magnitude, we can conduct z-score standardization<sup>69</sup> to make comparisons possible (eg, each variable is in 1 SD unit of continuous exposure). In the machine learning context, categorical variables are often required to be one-hot encoded (ie, transforming the levels of a categorical variable into new individual variables) prior to modeling.

### Other considerations

#### Analytic outputs: effect sizes, correlation, and odds ratios

In statistics, an effect size, or association size, quantitates the relationship between two variables. A simple example is correlation, which measures the relatedness between two variables, that is their tendency to vary together. Typical quantification metrics include Pearson’s product-moment coefficient ( $r$ ) for linear relationship and more generally Spearman’s rank correlation coefficient ( $r_s$ ) for any monotonic relationship. It is a unitless measure with a range of  $-1$  to  $1$ . A negative correlation suggests that variables are changing in opposite directions, and a zero correlation means there is no relationship. In ExWAS, absolute correlations are usually low, but heterogeneous, with values below  $0.25$ .<sup>50,51,70</sup> When comparing the effect sizes of two events, we can use absolute measures including mean risk difference, or relative terms such as relative risk and odds ratios. A magnitude equal to one for the relative measures indicates that the exposure does not affect the outcome, while a value smaller than one means a protective effect from the exposure, and a harmful effect for a bigger than one value. In a linear regression model, a regression coefficient denotes the average change in dependent variable per unit change in the corresponding independent variable.

#### Inference versus prediction

A statistical model is built to describe the relationships between the variables of interest and can be used to draw inference or to make predictions. The majority of public health studies focus on collecting representative samples from a population and constructing models in order to draw conclusions to support policy formulation. Conversely, in biomedical studies, models are often used to predict the outcomes of individuals through using their corresponding measurements as predictors. The predicted outcome can further aid in diagnosis and prognosis of diseases.

In order to build a regression model for inference in an ExWAS study, researchers first need to choose the right regression model to model a quantitative or binary outcome. Second, the analyst must decide if they want to transform the outcome variable so it adheres to the requirements of regression (eg, normally distributed for continuous outcomes). Then, one needs to identify suitable exposure and outcome variables for the study question, and include other potentially confounding variables based on domain knowledge to minimize distortion of the association between the variables of interest. For instance, secondhand smoke exposure was adjusted for the associations between short-term ozone exposure and platelet activation and blood pressure increases.<sup>71</sup> Next, a model is chosen based on the nature of the data and the hypothesis. Before fitting the model, the data must be inspected and cleaned to ensure a valid interpretation of the modeling statistics. The model is usually fitted using ordinary least squares or maximum likelihood estimation methods. Afterward, it is necessary to check model assumptions such as normality and

homoscedasticity through diagnostic tests and plots for every exposure phenotype association.

In predictive modeling, data are split into training and test sets with a ratio between 8:2 to 5:5. Using the training set, variables that have strong influence on the outcome are selected as predictors, and fit statistics such as  $R^2$  is calculated to assess how well the model describes the data. Such a model can be optimized iteratively, and the performance of candidate models is often evaluated through cross-validation, which is a resampling procedure utilizing different segments of the training dataset for testing and tuning a model across multiple iterations.<sup>72</sup> Increasingly, many complex machine learning algorithms are available, and they are often treated as “black boxes” when compared with linear-based regression models. Most of the time, research questions involve binary classification and the prediction characteristics are visualized with a receiver operating characteristic (ROC) curve.<sup>73</sup> Alternative models can be evaluated using performance indicators such as AUC. Model overfitting can occur and the generalizability of its predictive performance is assessed using the test dataset. The concept of bias-variance tradeoff emphasizes the importance of finding an optimal balance between simplicity (to prevent overfitting) and accuracy (to effectively capture the underlying patterns in the data) in a machine learning model.<sup>74</sup> No simple solution is known for this problem, but techniques such as cross-validation could help to detect it early in the model building process.

## Variable selection, reproducibility, and mitigation of the exposome-wide false discovery rate

In exposome research and ExWAS, the analyst is attempting to relate a vast array of environmental factors with an outcome, known as *variable selection*. However, the more tests in ExWAS, the higher the chance that some of the findings (indicating a significant effect) are actually just due to random chance. This is known as a false positive or a false discovery. False positives are a threat to *reproducibility* of associations. Traditional high-throughput inference techniques include the Bonferroni correction (simultaneous inference) or FDR control (inferring over the average of those that are selected). The prior adjusts the p value thresholds conservatively, attempting to reduce the probability of selecting one false positive. The former is more lenient, trying to ensure that the *average* of false discoveries among all variables selected is controlled.

Practically, these approaches address multiple hypothesis tests by correcting, or adjusting the significance threshold to account for multiple testing; however these estimates make inferences without accounting for the other exposome variables that are associated with the outcome. For example, the Bonferroni correction, known as a “family-wide error” rate correction, simply corrects the pvalue threshold from the standard 0.05 to 0.05 divided by the number of tests, or the number of exposome variables/factors that are being modeled. This leads to a challenging question: how many exposures can be analyzed in an ExWAS, and what would the denominator be?

