Causal Inference for High-Dimensional Exposures

Andrea Bellavia

Brigham and Women's Hospital
Harvard T.H. Chan School of Public Health
abellavia@bwh.harvard.edu

XI Congresso Nazionale SISMEC
September 16, 2021

Background: from association to causation

- A primary goal of epidemiologic studies is to identify possibly modifiable risk factors for a given health outcome.

- We know very well how to investigate associations between exposures of interests and health outcomes $[P(Y = y|X = x)]$

$$X \longrightarrow Y$$

- Nevertheless, our ultimate goal is generally to reduce the incidence/burden of a given disease, and we would like to translate our results into interventions or public health recommendations.

- Rather than an association, we really want to assess the effects of an intervention $[P(Y = y|do(X) = x)]$ [1]

$$do(X) \longrightarrow Y$$

[1] Additional layers of complexity for the definition of causal effects can be further identified

(Often under-looked) principles of causal analysis

Before getting into the aspects of statistical modeling, one should always take these aspects into account:[2]

1. A thorough assessment of causality is not only obtained by using specific statistical methods, but also by accurate study designs
2. Always consult the biological plausibility of your hypotheses and findings
3. Several methodologies and even schools of thoughts exist: do not stick to your favorite one but consider adopting a pluralistic approach
4. Be aware of assumptions that your methods required, and test them

[2] Dominici & Zigler, AJE 2017; Vanderbroucke et al., IJE 2016

Background (2): what do we mean by high-dimensional?

Epidemiologic analysis, and in particular the study of causal effects, has largely focused on evaluating risk factors one at the time. Several additional layers of complexity exist:

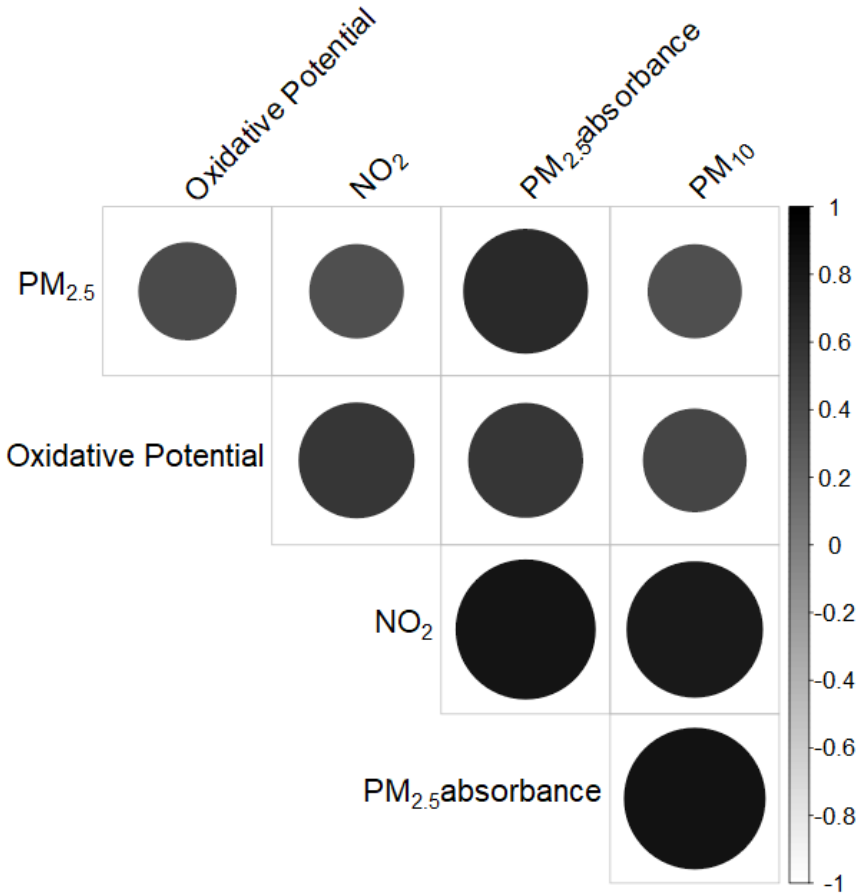
- There are 2 risk factors that jointly operate to determine the risk of a given disease, with a possible (causal) interaction also involved
- The disease or condition is affected by the complex interplay of several factors (exposome research)
- A set of biomarkers can be used to identify subjects at high risk of a given disease or condition (risk score modeling)
- Genetics data are evaluated as potential indicators of disease risk (polygenic risk scores)
- ...

