HHS Public Access

Author manuscript

Curr Environ Health Rep. Author manuscript; available in PMC 2025 March 20.

Published in final edited form as:

Curr Environ Health Rep. ; 12(1): 13. doi:10.1007/s40572-024-00467-2.

Exposome Burden Scores to Summarize Environmental Chemical Mixtures: Creating a Fair and Common Scale for Cross-study Harmonization, Report-back and Precision Environmental Health

Shelley H. Liu[1], **Katherine E. Manz**[2], **Jessie P. Buckley**[3], **Leah Feuerstahler**[4]

[1]Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2]Department of Environmental Health, University of Michigan, Ann Arbor, MI, USA

[3]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA

[4]Department of Psychology, Fordham University, Bronx, NY, USA

Abstract

Purpose of Review—Environmental health researchers are increasingly concerned about characterizing exposure to environmental chemical mixtures (co-exposure to multiple chemicals simultaneously). We discuss approaches for quantifying an overall summary score or index that reflects an individual's total exposure burden to components of the mixture. We focus on unsupervised methods, in which the summary score is not computed in relation to a pre-specified health outcome.

Recent Findings—Sum-scores and principal components analysis (PCA) are common approaches for quantifying a total exposure burden metric but have several limitations: 1) they require imputation when using exposure biomarkers with high frequency of non-detection, 2) they do not account for exposure heterogeneity, 3) sum-scores assume the same measurement error for all people, while there is no error term inherent to the PCA model as its primary purpose is dimension reduction, and 4) in pooled analyses, both approaches are limited to analyzing the set of exposure variables that are in common across all studies, potentially discarding valuable information. Meanwhile, item response theory (IRT) is a novel and promising alternative to calculate an exposure burden score that addresses the above limitations. It allows for the inclusion of exposure analytes with high frequency of non-detects without the need for imputation. It can account for exposure heterogeneity to calculate fair metrics for all people, through assessment

✉Shelley H. Liu, Shelley.liu@mountsinai.org.

Author Contribution S.H.L. and L.F. wrote the main manuscript text, prepared Table 1 and assisted with edits, K.E.M. wrote the future directions, prepared Figs. 1 and 2 and assisted with edits, and J.P.B. wrote the introduction and assisted with edits. All authors reviewed the manuscript.

Competing Interests The authors declare no competing interests.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

of differential item functioning and mixture IRT. IRT also quantifies measurement errors of the exposure burden score that are individual-specific, such that it appropriately assigns a larger standard error to an individual who has missing data on one or more exposure variables. Lastly, IRT enhances cross-study harmonization by enabling the creation of exposure burden calculators to set a common scale across studies, and allows for the inclusion of all exposure variables within a chemical class, even if they were only measured in a subset of participants.

Summary—Summarizing total exposure burden, through the creation of fair and informative index scores, is a promising tool for environmental health research as environmental exposures are increasingly used for biomonitoring and clinical recommendations.

Keywords

Environmental mixtures; Exposome; Precision environmental health; Item response theory; Exposure burden scores; Harmonization

Introduction

Environmental health scientists are increasingly concerned with characterizing exposure and health effects of the exposome, which reflects the reality that individuals are simultaneously exposed to multiple environmental agents at once [1]. The exposome encompasses the cumulative environmental exposures an individual experiences throughout their lifetime [2], capturing many environmental agents including chemicals exposures from sources such as air, water, and food, and non-chemical exposures such as social, psychological, and lifestyle factors [3–5]. Exposure to the chemical exposome occurs in mixtures, meaning that humans may encounter more than one harmful chemical at once [6]. Chemical mixtures that contribute to exposure can include organic chemicals, such as per- and polyfluoroalkyl substances (PFAS), pesticides, personal care product chemicals, plasticizers, and flame retardants, as well as inorganic chemicals, such as heavy metals. Because anthropogenic chemicals pose significant health risks, precision environmental health seeks to integrate this knowledge of the exposome with advances in environmental biochemistry and data science to refine our understanding of how chemical exposures affect an individual's health. Throughout this paper, we use the terms "biomarker" and "analyte" to refer to chemicals measured in human biospecimen. Although our focus is on chemical biomarker data, the methods discussed can also be broadly applied to other types of environmental mixtures data, such as external exposure data from environmental monitors and wristbands.

As statistical methods to address chemical mixtures have blossomed in recent years [7], so has the recognition that environmental "mixtures" research encompasses a broad array of questions that each requires its own tools [8, 9]. One such question relates to quantifying an overall summary exposure score or index that reflects an individual's total burden of exposure to chemicals within a specified mixture. Summary metrics representing multiple facets are frequently used in fields such as psychology (e.g., symptom severity scores) [10] and genetics (e.g., polygenic risk scores) [11, 12] to improve measurement precision and simplify risk prediction. However, scoring approaches to create summary metrics are currently underutilized in exposure mixtures research.

Many existing environmental mixture data science approaches are *supervised*, such that they estimate the independent, joint, interactive, or overall effects of a mixture of chemical biomarkers on a pre-specified health outcome [13–17]. These approaches jointly model the environmental mixture-health outcome relationship; therefore, the findings could differ across related health endpoints, even if they represent the same physiological system. These potentially inconsistent findings may make it difficult to understand how to intervene on modifiable mixtures to reduce health impacts. There are a dearth of studies that focus on quantifying cumulative exposure burden to environmental mixtures, independent of a pre-specified health outcome. This is known as an *unsupervised* approach, because the exposure burden to the mixture is modeled independently of a health outcome, so the exposure burden metric will remain constant across health outcomes.

A summary metric of cumulative exposure burden based on chemical biomarkers can be useful for biomonitoring, risk assessment, and health effects research. A summary measure can be used to identify individuals or groups of individuals with high exposure burden who may be most vulnerable to health effects and should be prioritized for intervention. Researchers may also be interested in comparing burden across race/ethnicity groups, socio-economic strata, and geographical regions. By using a consistent metric of exposure for the environmental mixture, researchers can compare the relative “contribution” of exposure burden across many health outcomes/endpoints. Calculating a single summary measure (or a set of summary metrics) can also be useful to collapse higher dimensional data for mediation analyses, confounding adjustment, and disease risk calculators when mixtures are of interest but not the primary research question. Another benefit of a summary score is the ability to compare a common measure across studies to aid harmonization or meta-analyses to make full use of all environmental chemical mixtures data. In some cases, a comparable summary score can be computed for a chemical class even when studies do not measure the same set of chemical exposure biomarkers within the class, whereas results from supervised mixtures approaches are not readily combined or compared when chemical exposure biomarkers are not common across studies.

