Statistical Approaches for Environmental Mixtures and Exposome-Wide Research

Andrea Bellavia, PhD

Department of Medicine, Harvard Medical School
TIMI Study Group, Brigham and Women's Hospital
Departments of Environmental Health, Harvard T.H. Chan School of Public Health
abellavi@hsph.harvard.edu

January 6, 2025

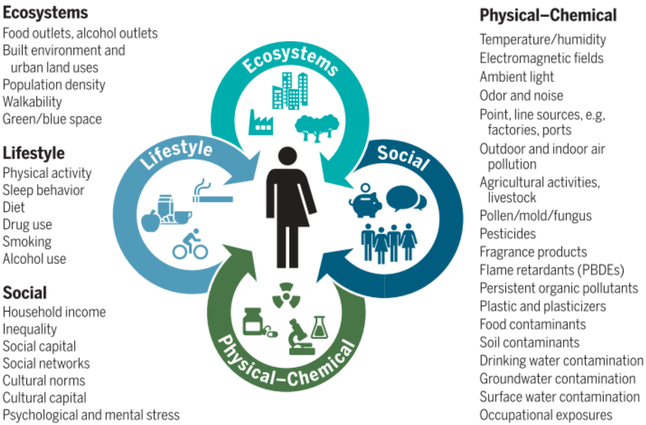
Outline

1. Introduction
2. Classical approaches and their limitations
3. Overview of modern approaches
4. Integrating metabolomics data within the exposome framework

1. Introduction

1.1 The Exposome

External exposome: *"the measure of all the exposures of an individual in a lifetime and how those exposures relate to health"*



Ecosystems
Food outlets, alcohol outlets
Built environment and urban land uses
Population density
Walkability
Green/blue space

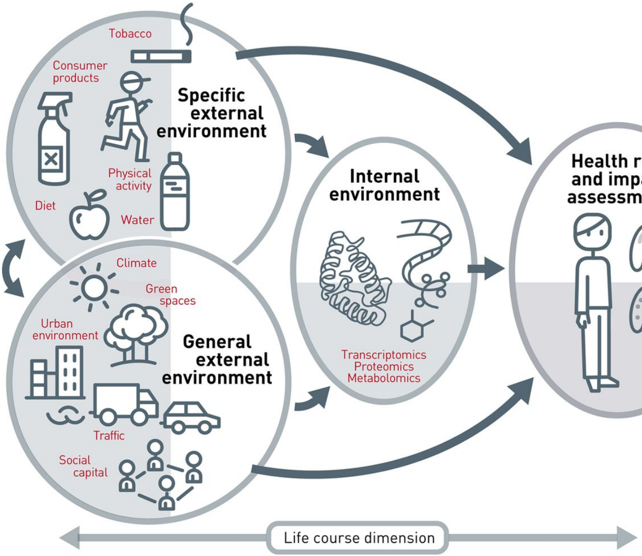
Lifestyle
Physical activity
Sleep behavior
Diet
Drug use
Smoking
Alcohol use

Social
Household income
Inequality
Social capital
Social networks
Cultural norms
Cultural capital
Psychological and mental stress

Physical–Chemical
Temperature/humidity
Electromagnetic fields
Ambient light
Odor and noise
Point, line sources, e.g. factories, ports
Outdoor and indoor air pollution
Agricultural activities, livestock
Pollen/mold/fungus
Pesticides
Fragrance products
Flame retardants (PBDEs)
Persistent organic pollutants
Plastic and plasticizers
Food contaminants
Soil contaminants
Drinking water contamination
Groundwater contamination
Surface water contamination
Occupational exposures

Vermeulen et al., 2020

External and internal exposome: complex mechanisms driven by complex exposures



Environmental Mixtures

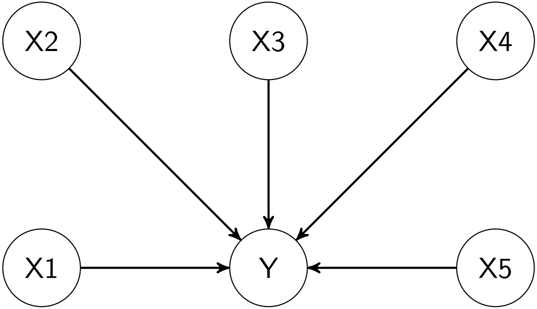
- Within the exposome framework, we can define environmental mixtures as groups of several factors of similar characteristics, often found together (e.g shared sources)

