Evaluating complex interactions in environmental and occupational epidemiology

Andrea Bellavia

Department of Environmental Health
Harvard T.H. Chan School of Public Health
abellavi@hsph.harvard.edu

RSA Seminar Series
December 6, 2019

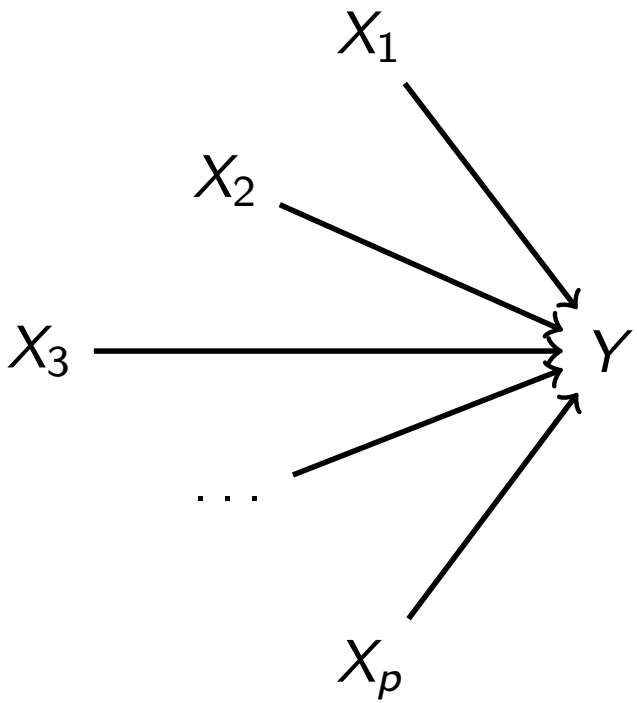
Identification of risk factors

- The primary goal of (most of) population-based studies is to identify possibly modifiable risk factors for a given health outcome.
- Over the last decades, research has primarily focused on evaluating risk factors one at the time (independently)

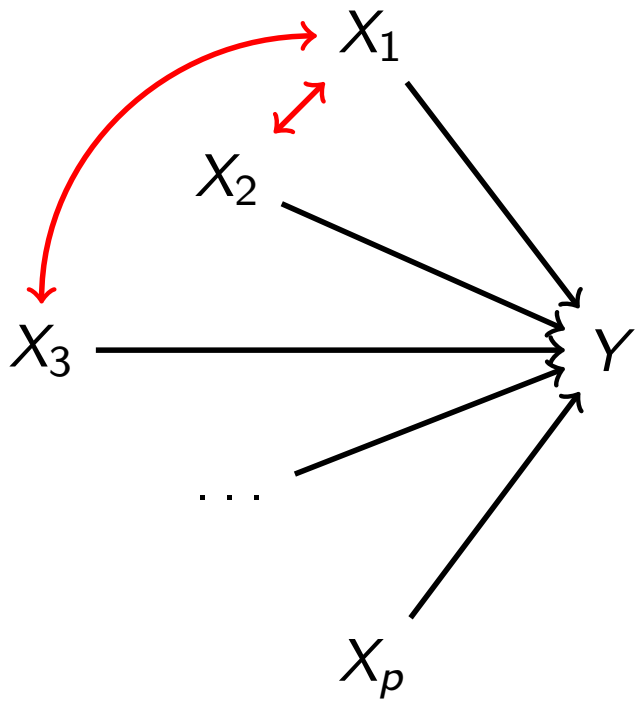
- We know how to study different types of outcomes (continuous, binary, count, survival . . .) and accommodate several additional features

From one to several risk factor

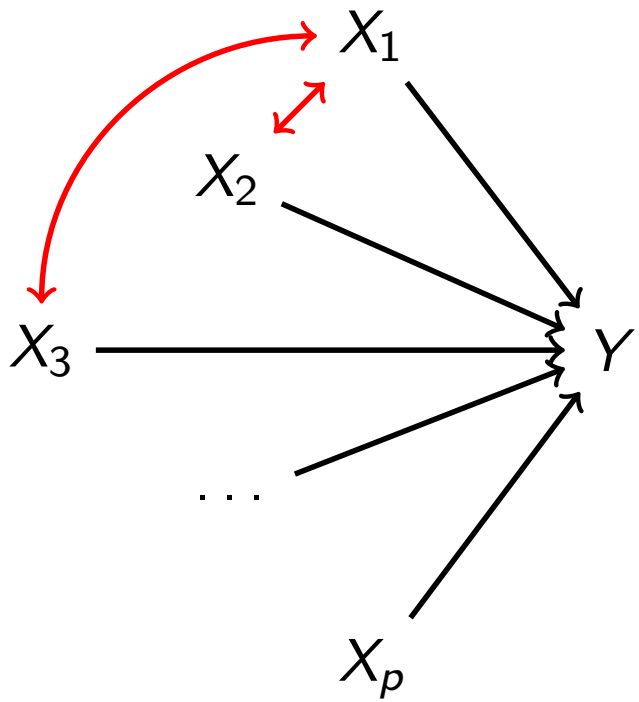
The one-at-the-time approach, however, does not well represent real-life exposures. Individuals, throughout their lifetime, are exposed to a large number of co-occurring exposures, which jointly operate to increase/decrease the risk of a given outcome of interest.