### False discovery rate estimation

The FDR method was introduced by Benjamini and colleagues.<sup>21,75</sup> The FDR is essentially the expected proportion of false discoveries among all the discoveries made. For instance, if you perform 100 tests and 10 of them show significant results, the FDR can help estimate how many of those 10 are likely to be

false positives. The FDR allows researchers to control the rate of these false discoveries, reducing the likelihood of mistakenly identifying an environmental factor as influential when it's not. Exposome factors being tested are often not independent of each other. For example, exposure to one pollutant might be correlated with exposure to another. Traditional FDR methods assume each test is independent, but this assumption doesn't hold in many practical scenarios. Recognizing this, more advanced FDR methods have been developed that take into account the correlation between tests. These methods understand that finding a significant result in tests that are correlated is different from finding one in tests that are independent. By factoring in these correlations, these methods provide a more nuanced and accurate estimation of the FDR.

For instance, if several environmental factors are correlated, a discovery in one may increase the likelihood of a discovery in another. Advanced FDR methods may be able to consider the correlation (eg, the Benjamini-Yekutieli approach,<sup>76</sup> or an empirical permutation-based approach<sup>77</sup>), ensuring that the overall rate of false discoveries remains controlled, even in the presence of these correlations.

The simplest approach to avoid false positives may be through validation in a held-out dataset, known as “sample splitting.”<sup>78</sup> In this procedure, a large dataset is split into at least two sub-datasets. Then, all variable selection procedures are executed in one of the datasets, and inference takes place in the second dataset.<sup>79</sup> Extensions of studies, such as “hierarchical” testing, for example, testing hypotheses at different levels of variables, may be especially appropriate in ExWAS, where exposure groups might be nested by behavior (eg, smoking behavior and biomarkers of smoking).

### Variable selection during prediction

A common machine learning task includes combining a “variable selection” procedure to identify groups of variables that maximize predictive power as a collective. Although predictive performance is generally correlated with the number of model variables, investigators should consider a balance between interpretability (simpler models) and accuracy of the final predictive model. One popular procedure in ‘omics research includes shrinkage approaches<sup>80</sup> (also known as regularization). The LASSO procedure (which stands for Least Absolute Shrinkage and Selection Operator)<sup>81,82</sup> is an algorithm that “shrinks” or even sets some of the less important feature variables to zero, essentially removing them from consideration. By focusing only on the most significant features and ignoring the less relevant ones, LASSO prevents our model from getting too attached to the noise or irrelevant details, or too many variables, in the data. This approach helps reduce overfitting, ensuring our analysis or model is more robust and generalizes better to new, unseen data. Procedures similar to LASSO ( $\ell_1$  penalty) include ridge ( $\ell_2$  penalty) and elastic net<sup>83,84</sup> regression ( $\ell_1 + \ell_2$  penalties). There are also other variants,<sup>85-87</sup> such as Group LASSO<sup>88</sup> (selecting groups of predictors rather than individual variables), Sparse LASSO (optimized to select a small number of critical variables in high-dimensional data), and Sparse Group LASSO (selecting important groups of variables and also variable within groups).

It is important to emphasize that the ExWAS study design, at best, yields exposures that are reliably correlated with, but not necessarily causal of, a phenotype of interest. Given the dense correlational relationships between exposure factors, phenotypes, as well as often pervasive biases in ascertainment,

sampling, and survey weighting, ExWASs are often just the first step in the process of winnowing a long list of exposures to then be analyzed for potential causal relationships with the phenotype of interest. Typically, after important variables are identified, one may conduct further studies to get better understanding between the selected variable and the outcome. These include but are not limited to: meta-analysis for reproducibility of the finding, mediation analysis for mechanistic insights, Mendelian randomization analysis for causal inference, and even molecular experiment to demonstrate the effects of the variables.

### Sample versus variable size

In omics studies, a very large number of variables (genes, proteins, and metabolites, etc.) are typically related to an outcome. Sample sizes range from hundreds to thousands; only recently have we seen large scale multi-omics in cohorts such as UK Biobank. In typical cohort scenarios, however, smaller sample sizes creates an analytics scenario known as “large p (number of variables), small n (sample size).” Statistical models built with data whose dimension is larger than the sample sizes are prone to overfitting (ie, fitting the noise rather than the underlying signals). While the models can perform well with the training data, they have low generalizability and typically fail to reproduce the performance when fed with new data. Other issues include the multiple testing (greater type I error rate) and collinearity (high correlation) between variables when selecting impacting variables. Mixture analysis might also be underpowered owing to the small effect size of individual exposures. Sample size requirement for an ExWAS depends on the numbers and general effect sizes of exposures of interest and can be estimated using simulation. For example, in a post hoc power analysis with over 120 endocrine disrupting chemicals, we estimated that a sample size of ~2700 is needed to achieve a statistical power of 0.8,<sup>89</sup> which means that if there is a true effect, the test has an 80% probability of detecting it.

Conversely, in nationwide scale analysis, various questionnaire-based and targeted external measurements are integrated through personal identifiers or geocodes, and sample size could be in the millions or more. Overpowered associations are a major issue in this scenario.<sup>90-92</sup> The huge sample size decreases the standard error, amplifying the ability to detect even miniscule differences in effect size. In extreme cases, the interpretation could overemphasize statistical differences in effect sizes that lack a clear biological relevance or may be “residually” confounded, eg, a 0.001% increase in an outcome for an unit change of an exposure. If the dataset is split and randomized, a pair of exposure-outcome relationships could be both statistically significant in the subsets, but with a flip of the directionality of effects, making it even harder to interpret the results.