In this review, we will discuss two common approaches for quantifying exposure burden based on biomarkers of environmental chemicals (sum-scores and principal components analysis, or PCA) and then discuss item response theory (IRT) as a novel and promising alternative. Our focus is on describing unsupervised methods, such that the summary score is not computed in relation to a pre-specified health outcome. We will describe prior applications of these methods in the environmental health literature and the strengths, assumptions, and limitations of each approach. We will focus on how these methods could be used to set a common exposure burden scale across studies to enable cross-study harmonization, pooled analyses, and meta-analyses. We will additionally address issues around dimensionality and how to determine whether a single summary metric or multiple summary metrics best fit the data. We will explore future directions for how these methods could be used to address pressing questions in the precision environmental health and environmental justice literature, around fairness and equity in computing exposure burden scores to ensure they are a fair and informative metric for all people.

Sum-scores

A common, simple approach for generating a measure of total exposure burden from chemical bioassay data is to calculate a sum of mixtures concentrations (Table 1), possibly incorporating weights that provide additional information [9]. For example, the National Academies of Science, Engineering and Medicine (NASEM) released clinical recommendations [18] for individual health monitoring based on the summed concentration of several PFAS analytes in a person's serum or plasma. Toxic equivalency factors (TEFs) are used by health agencies to calculate a toxicity-weighted total measure for dioxin and dioxin-like compounds, and researchers have proposed using molar sums [19] or androgen disrupting potency-weighted sums [20] for calculating total burden of exposure to phthalates [21, 22].

Summed approaches are easy to compute and have a straightforward interpretation. However, there are limitations to using a simple or weighted summation as a proxy of exposure burden. First, sum scores may not fairly represent the relative importance of each exposure analyte if different concentration ranges are observed for different exposure analytes. Even if these differences are accounted for using pre-specified weights, these weights may not be optimal and typically do not account for uncertainty in their determination. Another limitation of using sum-scores is that researchers do not often empirically test for dimensionality (e.g., whether a single sum-score can adequately represent the data, or whether multiple sum-scores are needed). As such, the sum-score may reduce construct validity (e.g., whether it accurately measures what it is supposed to). In addition, when using sum-scores, researchers implicitly assume that a single measurement model (e.g., a simple summation) adequately reflects exposure burden for all people. Due to heterogeneity in exposure sources and patterns in population subgroups, a simple summation could mask disparities in exposure burden across groups, because a simple summation could be a more valid measurement model for some population subgroups than for others. For example, Liu et al. [25] found that using a simple summation of PFAS analyte concentrations masked disparities in PFAS exposure burden between non-Hispanic Asians and non-Hispanic Whites, while these disparities were uncovered using a customized scoring algorithm of PFAS burden that accounted for heterogeneity in exposure profiles. Finally, using sum-scores can be limiting for pooled analyses because epidemiological studies often measure different sets of exposure analytes within a chemical class. In this setting the sum-score can only include exposure analytes assayed in *all* studies, discarding valuable information.

Principal Components Analysis

Principal components analysis (PCA) is another approach for dimensionality reduction, to represent multivariate data with a smaller number of scores [29]. Researchers who wish to use PCA scores must first decide on the number of principal components to retain [30]. This decision can be made through various criteria, including theoretical considerations, graphical methods such as scree plots, and numerical methods such as parallel analysis. If only one summary measure is needed, the first principal component provides an optimally weighted sum of the observed measurements such that no other weighted sum explains

greater variance. Unrotated PCA scores are commonly used and are uncorrelated with each other. Rotating PCA scores does not change the amount of information represented in the set of PCA scores, but it redistributes weights in a way that can improve interpretation [31]. In this setting, the first rotated component may no longer explain the largest amount of variance. As an example of PCA, Maresca et al. 2015 [32] used PCA for dimensionality reduction of urinary phthalate metabolite concentrations. Of note, PCA and its extensions, including the newly developed principal component pursuit [33], are often also used to identify exposure patterns.

There are several important limitations of the PCA approach. First, PCA analyses are typically based on Pearson correlations which reflect the strength of the *linear* relationship between variables. PCA with Pearson correlations cannot account for variables that are nonlinearly related. However, methods are available to conduct nonlinear PCA that can better account for variables that are nonlinearly related [34]. A potential reason for nonlinearity in environmental mixtures research is that some exposures often have high frequencies of non-detects. Directly using linear PCA in this case may be problematic, although principal component pursuit [33] and zero-inflated PCA methods [35] may be promising alternatives.

Item Response Theory

Latent variable models, including factor analysis and item response theory (IRT), are another option for estimating exposure burden scores. Factor analysis and IRT are structurally similar dimension-reduction techniques and assume that the observed exposures are indicators of the latent, underlying exposure burden. However, factor analysis has been recommended for research questions to examine the structure and construct validity of a scale (e.g. identifying the number of factors), while IRT has been recommended when the goal is to estimate scores for participants (for details, see [36, 37]). As recent advancements in factor analyses for environmental mixtures research [38] have been summarized elsewhere (see [7, 26]), our focus here is on describing IRT.

IRT was originally developed in educational testing for creating scales and scoring of high-stakes assessments (e.g., scoring college entrance exams), but has now been increasingly applied in the biomedical literature, such as for patient reported outcomes [39], cognitive outcomes [40], and allostatic load [41]. Most recently, IRT has been applied to create exposure burden scores to environmental mixtures, such as PFAS [25, 27, 28] and phthalate mixtures [23].

IRT routinely addresses issues that are conceptually similar to the challenges of constructing exposure burden scores. Specifically, IRT estimates exposure burden as a latent variable and the scoring algorithm accounts for both exposure analyte concentrations and exposure patterns to the environmental mixture to derive scores: The overall premise of IRT is that individual items differentially inform the latent variable, through data-driven nonlinear functions specific to that item. In our case, we view an “item” as any specific individual-level measurement, such as individual exposure or exposure analyte or biomarker. Analogous to how educational test items are indicators of a students’ latent (underlying)

cognitive ability, measured exposure analytes can be considered indicators for a latent environmental mixture burden. A simple sum or PCA only focus on measured exposure analytes, which may miss the bigger picture for exposure. Further, IRT uses information about the exposure analyte concentrations, and *exposure patterns* to the environmental mixture, to estimate burden scores. Two individuals may have very similar summed concentrations, but the underlying exposures that yield their individual sum scores could be quite different. Whereas there would be minimal differences in sum scores between these two individuals, their IRT derived exposure burden scores would likely be more different, as the IRT algorithm uses information both about the magnitudes of concentrations of exposure analytes as well as the pattern of exposure concentrations to derive scores. As IRT approaches are mostly developed for categorical data, when applied to environmental data, studies often discretize continuous exposure analyte data prior to IRT analysis, which is also more robust to outliers compared with using continuous data. For example, in developing a 2017–2018 PFAS exposure burden calculator whose scores can be interpreted relative to the general US population, we used survey-weighted decile cutoffs to discretize continuous PFAS biomarker data into up to ten categories per PFAS biomarker prior to IRT analysis [28].