- Framework and analytical tools mostly derived in the context of classical environmental exposures (e.g. chemicals, pollutants, metals..) but can be extended to any kind of exposures characterized by multiple related factors (e.g. nutrients, biomarkers, omics data)

- Example: plastic lunchbox in microwave ->food contaminated by a mixture of several endocrine disrupting chemicals ->hormonal disregulation mechanisms ->higher risk of gestational diabetes

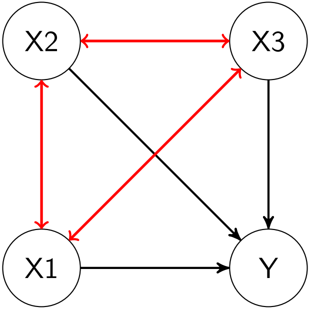
2. Statistical approaches for environmental mixtures and exposome-wide research: Classical approaches and their limitations

- Traditionally, epi studies have focused on one-at-the-time analysis (aka environment-wide association studies (EWAS))
- One regression model is evaluated for each mixture component (eventually adjusting for multiple comparison)



First problem: co-confounding

- The first problem with the application of EWAS in epi research is co-confounding: mixtures components are often associated with each other (e.g. they share a common source such as the plastic lunchbox) and therefore operate as confounders ->e.g: adjusting for X_2 is required to assess the real effect of X_1 on Y and viceversa



Second problem: non-additive effect

- AB pushes the car at 1 mph
- AB's friend pushes at 2 mph
- How fast do they go when they push together?

- 3mph : perfect additivity of effects
- more than 3mph: positive interaction
- less than 3mph: negative interaction

Perfect additivity is seldom met in biology. EWAS are not able to capture these mechanisms.

Straightforward solution: multiple regression

Mutually adjust for all predictors in the same statistical model (takes care of confounding)

$$f(Y) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta \cdot C$$

And eventually include product terms (takes care of non-additivity)

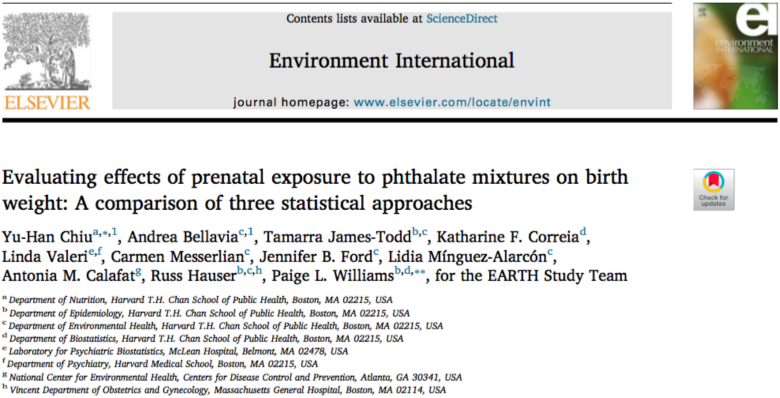
$$f(Y) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_2 \cdot X_3 + \beta \cdot C$$

New problems

Multiple regression allows adjusting for co-confounding and relaxing additivity assumptions. However, there are 2 new challenges:

- Potential collinearity: this may arise when exposures are highly correlated within each other
- Potential overfitting: as the number of covariates increases, the model perfectly describes the data at the expense of poor generalizability. This gets worse as we start relaxing assumptions (of additivity as well as of linearity)

Collinearity example



Contents lists available at ScienceDirect

Environment International

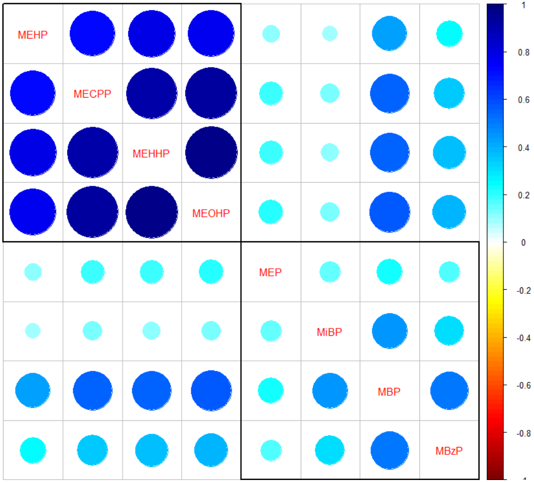
journal homepage: www.elsevier.com/locate/envint