How do we address causal questions in these settings?

Illustrative example: air pollution

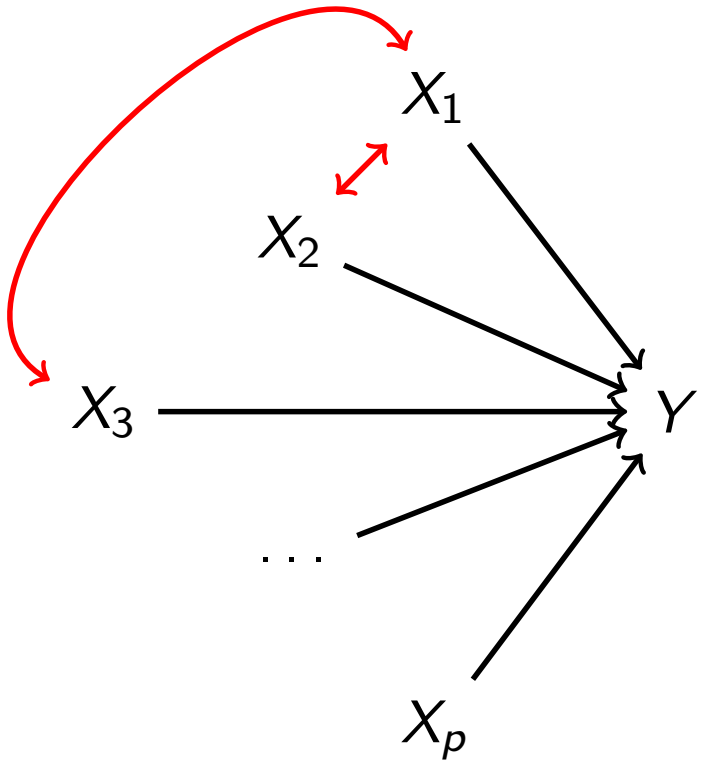
- Several studies have documented associations between air pollution and overall mortality, yet confirming the causal role of pollution is required to implement public health policies and regulations
- Both HEI and the NIEHS listed "Strengthening causal interpretation of evidence from existing cohorts" as one of the primary goals that pollution research should target in 2020-2025
- Methodological and analytical strategies, together with the challenges that this topic brings, have been presented and widely discussed [3]
- One major challenge is the need to incorporate the co-exposure to different pollutants

[3] Carone et al., Epidemiology 2020; Dominici & Zigler, AJE 2017



Correlation plot of pollutants in a population-based study of 90,000 Dutch individuals. Traini et al. Under review

Challenge #1: co-confounding



- The one-at-the-time approach will yield biased estimates (i.e. not reflecting causal effects) because it is subject to confounding bias.
- Multiple regression could be the way forward, but additional action is first needed to address multicollinearity and model complexity

Challenge # 2: interactions

When several covariates are simultaneously evaluated, interactions are likely to play a role. Identification of interactions has critical implications both from a biological as well as from a public health perspective

- A synergistic or antagonistic interaction can inform on biological mechanisms through which the predictors are jointly operating
- Moreover, it allows identification of subgroups in the population for which the effect is observed

- Very few papers have conceptually addressed the issue of high-dimensional (causal) interactions. An empirical yet effective way of addressing this challenge is to conduct a thorough screening of interactions in a preliminary phase of the analysis

- Machine learning techniques based on regression and classification trees, among others, can be very useful in this context, and a recommended approach is to use these methods first and evaluate relevant interactions in sequent regression modeling. [4]