It is reasonable to assume that the different factors of interest are associated within each other.



It is reasonable to assume that the different factors of interest are associated within each other.



This reasoning applies to the majority of lifestyle and behavioral factors (for example coffee/sugar, sleep/PA . . .)

Joint (cumulative) effect

- If that is the case, the 2 factors have be co-adjusted in the same statistical model

$$f(\text{life\_exp}) = \beta_0 + \beta_1 \cdot \text{coffee} + \beta_2 \cdot \text{sugar}$$

- However the two main effects may not be independent (so the joint effect does not correspond to a simple sum of main effects), but the effect of one exposure on the outcome will depend on the level of the other exposure of interest
- This is capture by what we called interaction effect, the additional effect when both predictors are present, in addition to the sum of the 2 main effects

$$f(\text{life\_exp}) = \beta_0 + \beta_1 \cdot \text{coffee} + \beta_2 \cdot \text{sugar} + \beta_3 \cdot \text{coffee} \cdot \text{sugar}$$

Environmental mixtures

In the specific context of environmental epidemiology, the issues we just presented have additional layers of complexity

- We are exposed to hundreds of environmental exposures at any single time point
- For example, Aylward et al (EHP, 2012) reported CDC data suggesting the presence of more than 400 environmental chemicals in human blood and urine

Environmental mixtures

In the specific context of environmental epidemiology, the issues we just presented have additional layers of complexity

- We are exposed to hundreds of environmental exposures at any single time point
- For example, Aylward et al (EHP, 2012) reported CDC data suggesting the presence of more than 400 environmental chemicals in human blood and urine
- Chemicals and pollutants are generally associated within each other (eg they share similar sources), and have to be accounted for as a mixture (defined by the NIEHS as having at least three distinct chemicals or chemical groups)
- To such end, specific methods have been developed and discussed, and the field of environmental epidemiology is gradually adopting a co-exposure assessment as a default

How do we deal with exposure to mixtures?

- Even with a moderate number of exposures, especially if high levels of correlations are present and we want to account for all possible pairs of 2-way interactions, conventional statistical methods such as GLM will be inappropriate.

How do we deal with exposure to mixtures?

- Even with a moderate number of exposures, especially if high levels of correlations are present and we want to account for all possible pairs of 2-way interactions, conventional statistical methods such as GLM will be inappropriate.
- Ideally, we would like to be able to estimate
  - The cumulative/joint effects
  - Detecting interactions
  - Identify the components of the mixture that are associated with the outcome (bad actors)
  - Describe the individual dose-response relationships

How do we deal with exposure to mixtures?

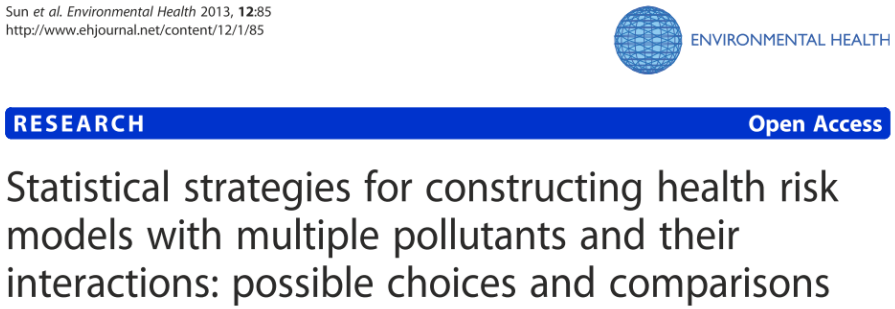
- Even with a moderate number of exposures, especially if high levels of correlations are present and we want to account for all possible pairs of 2-way interactions, conventional statistical methods such as GLM will be inappropriate.
- Ideally, we would like to be able to estimate
  - The cumulative/joint effects
  - Detecting interactions
  - Identify the components of the mixture that are associated with the outcome (bad actors)
  - Describe the individual dose-response relationships
- As we speak, unfortunately, a single approach with all these abilities does not exist.
- Methods generally target 1/2 of these aspects, broadly focusing on either variable selection or classification and prediction

Interaction analysis in environmental epidemiology

- Within the mixture complex synergistic as well as antagonistic interactions may be present, and there are likely high-dimension interactions at play