Evaluating effects of prenatal exposure to phthalate mixtures on birth weight: A comparison of three statistical approaches

Yu-Han Chiu[a,*,1], Andrea Bellavia[c,1], Tamarra James-Todd[b,c], Katharine F. Correia[d], Linda Valeri[e,f], Carmen Messerlian[c], Jennifer B. Ford[c], Lidia Mínguez-Alarcón[c], Antonia M. Calafat[g], Russ Hauser[b,c,h], Paige L. Williams[b,d,**], for the EARTH Study Team

[a] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA
[b] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA
[c] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA
[d] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA
[e] Laboratory for Psychiatric Biostatistics, McLean Hospital, Belmont, MA 02478, USA
[f] Department of Psychiatry, Harvard Medical School, Boston, MA 02215, USA
[g] National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA 30341, USA
[h] Vincent Department of Obstetrics and Gynecology, Massachusetts General Hospital, Boston, MA 02114, USA

Correlation plot



Results

	β (one at the time)	β (mutually adjusted)
MiBP	-20.0	-6.8
MBzP	-24.7	-18.7
MEOHP	-23.7	247.1
MnBP	-28.5	-6.5
MEHHP	-28.2	-127.4
MECPP	-32.6	-82.8
MEP	-27.1	25.0
MEHP	-36.8	-59.0

3. Statistical approaches for environmental mixtures and exposome-wide research: Overview of modern approaches

- 2015 NIEHS symposium to study pros and cons of statistical approaches for multiple exposures (focus on $\sim 10 - 20$ covariates)
- 2021 exposome data challenge, extending to settings of ~ 100 covariates, also covering machine learning approaches
- Several additional manuscripts discussing advantages and limitations of statistical approaches (see for example Hamra and Buckley)

Main conclusion:

- We don't have a win-it-all approach
- Addressing the complexity of multiple exposures requires a triangulation of several approaches to appreciate their different advantages and comprehensively assess associations under different perspectives
- There are several research questions of potential interest, with different methods being better suited for specific ones

Potential research questions

1. What are the most common exposure patterns?
2. What are the toxic agents? (sometimes called "bad actors")
3. What is the overall (cumulative) effect of a mixture?
4. Are there interactions (or even synergy) between environmental factors?
5. Are associations linear? What is the shape of the exposure-response curve for a given chemical?
6. What are the causal pathways (i.e. societal, behavioral, and biological mechanisms) through which environmental exposures affect human health?

Broad overview of mixtures/exposome approaches

- Indexing: summarizing the multiple exposures into one (or more) indexes, or summary scores
- Variable selection: Identify mixture components associated with the outcome
- Semi-ML and ML: required for complex settings, allow addressing overfitting when several assumptions must be relaxed (e.g. additivity and linearity)

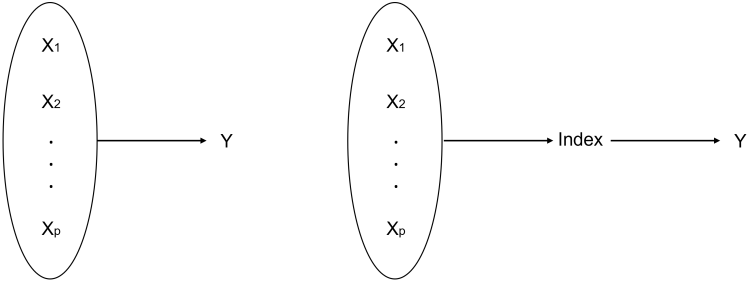
Key pre-analytical considerations

The more complex the data, the more time and effort should be given to pre-analytical and pre-processing phase

- Study design
- Skewedness and variance (often dealing with non-negative covariates)
- Missing values (often not at random)
- Zero values
- Correlation analysis
- Patterns identification. Can also include unsupervised ML such as principal component analysis and clustering approaches

Indexing

- Reduces complexity without variable selection
- Regression models will have no issues of collinearity or overfitting