## Advanced methods in analyzing the exposome

We have covered basic concepts for exposome study and conducting an ExWAS. However, more advanced topics and methods are available to gain insights on the complex exposure-disease relationships based on the study context and research questions. Many of them are extensions of, or involve the application of, the ExWAS approach and are discussed below.

## Methods to incorporate study design features: survey sampling, repeated measures, and time-to-events

### Survey sampling

Sampling is the process of selecting a representative subset of individuals from a population in order to obtain population estimates. Commonly used methods include simple random sampling and stratified sampling. In contrast to conducting phone interviews, large-scale nationwide studies involving physical interaction (eg, in person interview and examinations) will require significant resources and pose logistical challenges if simple random sampling is used to identify participants. Therefore, complex multistage survey design is employed. It is best to illustrate the concept with NHANES, a bi-yearly study conducted by the Centers for Disease Control and Prevention. To provide a representative sample of the US population, a 4-stage survey design is used. To begin, primary sampling units (PSUs) mostly in county level are first selected (Stage 1), and segments, generally in city block level, within PSUs are subsequently sampled (Stage 2). Households within segments are randomly drawn (Stage 3), and finally individuals were selected at random in households (Stage 4). To obtain correct population statistics, software or statistical packages designed to incorporate PSU, strata, and survey weight information of the study must be used. An example is the investigation of relationships between 27 physiological markers and mortality in multiple NHANES survey cycles by Nguyen et al.<sup>93</sup>

### Mixed linear modeling to account for repeated measures

Many statistical tests assume independence between observations. Violation generally causes the standard error and confidence interval of the estimates to be smaller, and thus deflating standard errors and increasing the chances of false findings. In practice, weak and random correlations are almost always observed, but this assumption is still largely valid. However, when sources of correlation are known in the study design, they can be considered by the model to control for spurious findings. Two typical correlational sources are clustering of individuals and repeated measurement of the same individuals over time. Clustering occurs when individuals are sampled through specific locations or institutions, such as enrollment of students for IQ tests through schools and patients for a disease study through hospitals. These designs have the advantages of reducing cost of sampling and variability of the data. Correlated data are analyzed with a mixed effect model<sup>94</sup> where the correlating hierarchical or repeating unit is modeled as the random effect and other variables of interest are called fixed effects, for example, outcome, predictor, and potentially confounding variables. Alternatively, a generalized estimating equation method can be used if only population-averaged effects are concerned.<sup>18,95,96</sup>

### Time to event outcomes

In a longitudinal study, individuals are followed over time and therefore time-to-event data is available. For example, a study on early life lead exposure and the later development of learning and behavior problems in children. We can conduct survival analysis<sup>97</sup> to understand the relationship between an exposure to a delayed outcome. Specifically, a Cox proportional-hazards model can handle multiple predictors and estimate the hazard ratio for each predictor. In addition, because Cox model is still a regression based method, it fits into the ExWAS analytical framework for conducting exposome-level analysis.

## Mixture analysis and additive polyexposure scores

Methods we previously introduced for variable selection can also be applied to identify important contributors to an outcome in mixture exposure settings. However, one of the significant limitations is that only the effects of individual exposures are known. This issue becomes more prominent when it is believed that concentration and impact of individual exposures are low but collectively they may have meaningful biological perturbation at molecular level or even an association to a clinical outcome.<sup>87</sup> To address this problem, we can apply weighted quantile sum regression.<sup>86</sup> The approach creates a summary score of the mixture for each individual and assesses the relationship between the scores and the outcome; however, the new score is a challenge to interpret. On the other hand, the model also estimates weights of individual exposures to the score, thus also enabling the identification of significant individual contributors to the overall mixture effects. Quantile-based g-Computation<sup>85</sup> is a technique integrating weighted quantile sum (WQS) regression with g-computation. Same as the original WQS regression method, it can estimate the overall mixture effect while the parameters are calculated using a marginal structural model instead of standard regression.

Similarly, polyexposure scores (PXS)<sup>73,98-101</sup> provide an alternative way to summarize the individual exposure risks to a disease, which are typically weak and nonstatistically significant, into a single predictive index for each subject. Building a PXS involves splitting the original data into three different subsets (training, validation, and testing) and employing multiple variable selection steps (eg, ExWAS and LASSO) to identify significant exposure factors to an outcome. Advanced methods are also available for different mixture exposure situations, such as Bayesian kernel machine regression<sup>80,102</sup> and boosted regression trees<sup>103</sup> for nonlinear response modeling and interaction screening. Identifying the optimal method for detecting health impacts of mixture exposures in the context of exposomics is an active research area and a wide range of methods has been discussed by others in a consortium setting.<sup>83,104-108</sup>

## Going forward: infrastructure to support standardization of ExWAS and terminology

Exposomics and ExWAS is a fast growing field. GWAS were made possible by the standardization of the ways that genetic factors are digitally represented (eg, as genotypes with standard identifiers) and the ease of which samples may be collected (eg, in a case-control fashion) due to the lack of unknown confounding and static nature of genetic factors and genotypes. Standardization of ExWASs will be possible, but require advancement in not only analytic standards but also will depend on study design characteristics, some of which we articulated in this paper.