Interaction analysis in environmental epidemiology

- Within the mixture complex synergistic as well as antagonistic interactions may be present, and there are likely high-dimension interactions at play
- To my knowledge, only 2 studies have discussed available approaches within the context of (2-way) interaction analysis


Sun *et al. Environmental Health* 2013, **12**:85
http://www.ehjournal.net/content/12/1/85

ENVIRONMENTAL HEALTH

RESEARCH **Open Access**

Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons


Barrera-Gómez *et al. Environmental Health* (2017) 16:74
DOI 10.1186/s12940-017-0277-6

Environmental Health

RESEARCH **Open Access**

CrossMark

A systematic comparison of statistical methods to detect interactions in exposome-health associations

Example: chemicals/metals mixture and pregnancy glucose

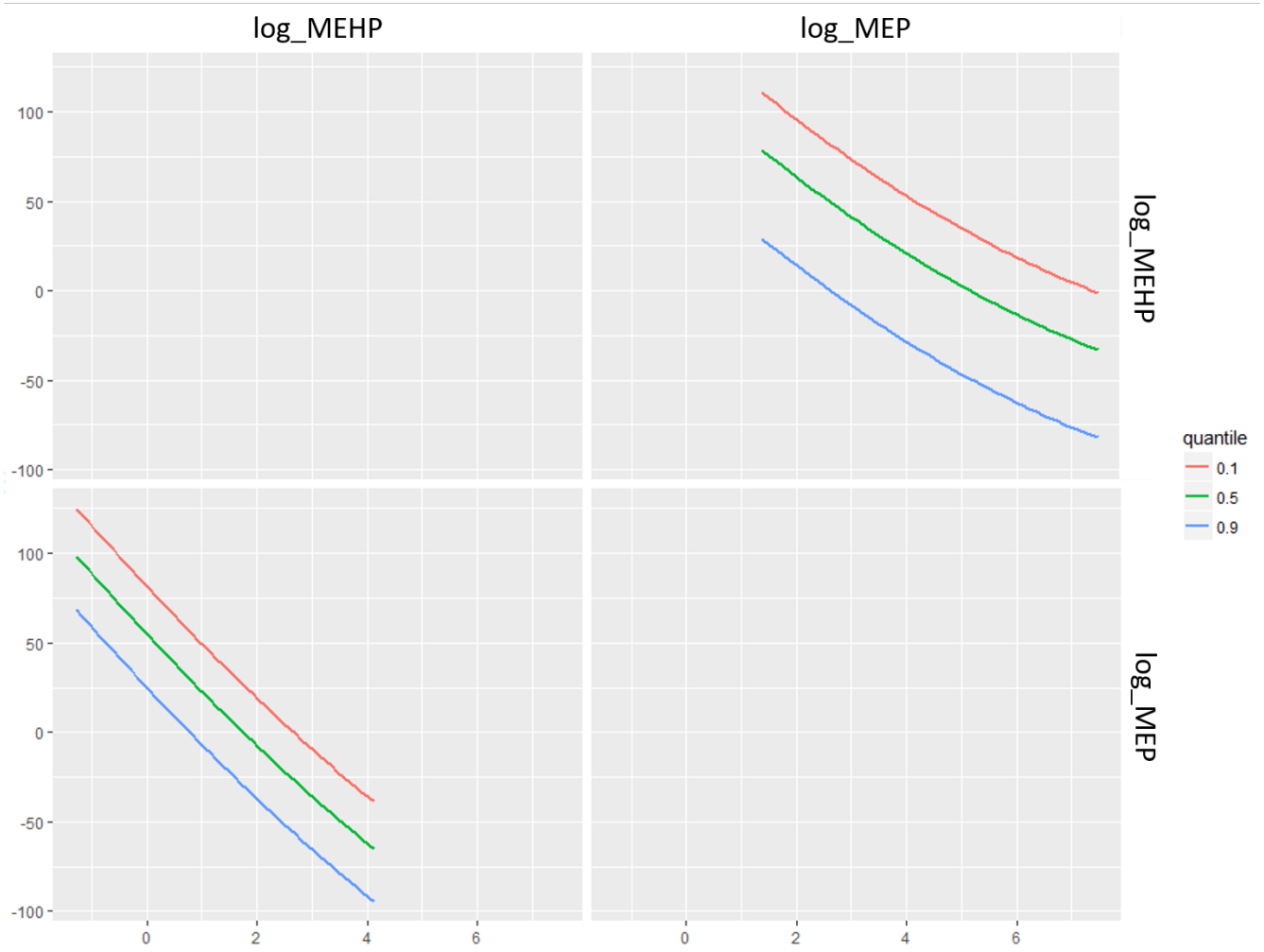
- We are applying some of these techniques to evaluate interactions between endocrine disrupting chemicals (phthalates/phenols) or metals, as they relate to reproductive and perinatal outcomes