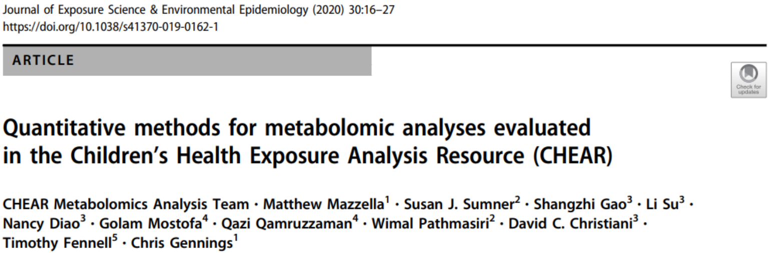
- Classical indexing approaches that do not take the exposure-outcome relationship into account
  - Molar sums of chemical exposures (e.g. DEHP metabolites)
  - Environmental Risk Scores
  - Microbiome indexes (e.g. Shannon)
- Supervised indexing approaches that allow estimating overall effect of the mixture as well as individual contributions (weights) of each mixture component to the index
  - Weighted Quantile Sum (WQS) regression
  - Quantile G-computation

Variable selection

Robust approaches that can conduct variable selection in complex settings (e.g. high correlations)

- Penalized regression (LASSO, elastic net)
- Bayesian approaches
- Partial least squares

Example



Journal of Exposure Science & Environmental Epidemiology (2020) 30:16–27
https://doi.org/10.1038/s41370-019-0162-1

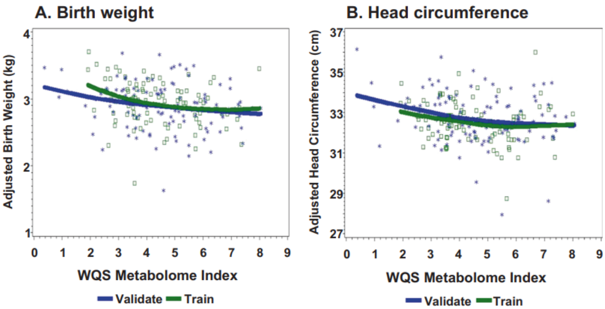
ARTICLE

Quantitative methods for metabolomic analyses evaluated in the Children's Health Exposure Analysis Resource (CHEAR)

CHEAR Metabolomics Analysis Team · Matthew Mazzella[1] · Susan J. Sumner[2] · Shangzhi Gao[3] · Li Su[3] · Nancy Diao[3] · Golam Mostofa[4] · Qazi Qamruzzaman[4] · Wimal Pathmasiri[2] · David C. Christiani[3] · Timothy Fennell[5] · Chris Gennings[1]

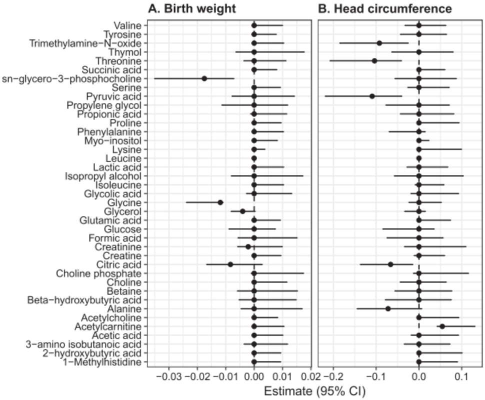
- 199 participants from a Bangladeshi birth cohort
- Associations between cord blood metabolites (H NMR measurements) and birth anthropometric measurements (birth weight and head circumference).
- Total of 39 metabolites in the analysis
- Illustrate research questions that can be addressed by different analytic methods

Example - WQS



"Significant negative associations were observed in validation datasets between WQS indices and birth weight (p = 0.03) and head circumference analyses (p = 0.01). Main contributors (>5%) to the birth weight WQS index include citric acid, formic acid, acetic acid, and leucine (Supplementary Table 3)"

Example - LASSO



”We identified two metabolites negatively associated with covariate-adjusted birth weight, glycine ($\beta = -0.01(-0.03, -0.01)$) and sn-glycero-3-phosphocholine ($\beta - 0.02(-0.04, -0.01)$)” (left panel)

Indexing vs variable selection

- Indexing approaches allow estimating the overall mixture effects, and provide information on bad actors (through ranking)
- With variable selection, original covariates are retained and covariate-specific effects can be estimated
- Both approaches are still making linearity and additivity assumptions[1]

[1] At least with default options

Complex methods for complex settings

- "Complex settings" can be broadly defined as:
  - High-dimensional settings (i.e. n. of covariates close or higher than n/ of individuals
  - Complex mechanisms (e.g. non-linear and non-additive)
- Machine Learning (ML) becomes of interest as it allows relaxing model assumptions and let the data inform on structures and associations
- Addressing overfitting becomes crucial

Hybrid approaches

Hybrid approaches (aka semi-ML) remove some modeling assumptions and incorporate validation tools

- WQS includes training/validation split and bootstrap procedures
- LASSO includes cross-validation to define selection thresholds