We foresee that a few emerging topics will be crucial to facilitate standardization of ExWAS studies in the future. First, the prevalence of open data and technologies to access these data, such as cloud computing will be beneficial. Since 2023, NIH requires all grant applications to submit a data management and sharing plan under a new policy.<sup>109</sup> It is becoming a standard for study funders/providers to share their data via FAIR Data Principles—Findability, Accessibility, Interoperability, and Reusability of digital assets.<sup>110</sup>

Second, administrative healthcare data, such as electronic health records and administrative claims, contains comprehensive, codified, and longitudinal information about an individual's health and disease status and drug prescription. These data have been instrumental in geospatial environmental health studies.<sup>9,111-114</sup> Unlike data collected for observational cohorts, these datasets are not created for research.<sup>115-117</sup> Data could be coming from a few major health care centers, and variations due to style of practices have to be considered. In addition, records are triggered by the severity of illnesses. An absence of disease records does not necessarily mean that the patient is disease/symptoms free. Repurposing these records for research requires that investigators have an understanding of the healthcare practice and coding system in order to draw valid conclusions from the analysis.

Third, the functional exposome encompasses a subset of biologically active exposome.<sup>5</sup> Semi-agnostic methods for functional exposomics, which are different from targeted and non-targeted measurement approaches, could become the mainstream driving molecular exposure and multi-omics analysis. These methods include semi-targeted analysis,<sup>118-120</sup> suspected screening,<sup>121-123</sup> adductomics,<sup>124-127</sup> and affinity-based measurement<sup>128</sup> and they are characterized by striking a balance between throughput and interpretability of measurement. An understanding of data generation processes is essential to ensure correct application of analytical methods and interpretation of results. Furthermore, the application of the functional exposome concept to the One Health concept<sup>129</sup> could enable holistic and integrated studies between humans, animals, and the external environment.

Finally, the term environment-wide association study (EWAS) was coined by Patel et al. in 2010,<sup>16</sup> but variants are common, including environmental-wide association study,<sup>130</sup> exposome-wide association study,<sup>89</sup> exposure-wide association study<sup>17</sup> with acronyms such as XWAS, EnWAS, and ExWAS. To further complicate the issue, EWAS is also an acronym for epigenome-wide association study, which appears in the literature around the same period of time. The ambiguity makes it challenging to search for relevant studies and creates confusion among researchers. In light of this, we propose to standardize the nomenclature with the term “exposome-wide association study, ExWAS”, pronounced as “x-wahz”, for any data-driven study that associates multiple and diverse exposome-based exposures with a phenotype or multiple phenotypes and involves correction for multiple comparisons and elucidation of replication. In discipline-specific analyses, such as nutrient-wide association study, and drug-wide association study, ExWAS could be tagged as a keyword for effective paper retrieval during literature reviews.

Uncovering the contribution of the environment to diseases presents a significant challenge and requires advancements in both measurement technologies and data analytical methods. From a data scientist's perspective, embracing the use of large cohorts and repurposed datasets, along with the application of the latest analytical techniques, could enable novel discovery of the elusive relationships between the environment and diseases.

## Acknowledgments

We thank Heidi Hanson and Ander Wilson for providing comments to improve the manuscript. We thank the participants of the NIEHS Exposome Workshop (<https://factor.niehs.nih.gov/>)

2022/9/feature/2-feature-exposomics-research), and the Exposomics Consortium (<https://www.exposomicsconsortium.org>)

## Funding

M.K.C. and C.J.P. were supported by grants from the U.S. National Institutes of Health through the National Institute for Environmental Health Sciences (ES032470, P30ES000002), the National Institute on Aging (AG074372), and from the U.S. National Science Foundation through the Northeast Big Data Innovation Hub. A.M.R., J.S.H., and F.S.A. are supported by intramural funds from the National Institute of Environmental Health Sciences. M.A.L. was supported by grants from the U.S. Environmental Protection Agency (G17D112354237), from the U.S. National Institutes of Health through the National Institute of Diabetes and Digestive and Kidney Diseases (R01DK125586), and from the U.S. Department of Veterans Affairs (HX002680).

## Author contributions

Ming Kei Chung (Conceptualization [equal], Project administration [equal], Writing—original draft [equal], Writing—review & editing [equal]), John S. House (Writing—original draft [equal], Writing—review & editing [equal]), Farida S. Akhtari (Writing—original draft [equal], Writing—review & editing [equal]), Konstantinos C. Makris (Conceptualization [equal], Writing—review & editing [equal]), Michael A. Langston (Writing—review & editing [equal]), Khandaker Islam (Writing—review & editing [equal]), Philip Holmes (Writing—review & editing [equal]), Marc Chadeau-Hyam (Conceptualization [equal], Writing—review & editing [equal]), Alex I. Smirnov (Writing—review & editing [equal]), Xiuxia Du (Writing—review & editing [equal]), Anne E. Thessen (Conceptualization [equal], Writing—review & editing [equal]), Yuxia Cui (Writing—review & editing [equal]), Kai Zhang (Conceptualization [equal], Writing—review & editing [equal]), Arjun K. Manrai (Conceptualization [equal], Writing—review & editing [equal]), Alison A. Motsinger-Reif (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Chirag J. Patel (Conceptualization [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Members of the Exposomics Consortium\* (Conceptualization [equal]).

## Data availability

This study does not generate new data or reanalyze any existing datasets.