Example: chemicals/metals mixture and pregnancy glucose

- We are applying some of these techniques to evaluate interactions between endocrine disrupting chemicals (phthalates/phenols) or metals, as they relate to reproductive and perinatal outcomes
- First, individual dose-responses association using general additive models GAM and Bayesian Kernel Machine Regression BKMR
- Second, Elastic-net to perform variable selection, using an extension GLINTERNET that allows extending the selection to all pairs of 2-way interactions
- Third, graphical assessment of interaction with BKMR

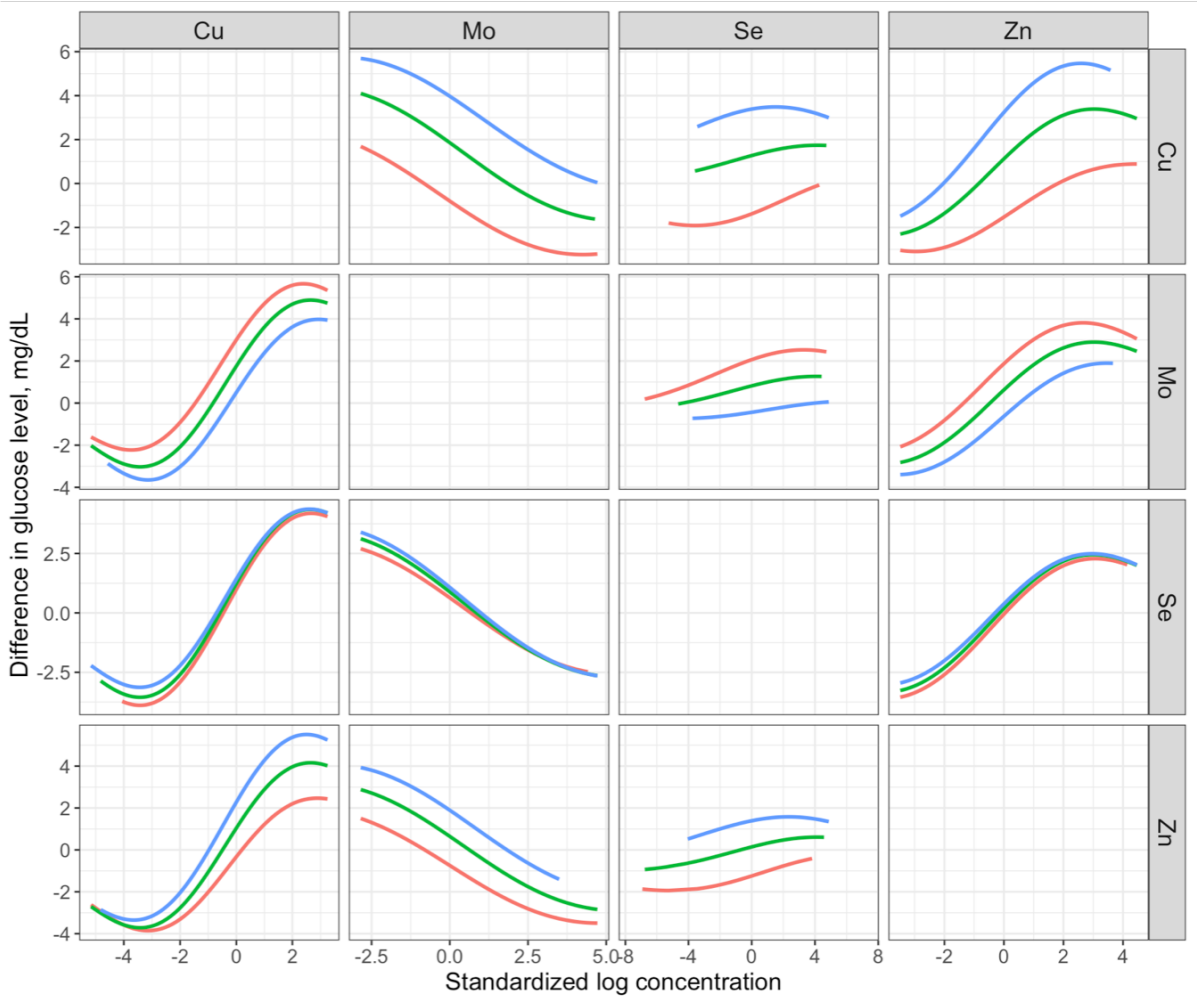
Example 1: phthalates/phenols and birth weight

$\sim$ 500 women from the Environment And Reproductive Health (EARTH) study, with data on $\sim$ 15 chemicals as they relate to birth weight



Example 2: metals and glucose

$\sim$ 2000 women with data on first trimester exposure to $\sim$ 20 metals as they relate to late pregnancy glucose levels