[4] Bellavia et al. IJHEH 2021; Lampa et al., Env. Health 2014

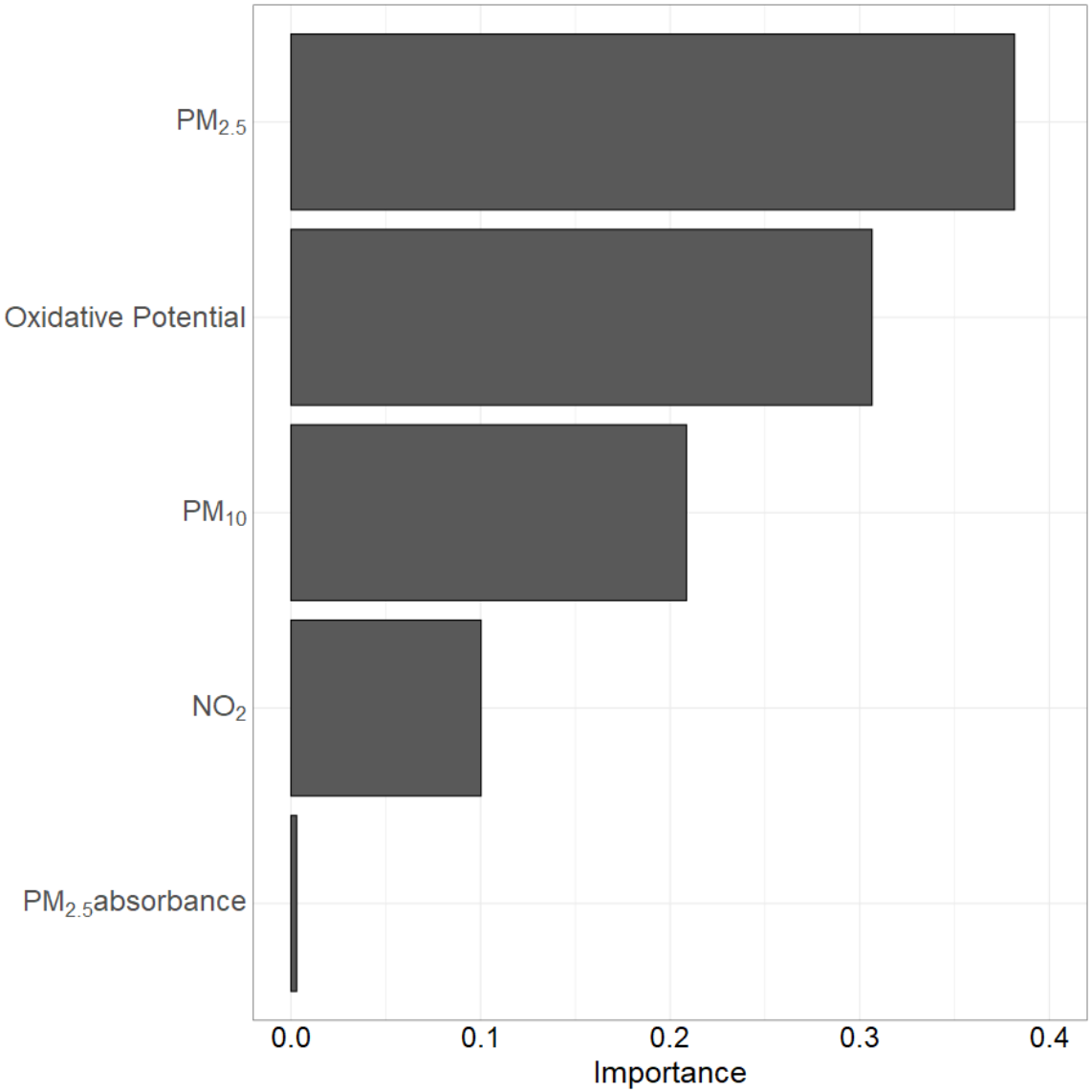
Challenge # 3: what is the role of each covariate?