## Conflict of interest statement

The authors declare that they have no conflicts of interest.

## References

- Wild CP. The exposome: from concept to utility. *Int J Epidemiol.* 2012;41(1):24–32.
- Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev.* 2005;14(8):1847–1850.
- Rappaport SM, Smith MT. Epidemiology. Environment and disease risks. *Science.* 2010;330(6003):460–461.
- Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: Where chemistry meets biology. *Science.* 2020;367(6476):392–396.
- Chung MK, Rappaport SM, Wheelock CE, 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.* 2021;129(8):85001.
- Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. *Toxicol Sci.* 2014;137(1):1–2.
- Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009;461(7265):747–753.
- Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. *Nat Rev Genet.* 2008;9(4):255–266.
- Lakhani CM, Tierney BT, Manrai AK, Yang J, Visscher PM, Patel CJ. Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nat Genet.* 2019;51(2):327–334.
- Polderman TJC, Benyamin B, de Leeuw CA, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet.* 2015;47(7):702–709.
- Hu X, Walker DI, Liang Y, et al. A scalable workflow to characterize the human exposome. *Nat Commun.* 2021;12(1):5575.
- Boyce M, Favela KA, Bonzo JA, et al. Identifying xenobiotic metabolites with in silico prediction tools and LCMS suspect screening analysis. *Front Toxicol.* 2023;5:1051483.
- Patel CJ, Claypool KT, Chow E, et al. The demographic and socioeconomic correlates of behavior and HIV infection status across sub-Saharan Africa. *Commun Med (Lond).* 2022;2:104.
- Rappaport SM. Biomarkers intersect with the exposome. *Biomarkers.* 2012;17(6):483–489.
- MacKinnon DP. *Introduction to Statistical Mediation Analysis (Multivariate Applications Series)*. 1st ed. Routledge; 2008.
- Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One.* 2010;5(5):e10746.
- Zhang Y, Chen SD, Deng YT, et al. Identifying modifiable factors and their joint effect on dementia risk in the UK Biobank. *Nat Hum Behav.* 2023;7(7):1185–1195.
- van de Weijer MP, Baselmans BML, Hottenga JJ, et al. Expanding the environmental scope: an environment-wide association study for mental well-being. *J Expo Sci Environ Epidemiol.* 2022;32(2):195–204.
- Shah RV, Steffen LM, Naylor M, et al. Dietary metabolic signatures and cardiometabolic risk. *Eur Heart J.* 2022;44(7):557–569.
- Smith GD, Ebrahim S. Data dredging, bias, or confounding. *BMJ.* 2002;325(7378):1437–1438.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 1995;57(1):289–300.
- Ioannidis JPA, Khoury MJ. Improving validation practices in “omics” research. *Science.* 2011;334(6060):1230–1232.
- Pearson TA, Manolio TA. How to interpret a genome-wide association study. *J Am Med Assoc.* 2008;299(11):1335–1344.
- Ioannidis JPA, Loy EY, Poulton R, Chia KS. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci Transl Med.* 2009;1(7):7ps8.
- Ragoussis J. Genotyping technologies for genetic research. *Annu Rev Genomics Hum Genet.* 2009;10:117–133.
- Uche UI, Suzuki S, Fulda KG, Zhou Z. Environment-wide association study on childhood obesity in the U.S. *Environ Res.* 2020;191:110109.