Increasing the number of exposures

- The methods presented in the previous example can be flexibly used with a moderate number of exposures ($\sim$3-100)
- What if we were interested in more complex settings, saying in the identification of risk factors and interactions among a set of $\sim$500-1000 or even more covariates?
- While some of the previous approaches may still provide some insight, the use of machine learning procedures is here required
- While several approaches exist (CARTs, logic regression, random forests, boosted algorithms, neural networks . . .), their use in population based-studies remains very sporadic

Example: Big data screening for identifying risk factors for amyotrophic lateral sclerosis (ALS)

- Different combinations of medications throughout the lifetime, or exposure to specific job occupations, may be associated with increased risk of ALS. As such, analytic procedures that specifically target combinations of exposures are required
- We have access to data from the Danish population and part of the Israeli population (with approximately 4300 ALS cases in total)
- We have been testing some proposed approaches, focusing on job occupations ($\sim$ 700 binary/continuous exposures) and health conditions as they relate to the incidence of ALS

- Logic regression finds combinations of binary covariates that have high predictive power for the response variable

- Logic regression finds combinations of binary covariates that have high predictive power for the response variable

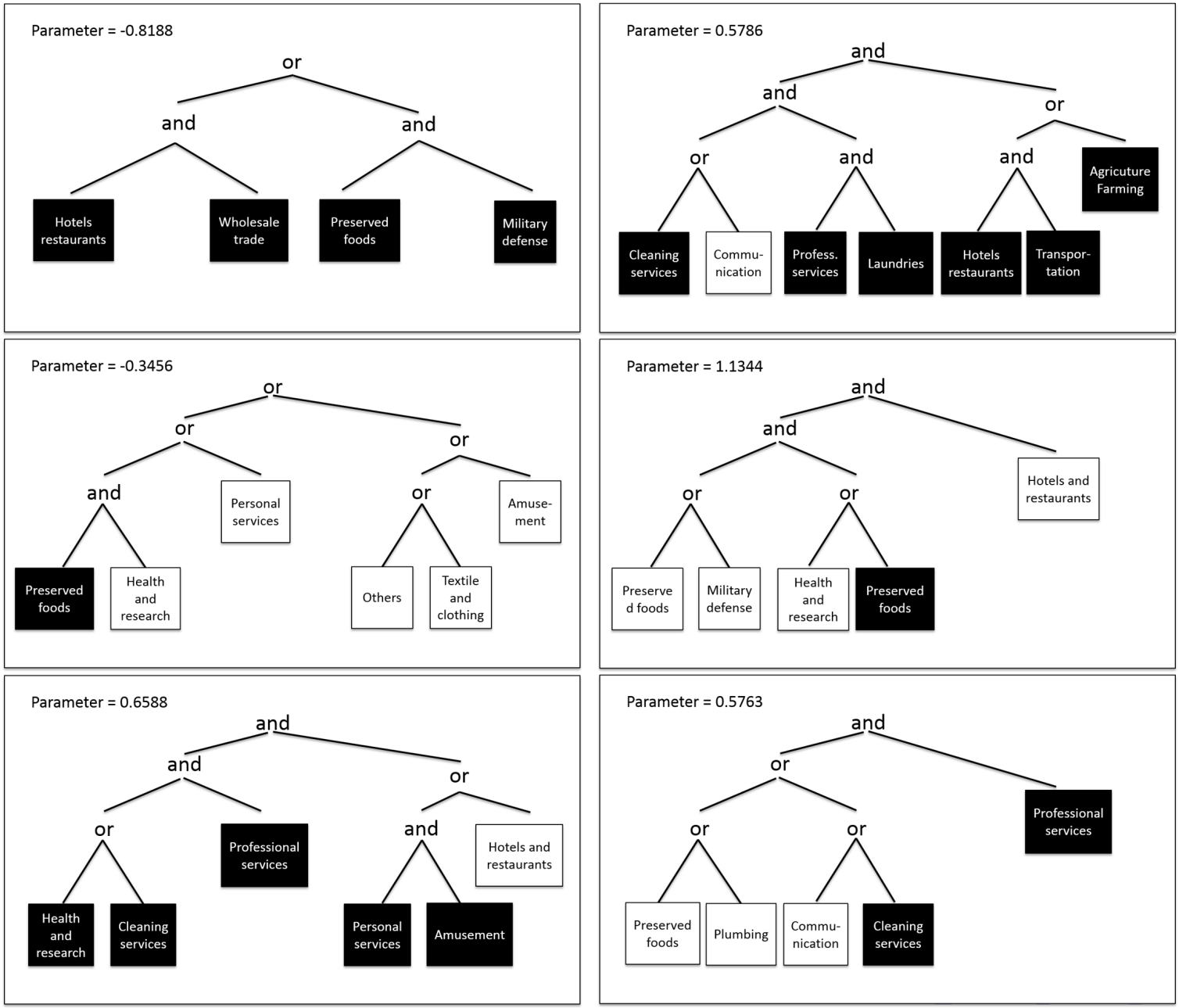
- Random forests were used as an initial screening to identify a selected sample of highly predictive covariates, and Boosted trees were then used to detect the relative importance of each predictor and interactions