Single Exposure Burden Scores vs. Multiple Exposure Burden Scores

At times, based on conceptual, theoretical or data-driven considerations, a single cumulative metric may be sufficient to represent the exposure burden, while at other times multiple exposure burden scores may be needed. For sum scores, these may be defined based on theoretical considerations; for example, researchers construct separate sums of low and high molecular weight phthalates instead of an overall sum. For PCA, researchers can decide to retain more than one principal component (PC), depending on the variance explained and interpretability. For example, Maresca et al. 2015 [32] retained two PCs in their analysis, finding that PC1 scores represented exposure to DEHP and PC2 scores could be interpreted as exposure to non-DEHP phthalates.

Multi-dimensional IRT can also be used to derive multiple summary scores to represent exposure to correlated exposome sub-classes [42], or a bifactor [43] or hierarchical structure [44]. For example, Chen et al. 2023 [23] compared multiple IRT measurement models to represent phthalate burden, and identified that a correlated factors model with three dimensions (low molecular weight, high molecular weight, and DEHP) fit the data best using established model fit statistics. Statistical methods to determine the number of dimensions to retain is a major topic in exploratory factor analysis (for details, see overview in Auerswald and Moshagen [45]). It is recommended to use multiple methods to find converging evidence for the appropriate number of factors to use, and to fit and compare models using different numbers of factors [45, 46]. For instance, one may wish to use a unidimensional latent variable model to estimate an overall score, even if multiple sub-factors are identified, especially if the sub-factors are correlated with each other [47, 48]. Finally, it is important to also consider whether the models make theoretical sense [49], as burden scores based on an implausible models may not validly represent true latent burdens.

Calculating Exposome Burden Scores in the Presence of Exposure Data with High Frequency of Non-detects

In environmental mixtures research, at times there are components of a chemical biomarker mixture with a high frequency of non-detects, for example in the case of contaminants that are uncommonly used by industry or exposure biomarker for which many study participants have levels below the limit of detection (LOD). When using sum-scores or PCA, researchers can use imputation, such as substituting by LOD/sqrt(2) or more complex approaches [50]. When using IRT, researchers do not need to impute – IRT accommodates mixed item types, meaning that researchers can use different numbers of categories for different exposures/biomarkers (e.g. using deciles for frequently detected exposures, and using binary detect/non-detect for infrequently detected exposures).

Implications for Cross-Study Harmonization and Meta-analyses: Creating a Common Exposure Burden Scale Across Studies

Data harmonization is often needed in consortia and other research to achieve large enough sample sizes to detect small effect sizes in pooled analysis, or to make inferences and comparisons across a heterogenous population. This may involve pooled analyses across studies, in which individual participant data is pooled across multiple cohorts to increase sample size, or meta-analyses. For example, the Environmental influences on Child Health Outcomes (ECHO) Program pools together exposure data previously measured for multiple cohorts to examine health effects of chemical exposures [51].

There are two challenges that occur in data harmonization of exposure data: 1) Different studies or laboratories may not measure the same set of chemical exposure biomarkers within a chemical class of interest, and it is important to not be limited to only analyzing the common set of exposure biomarkers measured across all cohorts; and 2) it is important to create a common exposure burden scale across studies, so that participants scores can be compared across studies and retain the same meaning. For example, because ECHO cohorts measured biomarkers in different laboratories at different times, pooled studies are typically limited to a common set of biomarkers measured across all cohorts. In a pooled analysis that investigated the association of gestational PFAS mixtures and childhood body mass index across eight ECHO cohorts, there were 14 PFAS analytes measured by at least one cohort that participated in the study, but the analyses only focused on the 7 PFAS analytes that had detection frequencies greater than 50% and were assayed in three or more cohorts [52].

When using sum-scores or PCA to quantify a cumulative exposure metric using the combined study data, researchers are generally limited to only using the set of exposure analytes that are common across studies. If researchers wish to include additional exposures only assayed in some studies, they may need to impute exposure data for the studies with missing exposure variables, which can introduce bias. If researchers instead decide to calculate study-specific sum-scores or PCA scores, in a setting where there are different sets of exposures assayed across studies, the participants' scores from one study cannot then be compared to participants' scores from another study, because they will not be on the same

scale. In other words, the same numerical exposure burden score does not retain the same meaning across studies, and thus cannot be used for pooled analyses or meta-analyses.

In contrast to other approaches, IRT can be used to facilitate cross-study harmonization, by creating a common exposure burden scale across studies, such that the scores retain the same meaning across studies even if they didn't measure the exact same set of exposure biomarkers, allowing for the full exposure data collected across studies to be used and not just limited to the common set. The exposures that were collected in common across studies are used as "anchor items" to place exposure burden scores onto a common scale across studies [53]. In this way, it is possible to estimate a common latent variable model for all studies [54]. For example, Liu et al. 2022 [28] developed a PFAS exposure burden score calculator based on 2017–2018 US PFAS biomonitoring reference ranges. Researchers can input PFAS analyte data from their studies to calculate PFAS burden scores on the same scale, which enables future harmonization and meta-analyses. We showed that this common scale can be used even if only a smaller set of PFAS analytes were assayed, and results were robust to associations with health outcomes. Even when the most informative PFAS analytes for the burden score were set to missing, we still found similar associations between the PFAS burden score and the health outcomes of interest. This finding may be in part because IRT takes into account an individual's exposure patterns in addition to concentrations of individual exposures, through data-driven nonlinear functions specific to that exposure, to derive exposure burden scores. Chen et al. 2022 [23] found that for phthalate summary scores, when a portion of participants in the study sample were missing phthalate metabolites, perhaps due to contamination of a batch, researchers would not be able to calculate molar sums on a common scale but would still be able to calculate phthalate burden scores on a common scale. Importantly, the simulation showed that consistent associations with health outcomes were identified even if half of the participants did not have measures of some phthalate metabolites (as might occur if different laboratory panels were used for different studies that did not contain the same analytes).