27. Milanlouei S, Menichetti G, Li Y, et al. A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease. *Nat Commun*. 2020;11(1):6074.
28. Amiri M, Lamballais S, Geenjaer E, et al. Environment-Wide Association Study (En WAS) of prenatal and perinatal factors associated with autistic traits: a population-based study. *Autism Res*. 2020;13(9):1582-1600.
29. Chung MK, Smith MR, Lin Y, et al. Plasma metabolomics of autism spectrum disorder and influence of shared components in proband families. *Exposome*. 2021;1(1):osab004. <https://doi.org/10.1093/exposome/osab004>.
30. Choi KW, Wilson M, Ge T, et al. Integrative analysis of genomic and exposomic influences on youth mental health. *J Child Psychol Psychiatry*. 2022;63(10):1196-1205.
31. Elhadad MA, Karavasiloglou N, Wulaningsih W, et al. Metabolites, nutrients, and lifestyle factors in relation to coffee consumption: an environment-wide association study. *Nutrients*. 2020;12(5), 1470. <https://doi.org/10.3390/nu12051470>.
32. Andrianou XD, Konstantinou C, Rodríguez-Flores MA, Papadopoulos F, Makris KC. Population-wide measures due to the COVID-19 pandemic and exposome changes in the general population of Cyprus in March-May 2020. *BMC Public Health*. 2022;22(1):2279.
33. Chadeau-Hyam M, Bodinier B, Elliott J, et al. Risk factors for positive and negative COVID-19 tests: a cautious and in-depth analysis of UK biobank data. *Int J Epidemiol*. 2020;49(5):1454-1467.
34. Jedynak P, Maitre L, Guxens M, et al. Prenatal exposure to a wide range of environmental chemicals and child behaviour between 3 and 7 years of age—An exposome-based approach in 5 European cohorts. *Sci Total Environ*. 2021;763:144115.
35. Amal S, Safarnejad L, Omiye JA, et al. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Front Cardiovasc Med*. 2022;9:840262.
36. Kline A, Wang H, Li Y, et al. Multimodal machine learning in precision health: A scoping review. *NPJ Digit Med*. 2022;5(1):171.
37. Patel CJ, Kerr J, Thomas DC, et al. Opportunities and challenges for environmental exposure assessment in population-based studies: exposures and gene-environment interaction. *Cancer Epidemiol Biomarkers Prev*. 2017;26(9):1370-1380. <https://aacrjournals.org/cebpa/article-abstract/26/9/1370/71218>.
38. Chung MK, Patel CJ. The exposome: An approach toward a comprehensive study of exposures in disease. In: Nriagu J (Ed.) *Encyclopedia of Environmental Health*. Elsevier; 2019:0-779.
39. Miller GW. Integrating the exposome into a multi-omic research framework. *Exposome* 2021;1(1):osab002. <https://doi.org/10.1093/exposome/osab002>.
40. CAS Databases. CAS. Accessed December 11, 2023. <https://www.cas.org/support/documentation/cas-databases>.
41. Naggara O, Raymond J, Guilbert F, et al. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *Am J Neuroradiol*. 2011;32(3):437-440.
42. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006;25(1):127-141.
43. Taylor JMG, Yu M. Bias and efficiency loss due to categorizing an explanatory variable. *J Multivar Anal*. 2002;83(1):248-263.
44. Patel CJ, Ioannidis JPA. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J Epidemiol Community Health*. 2014;68(11):1096-1100.
45. Patel CJ, Ioannidis JPA. Studying the elusive environment in large scale. *JAMA*. 2014;311(21):2173-2174.
46. Schauer JJ, Rogge WF, Hildemann LM, et al. Source apportionment of airborne particulate matter using organic compounds as tracers. *Atmos Environ*. 2007;41:241-259.
47. Chung MK, Hu R, Cheung KC, Wong MH. Pollutants in Hong Kong soils: polycyclic aromatic hydrocarbons. *Chemosphere*. 2007;67(3):464-473.
48. James RA, Hertz-Picciotto I, Willman E, Keller JA, Charles MJ. Determinants of serum polychlorinated biphenyls and organochlorine pesticides measured in women from the child health and development study cohort, 3-1967. *Environ Health Perspect*. 2002;110(7):617-624.
49. Wu XM, Bennett DH, Calafat AM, et al. Serum concentrations of perfluorinated compounds (PFC) among selected populations of children and adults in California. *Environ Res*. 2015;136:264-273.
50. Patel CJ, Manrai AK. Development of exposome correlation globes to map out environment-wide associations. *Pac Symp Biocomput*. 2015;231-242. [https://www.worldscientific.com/doi/abs/10.1142/9789814644730\\_0023](https://www.worldscientific.com/doi/abs/10.1142/9789814644730_0023).
51. Chung MK, Kannan K, Louis GM, Patel CJ. Toward capturing the exposome: exposure biomarker variability and coexposure patterns in the shared environment. *Environ Sci Technol*. 2018;52(15):8801-8810.
52. Makey CM, McClean MD, Sjödin A, et al. Temporal variability of polybrominated diphenyl ether (PBDE) serum concentrations over one year. *Environ Sci Technol*. 2014;48(24):14642-14649.
53. van der Meer TP, Chung MK, van Faassen M, et al. Temporal exposure and consistency of endocrine disrupting chemicals in a longitudinal study of individuals with impaired fasting glucose. *Environ Res*. 2021;197:110901.
54. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology*. 2016;6(2). <https://www.omicsonline.org/open-access/multicollinearity-in-regression-analyses-conducted-in-epidemiologic-studies-2161-1165-1000227.php?aid=69442>.
55. Llanos EJ, Leal W, Luu DH, et al. Exploration of the chemical space and its three historical regimes. *Proc Natl Acad Sci U S A*. 2019;116(26):12660-12665.
56. Aurich D, Miles O, Schymanski EL. Historical exposomics and high resolution mass spectrometry. *Exposome*. 2021;1(1). <https://academic.oup.com/exposome/article/1/1/osab007/6491249>.
57. Escher BI, Stapleton HM, Schymanski EL. Tracking complex mixtures of chemicals in our changing environment. *Science*. 2020;367(6476):388-392.
58. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The blood exposome and its role in discovering causes of disease. *Environ Health Perspect*. 2014;122(8):769-774.
59. Liew Z, Guo P. Human health effects of chemical mixtures. *Science*. 2022;375(6582):720-721.
60. Caporale N, Leemans M, Birgersson L, et al. From cohorts to molecules: Adverse impacts of endocrine disrupting mixtures. *Science*. 2022;375(6582):eabe8244.
61. Comprehensive R Archive Network (CRAN). *A Collection of Methods for Left-Censored Missing Data Imputation [R Package imputeLCMD Version 2.1]*. 2022. Accessed December 11, 2023. <https://cran.r-project.org/web/packages/imputeLCMD/index.html>.