- Logic regression finds combinations of binary covariates that have high predictive power for the response variable

- Random forests were used as an initial screening to identify a selected sample of highly predictive covariates, and Boosted trees were then used to detect the relative importance of each predictor and interactions

- We used these approaches as screening procedures and later included selected covariates and interactions in a final logistic regression model

Logic regression

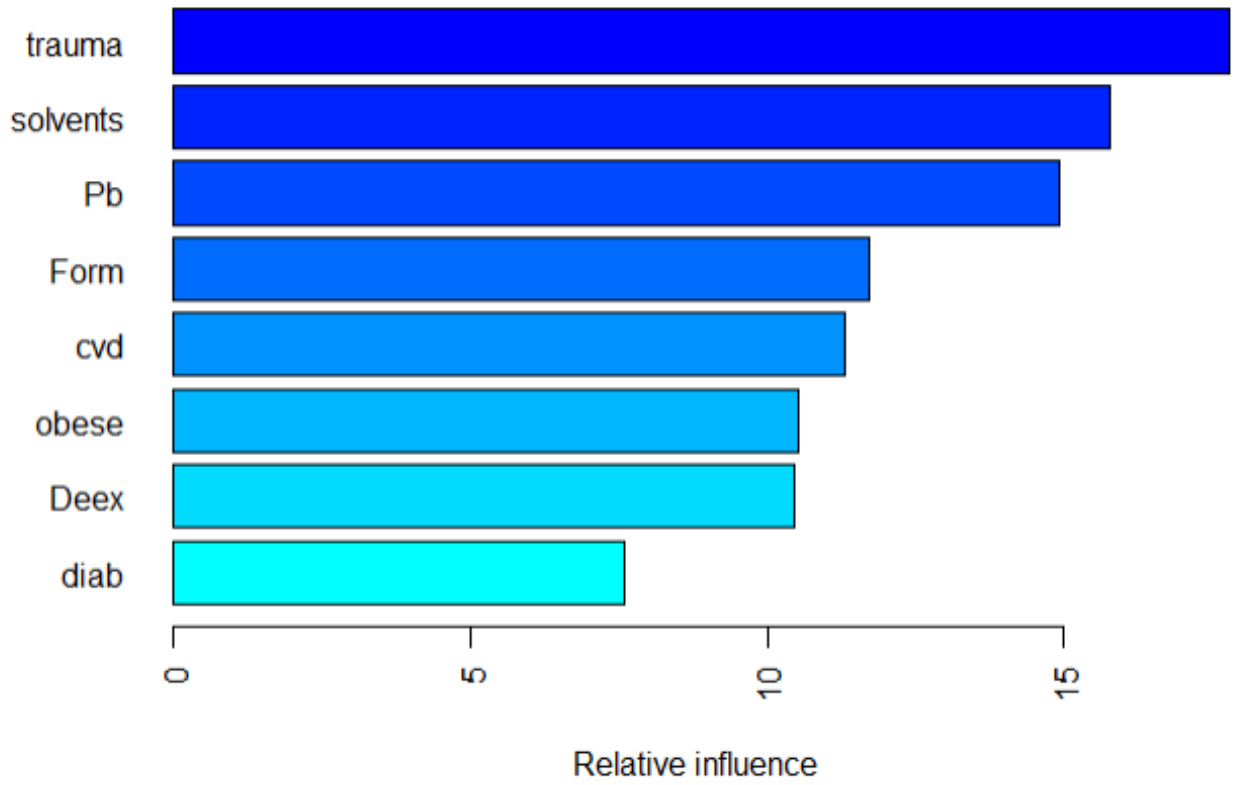


Logic regression - cont.