In Fig. 1, we provide an illustration of data harmonization challenges, in which we seek to do a pooled analysis of three studies, and there are five exposure variables (exposures 'A', 'B', 'C', 'D', 'E') measured in one or more of the studies. While Study 1 measures 'A', 'B', 'D'; Study 2 measures 'A', 'D', 'E'; and Study 3 measures 'A', 'B', 'C', 'E'; we see that only exposure 'A' is measured in common across the three studies. If using sum-scores or PCA, we are limited to only assaying exposure A, the common exposure, or we will need to impute exposure data for each study which introduces bias. Alternatively, study-specific sum-scores or PCAs cannot be used in a pooled analyses of the three studies because they are clearly not on the same scale; if they are used in a pooled analyses it may introduce substantial bias. Meanwhile, if using IRT, we can input data from the three studies together to estimate a common latent variable model. In this way, we can use the full set of exposure data measured across studies, and exposure 'A' can be used as the anchor item to set a common exposure burden scale across the three studies. The exposure burden scores estimated from the common latent variable model will be on the same common scale across the three studies, so that the participants' scores from each study retains the same meaning and can be compared and used in pooled analyses.

Ensuring the Exposure Burden Metric is Fair and Informative for all People to Facilitate Precision Environmental Health and Environmental Justice Research

A primary goal of precision environmental health is to deliver individually-tailored, precision interventions to reduce an individual's environmental exposure for primary disease prevention. In environmental justice, researchers may be interested in comparing exposure burden scores across subpopulations to identify and address exposure disparities. Both research goals necessitate having a fair and informative exposure burden metric for all people. One urgent concern, which has been understudied in environmental mixtures research, is whether it is appropriate to use a single measurement model to represent the cumulative burden metric for all people. The implicit assumption when using sum-scores, PCA, or a single IRT model is that a undimensional measurement model is sufficient for estimating exposure burden for all people. However, there may be a need for personalized exposure burden metrics to account for exposure heterogeneity, such as if different diets and behaviors predispose people to be exposed to different sets of exposures. As there may be systematic differences in exposure sources, we may need customized scoring algorithms for calculating exposure burden scores to ensure they are equitable and informative for all people.

Latent variable models can be used to identify if it is appropriate to use the same measurement model across groups (e.g. demographic subpopulations) to represent underlying exposure burden. These methods are known as *differential item functioning* (DIF) [55] or *measurement invariance* analysis [56] in latent variable modeling literature. If DIF is detected for a particular analyte, this suggests that for two individuals with the same underlying level of exposure burden, a participant from one subgroup has a different probability of having a certain concentration level of that exposure analyte compared to a participant from a different subgroup. DIF is a concern because if it is not accounted for in IRT modeling, the resulting exposure burden scores may be biased or will not be on a common scale that facilitates comparisons or pooling across heterogeneous subgroups. In environmental health, a potential cause of DIF could be exposure source heterogeneity, in which different socio-demographic subgroups have different diets or behaviors that expose them to different sets of exposures within the chemical class of interest. It is important to note that the presence/absence of group differences (i.e., overall differences in the distribution of exposure burden scores between two population subgroups) is different than the presence/absence of DIF. There are many formal tests to detect DIF, and if it is found to exist for certain exposure analyte(s) across groups, exposure burden scores can be derived by using different measurement models across groups, while also setting a common scale across groups through the identification of anchor items that do not function differently across groups, so that the exposure burden scores can be compared across groups while retaining the same meaning. If exposure burden scoring is hypothesized to be different across a known group (e.g. sex), a multiple group IRT approach may be used with anchor item(s) to set a common scale. If researchers hypothesize that an unknown combination of socio-demographic, diet and behavioral characteristics may affect exposure burden scoring, such as in the case of exposure source heterogeneity, a mixture IRT (MixIRT) method

combining IRT and latent class analysis could be used. The MixIRT method calculates customized exposure burden scores by simultaneously identifying latent subpopulations characterized by different exposure burden scoring algorithms and estimating those scoring algorithms for each latent subpopulation, while using anchor item(s) to set a common scale across latent subpopulations. The method predicts customized exposure burden scores that are based on a participants weighted likelihood of belonging to each latent subpopulation [25, 57]. We showed that by creating customized PFAS exposure burden scores using a mixture IRT approach, we were able to identify disparities in PFAS exposure burden across racial/ethnic groups that were masked when using summed concentrations as the summary metric [25].

Measurement Error in the Exposure Burden Metric and Accounting for Error in the Exposure Burden Scores when Examining Mixture-outcome Associations

A participant's cumulative exposure burden score is estimated with some error, because there is uncertainty about how closely an individual's true exposure burden can be estimated by the chosen measurement model (e.g. sum-scores, PCA or IRT). When exposure burden scores are used for biomonitoring or communicating exposure burden to participants, it is important to also relay the standard error of the exposure burden metric. Similarly, when investigating associations of the exposure burden metric with health outcomes, it is also important to conduct sensitivity analyses that incorporate the measurement error of the exposure burden metric to verify the robustness of findings.

Sum-scores allow for the estimation of measurement error of the sum-score, by a transformation of the reliability coefficient [58]. However, the standard errors estimated are not individual-specific; the standard error is the same for everyone in the dataset. While standard errors of the sum scores putatively account for the fact that an individual exposure/item is an incomplete measure of overall burden, because standard errors of sum-scores are the same for everyone in the dataset, this has disadvantages in that it does not reflect that a set of exposures may measure some individuals in the sample more accurately than it measures others. Second, standard errors computed for sum scores are considered to be sample-specific and may not generalize to different populations.

Meanwhile, PCA does not readily allow for the estimation of standard errors in the scores. As PCA is not a latent variable model [59], it does not explicitly account for errors or uncertainty in the scores themselves. In other words, there is no error term inherent to the PCA model as its primary purpose is dimension reduction. In contrast, both sum scores and IRT are based on latent variable models that include an error term from which standard errors may be computed [60].

IRT allows for the estimation of individual-specific standard errors of the exposure burden score, reflecting that the true exposure burden is measured better for certain individuals than for others. Specifically, standard errors of IRT scores account for the fact that an individual exposure analyte/item is an incomplete measure of overall burden. This framework also

appropriately assigns a larger standard error to an individual who has missing data on one or more exposures (e.g., there may have been contamination and certain exposure analytes may have been missing for a subset of the sample). This means that there is more uncertainty the fewer exposure analytes that are measured for an individual—this is automatically accounted for in estimating the individual-specific standard error. Another advantage of IRT-based scoring over other scoring methods is that standard errors can be predicted for new populations, so long as the model is verified to be appropriate for that new population. [61]. Researchers who wish to convey uncertainty in scores may also consider using IRT-based interval estimates [62].