62. Rinaldi E, Stellmach C, Rajkumar NMR, et al. Harmonization and standardization of data for a pan-European cohort on SARS-CoV-2 pandemic. *NPJ Digit Med.* 2022;5(1):75.
63. Bennett SN, Caporaso N, Fitzpatrick AL, et al.; GENEVA Consortium. Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. *Genet Epidemiol.* 2011;35(3):159-173.
64. Schaap LA, Peeters GM, Dennison EM, et al.; EPOSA Research Group. European Project on OsteoArthritis (EPOSA): methodological challenges in harmonization of existing data from five European population-based cohorts on aging. *BMC Musculoskelet Disord.* 2011;12:272.
65. Ramsey PH, Ramsey PP. Optimal trimming and outlier elimination. *J Mod Appl Stat Methods.* 2007;6(2):2.
66. Sakia RM. The box-cox transformation technique: a review. *Statistician.* 1992;41(2):169.
67. Osborne J. Improving your data transformations: applying the Box-Cox transformation. *Prac Assess Res Evaluation.* 2019;15(1):12.
68. Beasley TM, Erickson S, Allison DB. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav Genet.* 2009;39(5):580-595.
69. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. *J Mol Diagn.* 2003;5(2):73-81.
70. Tamayo-Uria I, Maitre L, Thomsen C, et al. The early-life exposome: description and patterns in six European countries. *Environ Int.* 2019;123:189-200.
71. Day DB, Xiang J, Mo J, et al. Association of ozone exposure with cardiorespiratory pathophysiologic mechanisms in healthy adults. *JAMA Intern Med.* 2017;177(9):1344-1353.
72. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol.* 2022;23(1):40-55.
73. He Y, Lakhani CM, Rasooly D, et al. Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of type 2 diabetes. *Diabetes Care.* 2021;44(4):935-943.
74. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc Natl Acad Sci U S A.* 2019;116(32):15849-15854.
75. Noble WS. How does multiple testing correction work? *Nat Biotechnol.* 2009;27(12):1135-1137.
76. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 2001;29(4):1165-1188. <https://doi.org/10.1214/aos/1013699998>.
77. Efron B. *Large-Scale Inference.* Cambridge University Press; 2010.
78. Wasserman L, Roeder K. High dimensional variable selection. *Ann Stat.* 2009;37(5A):2178-2201.
79. Benjamini Y. Selective inference: the silent killer of replicability. 2020;2(4). [https://hdsr.mitpress.mit.edu/pub/139rpgyc/lease/3](https://hdsr.mitpress.mit.edu/pub/139rpgyc/release/3).
80. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R.* Springer New York; 2013.
81. Tibshirani RJ, Taylor J. The solution path of the generalized lasso. *Ann Stat.* 2011;39(3):1335-1371.
82. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Series B Stat Methodol.* 1996;58(1):267-288.
83. Agier L, Portengen L, Chadeau-Hyam M, et al. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. *Environ Health Perspect.* 2016;124(12):1848-1856.
84. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol.* 2005;67(2):301-320.
85. Keil AP, Buckley JP, O'Brien KM, et al. A quantile-based g-computation approach to addressing the effects of exposure mixtures. *Environ Health Perspect.* 2020;128(4):47004.
86. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J Agric Biol Environ Stat.* 2015;20(1):100-120.
87. Goodrich JA, Walker DI, He J, et al. Metabolic signatures of youth exposure to mixtures of per- and polyfluoroalkyl substances: a multi-cohort study. *Environ Health Perspect.* 2023;131(2):27005.
88. Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization. *J Comput Graph Stat.* 2015;24(3):627-654.
89. Chung MK, Buck Louis GM, Kannan K, Patel CJ. Exposome-wide association study of semen quality: systematic discovery of endocrine disrupting chemical biomarkers in fertility require large sample sizes. *Environ Int.* 2018;125:505-514.
90. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology.* 2008;19(5):640-648.
91. Patel CJ, Ji J, Sundquist J, Ioannidis JPA, Sundquist K. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. *Sci Rep.* 2016;6:31308.
92. Khoury MJ, Ioannidis JPA. *Medicine. Big data meets public health.* Science. 2014;346(6213):1054-1055.
93. Nguyen VK, Colacino J, Chung MK, et al. Characterising the relationships between physiological indicators and all-cause mortality (NHANES): a population-based cohort study. *Lancet Healthy Longev.* 2021;2(10):e651-e662.
94. Galecki A, Burzykowski T. *Linear Mixed-Effects Models Using R: A Step-by-Step Approach.* Springer Science & Business Media; 2013.
95. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics.* 1988;44(4):1049-1060.
96. Chung MK, Caboni M, Strandwitz P, et al. Systematic comparisons between Lyme disease and post-treatment Lyme disease syndrome in the U.S. with administrative claims data. *EBioMedicine.* 2023;90:104524.
97. Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text,* 3rd ed. Springer; 2011.
98. Akhtari FS, Lloyd D, Burkholder A, et al. Questionnaire-based polyexposure assessment outperforms polygenic scores for classification of type 2 diabetes in a Multiancestry Cohort. *Diabetes Care.* 2023;46(5):929-937.
99. He Y, Patel CJ. Software application profile: PXStools—an R package of tools for conducting exposure-wide analysis and deriving polyexposure risk scores. *Int J Epidemiol.* 2022;52(2):633-640. <https://doi.org/10.1093/ije/dyac216>.
100. He Y, Patel CJ. Shared exposure liability of type 2 diabetes and other chronic conditions in the UK Biobank. *Acta Diabetol.* 2022;59(6):851-860. <https://doi.org/10.1007/s00592-2-01864-5>.
101. He Y, Qian DC, Diao JA, et al. Prediction and stratification of longitudinal risk for chronic obstructive pulmonary disease across smoking behaviors. *Nat Comm* 2023;14:8297. <https://www.nature.com/articles/s41467-023-44047-8>.
102. Bobb JF, Valeri L, Claus Henn B, et al. Bayesian Kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics.* 2015;16(3):493-508.

103. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol*. 2008;77(4):802-813.
104. Maitre L, Guimbaud JB, Warembourg C, et al.; Exposome Data Challenge Participant Consortium. State-of-the-art methods for exposure-health studies: results from the exposome data challenge event. *Environ Int*. 2022;168:107422.
105. Taylor KW, Joubert BR, Braun JM, et al. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. *Environ Health Perspect*. 2016;124(12):A227-A229.
106. Chiu YH, Bellavia A, James-Todd T, et al.; EARTH Study Team. Evaluating effects of prenatal exposure to phthalate mixtures on birth weight: a comparison of three statistical approaches. *Environ Int*. 2018;113:231-239.
107. Barrera-Gómez J, Agier L, Portengen L, et al. A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environ Health*. 2017;16(1):74.
108. Lazarevic N, Barnett AG, Sly PD, Knibbs LD. Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: a review of existing approaches and new alternatives. *Environ Health Perspect*. 2019;127(2):26001.
109. Data Management and Sharing Policy. Accessed December 11, 2023. <https://sharing.nih.gov/data-management-and-sharing-policy>.
110. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
111. Jbaily A, Zhou X, Liu J, et al. Air pollution exposure disparities across US population and income groups. *Nature*. 2022;601(7892):228-233.
112. Yazdi MD, Wang Y, Di Q, et al. Long-term effect of exposure to lower concentrations of air pollution on mortality among US Medicare participants and vulnerable subgroups: a doubly-robust approach. *Lancet Planet Health*. 2021;5(10):e689-e697.
113. Zhou X, Josey K, Kamareddine L, et al. Excess of COVID-19 cases and deaths due to fine particulate matter exposure during the 2020 wildfires in the United States. *Sci Adv*. 2021;7(33). <https://www.science.org/doi/10.1126/sciadv.abi8789>.
114. Di Q, Wang Y, Zanobetti A, et al. Air pollution and mortality in the medicare population. *N Engl J Med*. 2017;376(26):2513-2522.
115. Kohane IS, Aronow BJ, Avillach P, et al.; Consortium For Clinical Characterization Of COVID-19 By EHR (4CE). What every reader should know about studies using electronic health record data but may be afraid to ask. *J Med Internet Res*. 2021;23(3):e22219.
116. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;361:k1479.
117. Rassen JA, Bartels DB, Schneeweiss S, Patrick AR, Murk W. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. *Clin Epidemiol*. 2019;11:1-15.
118. Heinemann JA, Kurenbach B, Quist D. Molecular profiling—a tool for addressing emerging gaps in the comparative risk assessment of GMOs. *Environ Int*. 2011;37(7):1285-1293.
119. Yu CT, Farhat Z, Livinski AA, Loftfield E, Zanetti KA. Characteristics of cancer epidemiology studies that employ metabolomics: a scoping review. *Cancer Epidemiol Biomarkers Prev*. 2023;32(9):1130-1145.
120. Viant MR, Ebbels TMD, Beger RD, et al. Use cases, best practice and reporting standards for metabolomics in regulatory toxicology. *Nat Commun*. 2019;10(1):3041.
121. Wang A, Gerona RR, Schwartz JM, et al. A suspect screening method for characterizing multiple chemical exposures among a demographically diverse population of pregnant women in San Francisco. *Environ Health Perspect*. 2018;126(7):077009.
122. Wang A, Abrahamsson DP, Jiang T, et al. Suspect screening, prioritization, and confirmation of environmental chemicals in maternal-newborn pairs from San Francisco. *Environ Sci Technol*. 2021;55(8):5037-5049.
123. Phillips KA, Yau A, Favela KA, et al. Suspect screening analysis of chemicals in consumer products. *Environ Sci Technol*. 2018;52(5):3125-3135.
124. Chung MK, Regazzoni L, McClean M, Herrick R, Rappaport SM. A sandwich ELISA for measuring benzo[a]pyrene-albumin adducts in human plasma. *Anal Biochem*. 2013;435(2):140-149.
125. Chung MK, Riby J, Li H, et al. A sandwich enzyme-linked immunosorbent assay for adducts of polycyclic aromatic hydrocarbons with human serum albumin. *Anal Biochem*. 2010;400(1):123-129.
126. Chung MK, Grigoryan H, Iavarone AT, Rappaport SM. Antibody enrichment and mass spectrometry of albumin-Cys34 adducts. *Chem Res Toxicol*. 2014;27(3):400-407.
127. Rappaport SM, Li H, Grigoryan H, Funk WE, Williams ER. Adductomics: characterizing exposures to reactive electrophiles. *Toxicol Lett*. 2012;213(1):83-90.
128. Rinschen MM, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. *Nat Rev Mol Cell Biol*. 2019;20(6):353-367.
129. Gao P. The Exposome in the Era of One Health. *Environ Sci Technol*. 2021;55(5):2790-2799.
130. Lind PM, Riserus U, Salihovic S, van Bavel B, Lind L. An environmental wide association study (EWAS) approach to the metabolic syndrome. *Environ Int*. 2013;55:1-8.