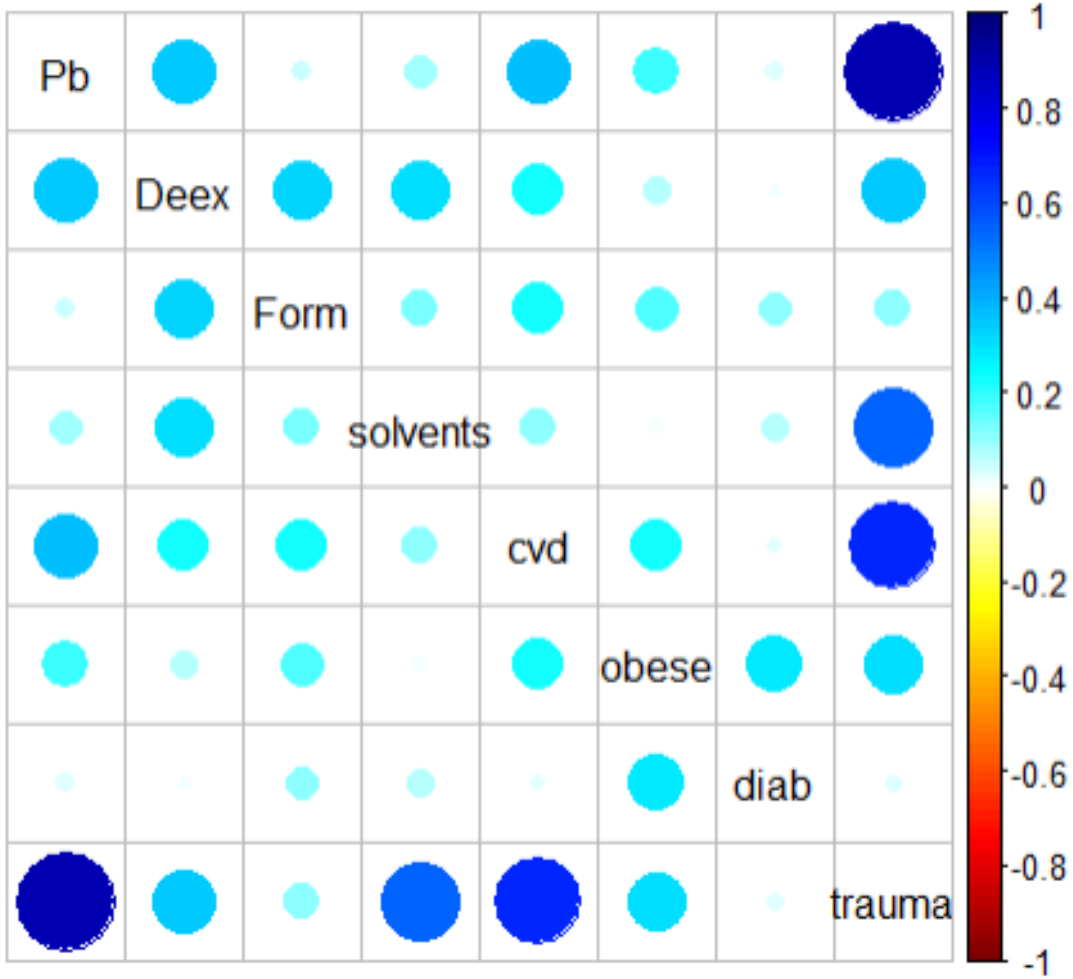
- Of interest, we identified and discussed advantages and (especially) limitations of using this approach with epidemiologic data
- For example, if several exposures have a weak effect, results will be very inconsistent and subject to noise
- Moreover, the interpretation of the boolean combination is anything but straightforward
- At the same time, this technique provided critical information as a screening technique, identifying main predictors and interactions to be included in further statistical models

Random forests and Boosted regression trees

These approaches were used on continuous occupational exposures, estimated from historical job exposure matrices, combined with binary indicators of previous health conditions



(Relative importance of predictors of ALS among male, estimated from gradient boosted models)



(Relative importance of 2-way interactions (H statistic) among male, estimated with gradient boosted models)