When verifying the robustness of associations between the exposure burden metric and the health outcome, researchers have different options. One option is to simultaneously estimate the measurement model (such as IRT) and the structural model (such as regressions with other outcomes), using structural equation modeling (SEM) [63]. SEM explicitly allows users to investigate the relationships of latent variables with other outcomes. As another option, if using IRT estimates that have been estimated in a separate step (such as through the PFAS burden score calculator), it is possible to appropriately account for this uncertainty through the use of plausible value imputation, as we have shown [23, 27, 64]. The premise is that the point estimates of the latent variable scores (e.g. estimated exposome burden scores) are an imperfect estimate of the true unknown exposome burden, so plausible values provide multiple imputations of the score values and are available in most software packages for fitting IRT models. For example, Chen et al. 2022 [23] used plausible value imputation [65] to verify the association between multi-dimensional phthalate burden scores and insulin resistance in sensitivity analyses. They imputed sets of plausible values of the factor scores using their calibrated IRT model, which accounts for errors in the estimation of the IRT model scores, and estimated associations with the health outcome, and repeated this process many times to plot the effect sizes and p-values.

Future Directions and Challenges: Exposome Burden Scores for Non-targeted Chemical Exposomics Data

Traditionally, the exposome has been measured by analytical chemistry laboratories using targeted chemical analysis, which focuses on a set of predefined chemicals that are suspected to have toxicological or adverse effects [66]. This has led to gaps in exposure assessment as this approach often excludes newly introduced chemical substances that may also pose substantial health risks. Thus, regrettable substitution – the practice of replacing a known toxic chemical with another chemicals that is later determined to be equally or more toxic – has made the exposome more difficult to measure as the scope of chemical exposures becomes more complex and variable [67]. Non-targeted analysis (NTA) is a discovery-based analytical chemistry approach to detect organic chemicals that are not included in targeted chemical analysis [3]. NTA relies on high resolution mass spectrometry (HRMS) and provides unitless peak areas for each feature (mass-to-charge ratio detected by the HRMS paired with a chromatography retention time – if enough data is collected by HRMS, this feature can be matched to a chemical identify or chemical formula). When paired with data

dimension reduction approaches, such as burden scores, NTA can help identify cumulative exposures to new chemicals that arise as a result of regrettable substitution.

However, there are multiple data challenges for applying exposome burden scores to non-targeted data. Within a cohort, non-targeted exposome data may be sparse (due to unusual or uncommon chemical structures only encountered in a small group of people) and exhibit multi-dimensionality that complicates analysis and interpretation. Further, there are significant challenges for applying exposome burden scores for cross-study harmonization when using non-targeted data. Because NTA provides non-standardized unitless peak areas than can vary from instrument to instrument, comparing intensities across laboratories or even study populations is challenging. Further, in biospecimens, it can be challenging to detect low-abundant chemical exposures when signals from metabolites and other small biomolecules are much higher, potentially masking the exposome. There are ongoing standardization and harmonization efforts in the NTA community for detection of environmental chemicals (e.g., Best Practices for Non-Targeted Analysis—BP4NTA —https://nontargetedanalysis.org/). [68, 69] Data dimension reduction approaches for detecting low-abundance, potentially sparse, yet highly important or toxic chemicals in these data sets is critical for improving our understanding of the exposome, so that researchers can better understand and mitigate the health impacts of chemical exposures and regrettable substitution.

Conclusion

Summarizing exposure burden, through the creation of fair and informative index scores, is increasingly an important part of environmental health research as environmental exposures are used for clinical recommendations and biomonitoring. Advanced psychometric methods such as IRT may be a useful way to address pressing questions in cross-study harmonization and fairness in exposomics research. In the research community, exposome burden scores can be used for multiple uses (see Fig. 2), including: 1) for cross-study harmonization, such as creating exposure burden calculators so that exposures can be placed onto a common scale across studies, even if they didn’t measure the same set of exposures within that class; 2) for health effects quantification; that is, to model the impact of cumulative exposure on multiple health outcomes, while keeping the exposure metric fixed so that the relative impact on each health outcome can be compared; as a cumulative environmental index for studying gene-environment interactions on health outcomes, and as a exposure mediator variable, for example to assess if the impact of a predictor on the health outcome is mediated through the environment; 3) for biomonitoring and environmental justice, towards identifying high-risk groups for earlier intervention/monitoring for precision environmental health and primary disease prevention; and 4) for report-back to participants in community engaged research, in which researchers can report-back an index of exposure to research participants, and if the index was calibrated based on nationally representative biomonitoring data, participants can compare their own exposure burden scores relative to those of the US population to understand their own exposure risks. There are myriad opportunities for adapting use, additional methodology development, and dissemination of exposome burden scores in environmental health research.

Acknowledgements

S.H.L. was supported by National Institute for Environmental Health Sciences (NIEHS) R03ES033374 and National Institute of Child Health and Human Development (NICHD) K25HD104918. J.P.B. was supported by NIEHS R01ES030078 and R01ES033252.

K.E.M. was supported by NIEHS K01ES035398. Figures 1 and 2 were created in BioRender: Manz, K. (2024) BioRender.com/i61g627. The authors have no conflict of interest to declare.

References

1. 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):A227–9. 10.1289/EHP547. [PubMed: 27905274]
2. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012;41(1):24–32. 10.1093/ije/dyr236. [PubMed: 22296988]
3. Manz KE, Feerick A, Braun JM, et al. Non-targeted analysis (NTA) and suspect screening analysis (SSA): a review of examining the chemical exposome. J Expo Sci Environ Epidemiol. 2023;33(4):524–36. 10.1038/s41370-023-00574-6. [PubMed: 37380877]
4. Zhang P, Carlsten C, Chaleckis R, et al. Defining the scope of exposome studies and research needs from a multidisciplinary perspective. Environ Sci Technol Lett. 2021;8(10):839–52. 10.1021/acs.estlett.1c00648. [PubMed: 34660833]
5. Wild CP. Complementing the genome with an “Exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev. 2005;14(8):1847–50. 10.1158/1055-9965.Epi-05-0456.
6. Sexton K, Hattis D. Assessing cumulative health risks from exposure to environmental mixtures —three fundamental questions. Environ Health Perspect. 2007;115(5):825–32. 10.1289/ehp.9333. [PubMed: 17520074]
7. Joubert BR, Kioumourtzoglou MA, Chamberlain T, et al. Powering research through innovative methods for mixtures in epidemiology (PRIME) program: novel and expanded statistical methods. Int J Environ Res Public Health. 2022;19(3). 10.3390/ijerph19031378 [PubMed: 36612341]
8. Gibson EA, Goldsmith J, Kioumourtzoglou MA. Complex mixtures, complex analyses: an emphasis on interpretable results. Curr Environ Health Rep. 2019;6(2):53–61. 10.1007/s40572-019-00229-5. [PubMed: 31069725]
9. Hamra GB, Buckley JP. Environmental exposure mixtures: questions and methods to address them. Curr Epidemiol Rep. 2018;5(2):160–5. 10.1007/s40471-018-0145-0. [PubMed: 30643709]
10. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606–13. 10.1046/j.1525-1497.2001.016009606.x. [PubMed: 11556941]
11. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic risk scores for diabetes diagnosis and precision medicine. Endocr Rev. 2019;40(6):1500–20. 10.1210/er.2019-00088. [PubMed: 31322649]
12. Ruchat SM, Vohl MC, Weisnagel SJ, Rankinen T, Bouchard C, Perusse L. Combining genetic markers and clinical risk factors improves the risk assessment of impaired glucose metabolism. Ann Med. 2010;42(3):196–206. 10.3109/07853890903559716. [PubMed: 20384434]
13. 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. 10.1093/biostatistics/kxu058. [PubMed: 25532525]
14. Carrico G, 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):100–20. [PubMed: 30505142]
15. Liu SH, Bobb J, Claus Henn B, et al. Bayesian varying coefficient kernel machine regression to assess neurodevelopmental trajectories associated with exposure to complex mixtures. Stat Med. 2018;37:4680–94. [PubMed: 30277584]