Additional approaches include:

- Generalized Additive Models (GAM). Rely on classical regression structure but use splines to study non-linear associations
- Bayesian Kernel Machine Regression (BKMR). Incorporates variable selection through a Bayesian (MCMC) procedure modeling mixture-outcome associations in a non-parametric fashion. Allows providing graphical display of exposure-outcome relationships and qualitative interaction assessment

ML in exposome-wide research

- We could consider full ML for exposome-wide analysis when the data is too complex for even hybrid approaches to work. For example:
  - High number of covariates, sometimes even larger than n
  - High number of interactions and potential interaction levels (not necessarily high p)
  - High number of individuals with complex mechanisms

- Powerfool tools of potential interest include random forests and gradient boosting machines (e.g. GBM, xgboost)

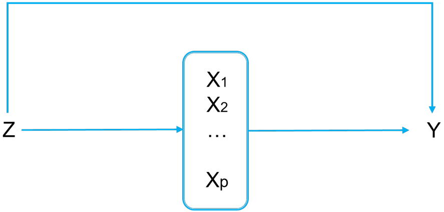
Additional notes for the application of ML in epi studies

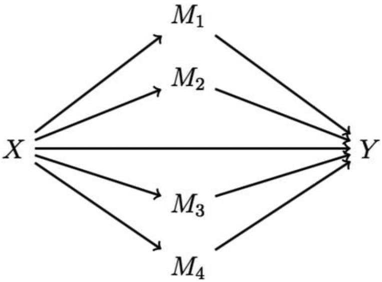
- Several applications on -omics data and recent extensions in exposome-wide research
- Issues of interpretability vs prediction accuracy (Gibson 2019)
- Inferring causal effects from ML methods is not straightforward

4. Integrating metabolomics data within the external/internal exposome framework

- Approaches for environmental mixtures provide considerable advantages for omics data whenever there is a need to overcome the limitations of GWAS/EWAS
- JESEE paper discussed earlier is a great illustrative introduction to the topic in the context of metabolomics data
- In addition, the exposome framework outlines the causal connection between external exposome (e.g. environmental exposures), internal exposome (e.g. omics) and health

- This causal relationship between external exposome, internal, and outcome, can be evaluated with recent extensions of mediation analysis approaches
- Assessing the extent to which a given association (from an epi study - eg. pollution and CVD health) can be explained by specific mechanisms (assessed at the internal exposome level)
- These "mechanisms" are assessed as a complex mixture of multiple metabolites with the challenges previously outlined





- It is possible to integrate previously discussed approaches (e.g. BKMR, LASSO, WQS) within this framework (Bellavia et al 2019)
- Limited applications in the context of omics data

Recap and discussion (1)

- Integrating multiple exposures in epidemiologic and toxicologic studies often requires a step beyond EWAS/GWAS
- Statistical approaches for multivariable complex data (mixtures) are available and have become mainstream in fields such as environmental epidemiology. Applications in other settings (e.g. omics data) are less frequent but growing

Recap and discussion (2)

- The exposome framework connects external exposures and omics data, with the potential of identifying mechanisms of action at individual and population level. Methods to address these complex questions are available
- Complex data require additional effort in the phases of study design, data collection and cleaning
- Complex methods have potential but their application in epi studies might present challenges in terms of results interpretation and translation

Software Resources

R material and additional links:
`https://github.com/andreabellavia/statsmixtures`

References

- Bellavia. "Statistical Methods for Environmental Mixtures: A Primer in Environmental Epidemiology." Springer (2025)
- Bellavia et al. "Approaches for incorporating environmental mixtures as mediators in mediation analysis." Env int 123 (2019): 368-374.
- Chiu et al. "Evaluating effects of prenatal exposure to phthalate mixtures on birth weight: a comparison of three statistical approaches." Env int 113 (2018): 231-239.
- Gibson et al. "Complex mixtures, complex analyses: an emphasis on interpretable results." Current environmental health reports 6 (2019): 53-61.
- Maitre et al. "State-of-the-art methods for exposure-health studies: results from the exposome data challenge event." Env int 168 (2022): 107422.
- Mazzella et al. "Quantitative methods for metabolomic analyses evaluated in the Children's Health Exposure Analysis Resource." JESEE 30.1 (2020): 16-27.
- Taylor et al. "Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop." EHP 124.12 (2016): A227-A229.
- Vermeulen et al. "The exposome and health: Where chemistry meets biology." Science 367.6476 (2020): 392-396.