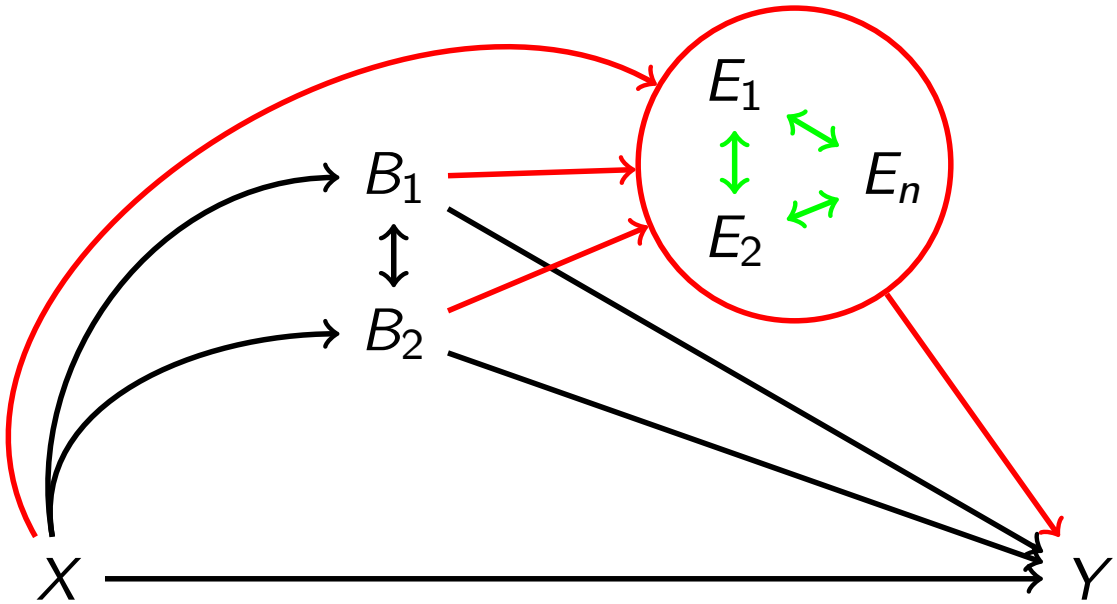
Future directions

1. This ongoing study is one of the first applications of these machine learning techniques for big data screening in population-based studies.

   Other approaches can be investigated, and a formal comparison of techniques is required.

   Furthermore, applications are needed to understand advantages and limitations in real settings

2 Interactions do not only occur within a class of exposures, but throughout the exposome. For example, integrating in a single framework lifestyle and behavioral factors (B), social constructs (X), and environmental exposures (E), may be complex but has the potential to elucidate mechanisms through which diseases are caused.



Bellavia et al. Env. Epi. 2018

3 Exposures, moreover, change over time, either because of public health implementation, lifestyle changes, aging or other reasons. It is important that future studies will be able to integrate this time-varying nature in evaluating individual risk factors as well as high-dimension joint and interactive effects

4 How do we identify causal mechanisms when focusing on high-dimensional data? Integrating methods for co-exposures in mediation analysis is a crucial step in this process

American Journal of Epidemiology

DOI: 10.1093/aje/kwx355

Practice of Epidemiology

Decomposition of the Total Effect in the Presence of Multiple Mediators and Interactions

Andrea Bellavia and Linda Valeri*

* Correspondence to Dr. Linda Valeri, Psychiatric Biostatistics Laboratory, McLean Hospital, Belmont campus – North Belknap, Room 310A, 115 Mill Street, Belmont, MA 02478 (e-mail: lvaleri@mclean.harvard.edu).

Statistics > Methodology

Bayesian kernel machine regression-causal mediation analysis

Katrina L. Devick, Jennifer F. Bobb, Maitreyi Mazumdar, Birgit Claus Henn, David C. Bellinger, David C. Christiani, Robert O. Wright, Paige L. Williams, Brent A. Coull, Linda Valeri

(Submitted on 26 Nov 2018 (v1), last revised 16 Aug 2019 (this version, v2))

Contents lists available at ScienceDirect

Environment International

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

Review article

Approaches for incorporating environmental mixtures as mediators in mediation analysis

Andrea Bellavia[a,*], Tamarra James-Todd[a,b,d], Paige L. Williams[b,c]

Summary

- For a correct representation of reality, population-based studies should focus on evaluating the joint effects of several exposures

Summary

- For a correct representation of reality, population-based studies should focus on evaluating the joint effects of several exposures
- Since co-exposures do not act independently, this implies accounting for high dimensional interactions

Summary

- For a correct representation of reality, population-based studies should focus on evaluating the joint effects of several exposures
- Since co-exposures do not act independently, this implies accounting for high dimensional interactions
- Statistical and machine learning approaches for assessing high-dimension interactions are potentially available

Summary

- For a correct representation of reality, population-based studies should focus on evaluating the joint effects of several exposures
- Since co-exposures do not act independently, this implies accounting for high dimensional interactions
- Statistical and machine learning approaches for assessing high-dimension interactions are potentially available
- All applications and further developments will need to account for the progresses made in the field of causal inference

Acknowledgements

- Department of Environmental Health
  - Tamarra James-Todd
  - Marc Weisskopf
  - Ran Rotem
  - Yinnan Zheng
- Department of Biostatistics
  - Paige Williams
  - Brent Coull

References

- Barrera-Gomez J et al. A systematic comparison of statistical methods to detect interactions in exposome-health associations. Environmental Health. 2017 Dec;16(1):74.
- Bellavia A, Valeri L. Decomposition of the total effect in the presence of multiple mediators and interactions. American journal of epidemiology. 2017 Nov 13;187(6):1311-8.
- Bellavia A et al. Approaches for incorporating environmental mixtures as mediators in mediation analysis. Environment international. 2019 Feb 1;123:368-74.
- Lampa E et al. The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. Environmental health. 2014 Dec;13(1):57.
- Ruczinski I et al. Logic regression. Journal of Computational and graphical Statistics. 2003 Sep 1;12(3):475-511.
- Sun Z et al. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environmental Health. 2013 Dec;12(1):85.
- Taylor KW et al. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. Environmental health perspectives. 2016 Dec;124(12):A227.