16. Liu SH, Bobb J, Schnaas L, et al. Modeling the health effects of time-varying complex environmental mixtures: Mean field variational Bayes for lagged kernel machine regression. Environmetrics. 2018;29:e2504. [PubMed: 30686915]
17. Liu SH, Bobb JF, Lee KH, et al. Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. Biostatistics. 2017;19(3):325–41. 10.1093/biostatistics/kxx036.
18. National Academies of Sciences E, and Medicine. Guidance on PFAS Exposure, Testing, and Clinical Follow-Up. 2022;(Washington, DC: The National Academies Press.) 10.17226/26156
19. Wolff MS, Engel SM, Berkowitz GS, et al. Prenatal phenol and phthalate exposures and birth outcomes. Environ Health Perspect. 2008;116(8):1092–7. 10.1289/ehp.11007. [PubMed: 18709157]
20. Varshavsky JR, Zota AR, Woodruff TJ. A Novel Method for Calculating Potency-Weighted Cumulative Phthalates Exposure with Implications for Identifying Racial/Ethnic Disparities among U.S. Reproductive-Aged Women in NHANES 2001–2012. Environ Sci Technol. 2016;50(19):10616–24. 10.1021/acs.est.6b00522. [PubMed: 27579903]
21. DeVito M, Bokkers B, van Duursen MBM, et al. The 2022 world health organization reevaluation of human and mammalian toxic equivalency factors for polychlorinated dioxins, dibenzofurans and biphenyls. Regul Toxicol Pharmacol. 2024;146:105525. 10.1016/j.yrtph.2023.105525. [PubMed: 37972849]
22. United States Environmental Protection Agency. Dioxin and Dioxin-Like Compounds Toxic Equivalency Information. 2024. 2024. https://www.epa.gov/toxics-release-inventory-tri-program/dioxin-and-dioxin-compounds-toxic-equivalency-information
23. Chen Y, Feuerstahler L, Martinez-Steele E, Buckley JP, Liu SH. Phthalate mixtures and insulin resistance: an item response theory approach to quantify exposure burden to phthalate mixtures. J Expo Sci Environ Epidemiol. 2023. 10.1038/s41370-023-00535-z.
24. Greenfield BK, Rajan J, McKone TE. A multivariate analysis of CalEnviroScreen: comparing environmental and socioeconomic stressors versus chronic disease. Environmental Health. 2017;16(1):131. 10.1186/s12940-017-0344-z. [PubMed: 29237504]
25. Liu SH, Feuerstahler L, Chen Y, Braun JM, Buckley JP. Toward advancing precision environmental health: developing a customized exposure burden score to pfas mixtures to enable equitable comparisons across population subgroups, using mixture item response theory. Environ Sci Technol. 2023;57(46):18104–15. 10.1021/acs.est.3c00343. [PubMed: 37615359]
26. Liu SH, Chen Y, Kuiper JR, Ho E, Buckley JP, Feuerstahler L. Applying latent variable models to estimate cumulative exposure burden to chemical mixtures and identify latent exposure subgroups: a critical review and future directions. Stat Biosci. 2024. 10.1007/s12561-023-09410-9.
27. Liu SH, Chen Y, Feuerstahler L, et al. The U.S. PFAS exposure burden calculator for 2017–2018: Application to the HOME Study, with comparison of epidemiological findings from NHANES. Neurotoxicol Teratol. 2024;102:107321. 10.1016/j.ntt.2024.107321. [PubMed: 38224844]
28. Liu SH, Kuiper JR, Chen Y, Feuerstahler L, Teresi J, Buckley JP. Developing an exposure burden score for chemical mixtures using item response theory, with applications to PFAS mixtures. Environ Health Perspect. 2022;130(11):117001. 10.1289/EHP10125. [PubMed: 36321842]
29. Agay-Shay K, Martinez D, Valvi D, et al. Exposure to endocrine-disrupting chemicals during pregnancy and weight at 7 years of age: a multi-pollutant approach. Environ Health Perspect. 2015;123(10):1030–7. 10.1289/ehp.1409049. [PubMed: 25956007]
30. Zwick W, Velicer W. Comparison of five rules for determining the number of components to retain. Psychol Bull. 1986;99(3):432.
31. Wickens TD. The geometry of multivariate statistics. Psychology Press; 2014:chap 9: Principal component analysis.
32. Maresca MM, Hoepner LA, Hassoun A, et al. Prenatal exposure to phthalates and childhood body size in an urban cohort. Environ Health Perspect. 2015;124(4):514–20. [PubMed: 26069025]
33. Gibson EA, Zhang J, Yan J, et al. Principal component pursuit for pattern identification in environmental mixtures. Environ Health Perspect. 2022;130(11). 10.1289/EHP10479
34. Linting M, van der Kooij A. Nonlinear principal components analysis with CATPCA: a tutorial. J Pers Assess. 2012;94(1):12–25. [PubMed: 22176263]

35. Zeng Y, Li J, Wei C, Zhao H, Tao W. mbDenoise: microbiome data denoising using zero-inflated probabilistic principal components analysis. Genome Biol. 2022;23(1):1–29. [PubMed: 34980209]
36. Bean GJ, Bowen NK. Item response theory and confirmatory factor analysis: complementary approaches for scale development. J Evid Based Soc Work. 2021;18(6):597–618.
37. Wirth RJ, Edwards MC. Item factor analysis: current approaches and future directions. Psychol Methods. 2007;12(1):58–79. 10.1037/1082-989X.12.1.58. [PubMed: 17402812]
38. Roy A, Lavine I, Herring AH, Dunson DB. Perturbed factor analysis: accounting for group differences in exposure profiles. Ann Appl Stat. 2021;15:1386–404. 10.1214/20-AOAS1435. [PubMed: 36324423]
39. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. Qual Life Res. 2007;16(Suppl 1):43–68. 10.1007/s11136-007-9186-4. [PubMed: 17484039]
40. Liu SH, Chen Y, Bellinger D, et al. Pre-natal and early life lead exposure and childhood inhibitory control: An item response theory approach to improve measurement precision of inhibitory control. Environ Health 2023; 10.1186/s12940-023-01015-5
41. Liu SH, Juster RP, Dams-O'Connor K, Spicer J. Allostatic load scoring using item response theory. Compr Psychoneuroendo-crinology. 2021;5:100025. 10.1016/j.cpnec.2020.100025.
42. Reckase MD. Multidimensional item response theory. Springer Science & Business Media; 2009. 10.1007/978-0-387-89976-3
43. Reise SP. The rediscovery of bifactor measurement models. Multivar Behav Res. 2012;47(5):667–96.
44. Yung YF, Thissen D, McLeod LD. On the relationship between the higher-order factor model and the hierarchical factor model. Psychometrika. 1999;64:113–28.
45. Auerswald M, Moshagen M. How to determine the number of factors to retain in exploratory factor analysis: a comparison of extraction methods under realistic conditions. Psychol Methods. 2019;24(4):468–91. [PubMed: 30667242]
46. Maydeu-Olivares A, Cai L, Hernandez A. Comparing the fit of item response theory and factor analysis models. Struct Equ Model. 2011;18:333–56.
47. Drasgow F, Parson CK. Application of unidimensional item response theory models to multidimensional data. Appl Psychol Meas. 1983;7(2):189–99.
48. Reckase MD. Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests. presented at: Americal Educational Research Association; 1990; Boston.
49. Matsunaga M How to factor analyze your data right: do's don'ts and how-to's. Int J Psychol Res. 2010;3:97–110.
50. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40–9. 10.1002/mpr.329. [PubMed: 21499542]
51. Buckley JP, Barrett ES, Beamer PI, et al. Opportunities for evaluating chemical exposures and child health in the United States: the Environmental influences on Child Health Outcomes (ECHO) Program. J Expo Sci Environ Epidemiol. 2020;30(3):397–419. 10.1038/s41370-020-0211-9. [PubMed: 32066883]
52. Liu Y, Wosu AC, Fleisch AF, et al. Associations of gestational perfluoroalkyl substances exposure with early childhood BMI z-scores and risk of overweight/obesity: results from the ECHO cohorts. Environ Health Perspect. 2023;131(6):67001. 10.1289/EHP11545. [PubMed: 37283528]
53. Chan KS, Gross AL, Pezzin LE, Brandt J, Kasper JD. Harmonizing measures of cognitive performance across international surveys of aging using item response theory. J Aging Health. 2015;27(8):1392–414. 10.1177/0898264315583054. [PubMed: 26526748]
54. Kim S-H, Cohen AS. A comparison of linking and concurrent calibration under item response theory. Appl Psychol Meas. 1998;22(2):131–43.
55. Holland PW, Wainer H. Differential Item Functioning. Routledge; 2012. 10.4324/9780203357811
56. Millsap RE. Statistical Approaches to Measurement Invariance. Routledge; 2012. 10.4324/9780203821961

57. Liu SH, Feuerstahler L, Chen HY, Braun JM, Buckley JP. Towards advancing precision environmental health: Developing a customized exposure burden score to PFAS mixtures to enable equitable comparisons across population subgroups, using mixture item response theory. presented at: International Society for Exposure Science; 2023; Chicago, IL. 10.1021/acs.est.3c00343
58. McNeish D, Wolf MG. Thinking twice about sum scores. Behav Res Methods. 2020;52:2287–305. 10.3758/s13428-020-01398-0. [PubMed: 32323277]
59. Schneeweiss H, Mathes H. Factor analysis and principal components. J Multivar Anal. 1995;55(1):105–24.
60. Bandalos D Measurement theory and applications for the social sciences. Guilford Publications; 2018.
61. Kim E, Yoon M. Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. Struct Equ Model. 2011;18(2):212–28.
62. Liu Y, Yang J. Bootstrap-calibrated interval estimates for latent variable scores in item response theory. Psychometrika. 2018;83:333–54. [PubMed: 28879431]
63. Bedeian A, Day D, Kelloway E. Correcting for measurement error attenuation in structural equation models: some important reminders. Educ Psychol Measur. 1997;57(5):785–99.
64. Khorramdel L, von Davier M, Gonzalez E, Yamamoto K. Plausible values: principles of item response theory and multiple imputations. Large-Scale Cognitive Assessment: Analyzing PIAAC Data,. 2020:27–47.
65. Khorramdel L, von Davier M, Gonzalez E, Yamamoto K. Plausible Values: Principles of Item Response Theory and Multiple Imputations. In: Maehler DB, Rammstedt B, eds. Large-Scale Cognitive Assessment : Analyzing PIAAC Data. Springer International Publishing; 2020:27–47.
66. Lai Y, Koelmel JP, Walker DI, et al. High-resolution mass spectrometry for human exposomics: expanding chemical space coverage. Environ Sci Technol. 2024;58(29):12784–822. 10.1021/acs.est.4c01156. [PubMed: 38984754]
67. Maertens A, Golden E, Hartung T. Avoiding regrettable substitutions: green toxicology for sustainable chemistry. ACS Sustain Chem Eng. 2021;9(23):7749–58. 10.1021/acssuschemeng.0c09435. [PubMed: 36051558]
68. Best Practices for Non-Targeted Analysis (BP4NTA) https://nontargetedanalysis.org
69. Place BJ, Ulrich EM, Challis JK, et al. An introduction to the benchmarking and publications for non-targeted analysis working group. Anal Chem. 2021;93(49):16289–96. [PubMed: 34842413]

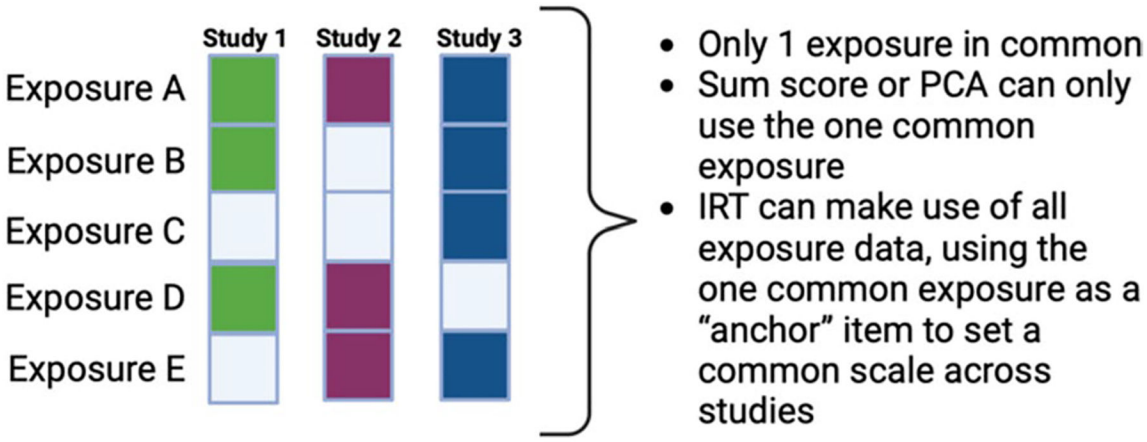

Fig. 1.
Illustration of cross-study harmonization

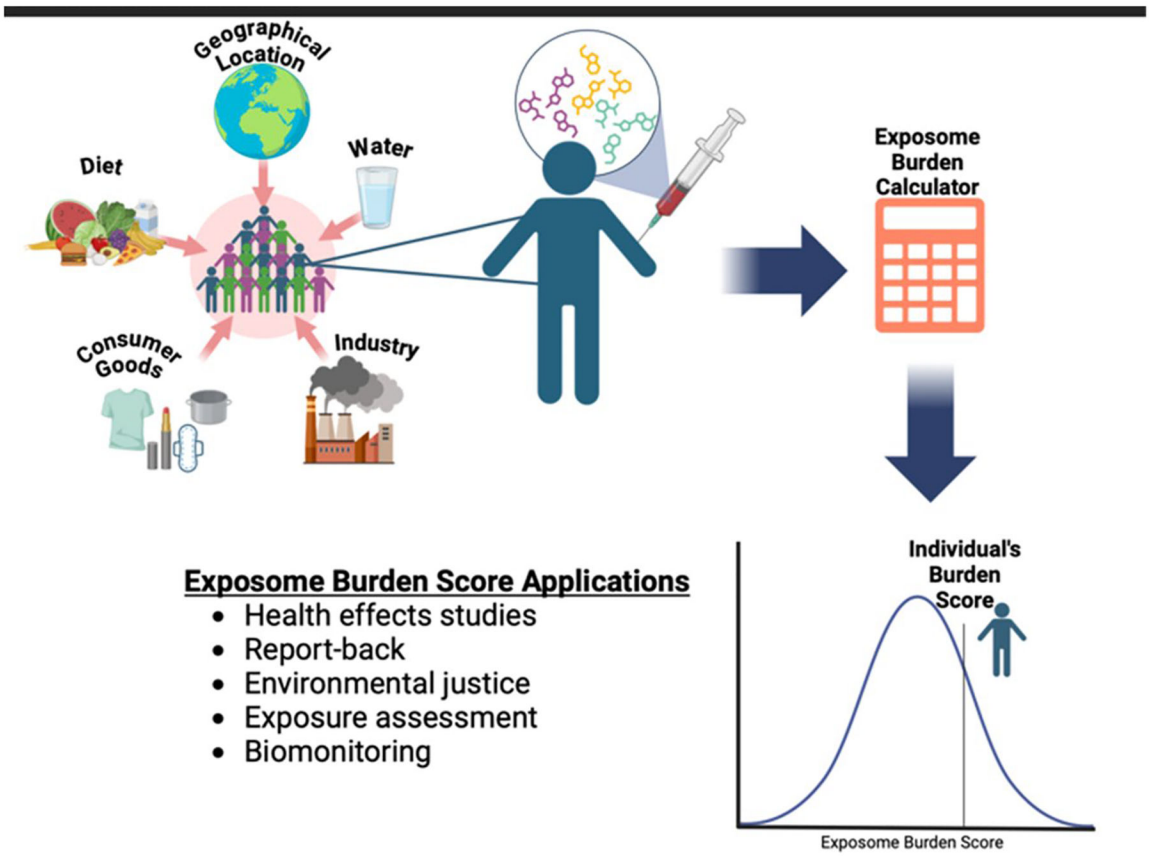

Fig. 2.
Illustration of downstream uses of exposome burden scores

Table 1

Overview of methods to quantify an exposure burden metric

	Sum-scores	Principal components analysis	Item response theory
Data-driven or theoretically-driven	Theoretically-driven	Data-driven	Both theoretically- and data-driven
Missing data for some exposure analytes	Complete-case analysis only or requires imputation	Complete-case analysis only or requires imputation	Allows for missing data using full information maximum likelihood without imputation
Uses continuous or categorical data	Continuous	Continuous	Categorical, exposure variables can have different numbers of categories (e.g. mixed item types)
Allow for multiple indices	Yes; is pre-defined by the researcher	Yes; each principal component is uncorrelated to all other PCs	Yes; can test for number of dimensions; factors can be correlated
Exposure analytes with high frequency of non-detects	Could be included after imputation	Could be included after imputation or using zero-inflated PCA methods	Can use mixed item types (e.g., deciles for frequently detected exposure analytes, binary – detect/nondetect for infrequently detected exposure analytes)
Creating exposure burden calculators to set a common scale across studies, for cross-study harmonization	Yes; but limited to the set of exposure analytes that are in common across studies	Yes; but limited to the set of exposure analytes that are in common across studies	Yes; can use common exposure analytes as anchor items, to make full use of all exposure analytes while setting a common scale
Allows for the creation of a fair burden score for equitable comparisons across population subgroups on a common scale	No; assumes a single measurement model for all people	No; assumes a single measurement model for all people	Yes; through accounting for differential item functioning *(known groups)* and mixture IRT *(unknown groups)*, while using anchor items to set a common scale
Estimating measurement error of exposure burden scores	Not typically done; constant error across all scores	There is no error term inherent to the PCA model	Measurement errors are easily estimated and are individual-specific and automatically adjust for missing data
Examples	Molar sum of phthalate metabolites; [23] Summed PFAS concentrations [18]	PCA for dimension reduction of census tractlevel environmental hazard variables [24]	PFAS exposure burden scores; [25–28] Multi-dimensional phthalate burden scores [23]
Software	No special software/packages needed	Many software/packages, including “stats”, “LearnPCA” in R	“ltm”, “mirt” in R, “MPlus”