- A critical step in causal modeling is the a-priori identification of the role played by different factors
- Whether a covariate is a predictor, a confounder, an effect modifier or a mediator, will impact the way analyses are conducted
- In the previous example, we had information on other environmental factors such as greenness temperature and noise, lifestyle factors, anthropometrics, social, and demographic. How does the set of potential (and, often, high-dimensional as well) set of covariates relate to the set of high-dimensional exposures?
- The use of direct acyclic graph (DAGs) in this context becomes extremely important

Illustrative example: preliminary analyses

These are the steps we followed in our air pollution study:

- Pre-processing check of the high-dimensional air pollution exposure (correlation plots and network analysis, missingness evaluation . . .)
- DAG modeling to identify potential confounders of the association between air pollution and mortality
- Unadjusted tree based modeling (boosted regression trees - BRT) to detect variables importance and departures from linearity, and screening for potential interactions
- Application of a specific method developed for multiple pollutants, the weighted quantile sum (WQS), to identify variables importance within a set of correlated environmental exposures. This helped us confirming results from BRT and informing variable selections in following steps



E. Traini et al. Under review

Illustrative example: causal modeling

- Preliminary analyses showed that there were 5 relevant exposures presenting a linear and additive association with the outcome
- We first ran a set of logistic and Cox regression models to select confounders that would be included in later steps
- For this study, we used an extension of propensity score for a set of multiple continuous covariates (mvGPS)
  - When the number of exposures and confounders is not extremely large, this method is a great option to evaluate causal effects of multiple exposure jointly evaluated
  - It requires a multivariate normal distribution for the set of continuous exposures
  - Several tools for balance assessment and bias reduction are available, and a user-friendly R package was developed (`mvGPS`)
- Results confirmed a causal effect of PM2.5 and PM10, with the first one being the most relevant pollutant

Back to the general situation

In a general setting, how do we assess causal effects when multiple exposures are jointly of interest?

1. By spending a lot of time on the pre-analytical phases
   - Is the study design accurate to assess causality?
   - Is there a biological plausibility for each of the variables involved
   - Are assumptions for causal inference met for all possible causes?
2. By drawing a DAG that could capture the complexity of the associations, identify confounders for each association of interest, and distinguish mediators

3. Statistical analyses
   - The more complex the set of exposures of interest, the more time will be required on pre-processing analyses (in ML language, this is referred to as unsupervised analysis)
   - After this, conducting several layers of analyses is recommended (e.g. screening+causal modeling). Tree-based modeling approaches can be very useful in this screening phase
   - Adopt a pluralistic approach. Confirm results with more than one technique and, if possible, choose more than one causal inference technique
   - If possible, go for a regression-based approach. This allows drawing from a set of well-documented familiar techniques without having to reinvent the wheel (e.g. propensity score, IPW, double-robust estimation, marginal structural models, difference in differences, g-methods . . .)

Caveat #1: what if my setting is too complex?

In the air pollution example we got somehow lucky. We had linearities, additivities, and we could use a causal technique based on a simple regression model. Other settings may be way more complex and may require considering other approaches:

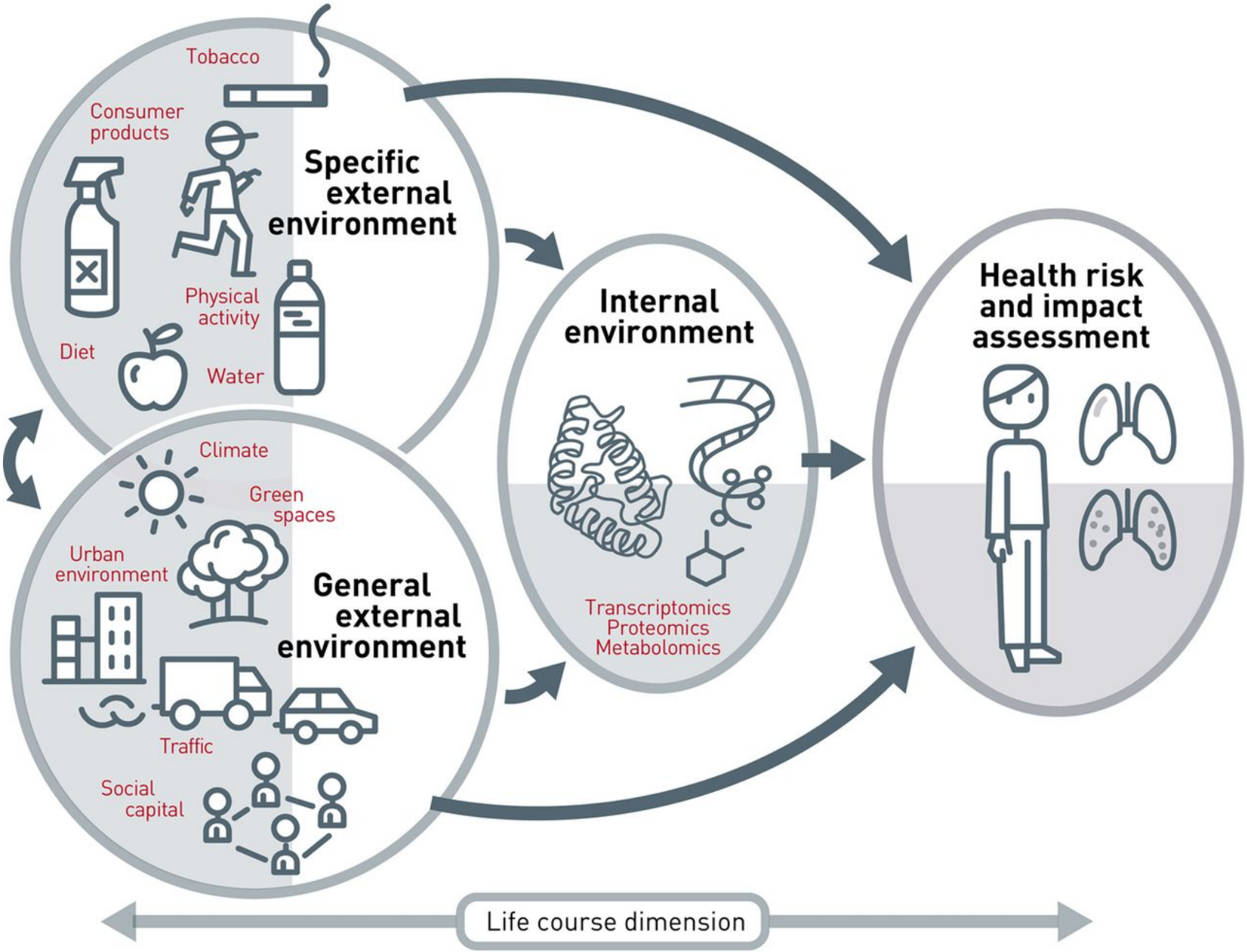
- Adaptive LASSO or other methods for variable selection can be used to reduce the number of covariates within a causal inference framework [5]
- Averaging causal estimators is another valid technique to reduce dimension, especially when hampered by high-dimensional confounders [6]
- Alternative approaches might be available for specific settings

[5] Shortreed et al., Biometrics 2017

[6] Antonelli & Cefalu, Biometrics 2020

Caveat #2: what if I have mediators

Mediation analysis, one of the major areas of research in causal inference, should also take into account the complex high-dimensional nature of exposures and mediators. This, for example, is the challenge in exposome analysis, which aims to evaluate the complex relationships between the general external exposome (E_g), the specific external exposome (E_s), the internal exposome (I), and health outcomes (Y)



Vrijheid et al. Thorax. 2014

- Needless to say, high-dimensional mediation adds some further complexity
- Several papers have addressed the causal definition of direct and indirect effects when multiple mediators and interactions are present [7]
- A comprehensive review on the topic was recently published on EHP [8]
- In general, most settings can be evaluated as long as the high-dimensional mediator is composed by non-sequential components
- If mediators are not independent, estimation will be very complicated

[7] Vanderweele & Vansteelandt, Epi. Methods 2014; Daniel et al., Biometrics 2015; Bellavia & Valeri, AJE 2018

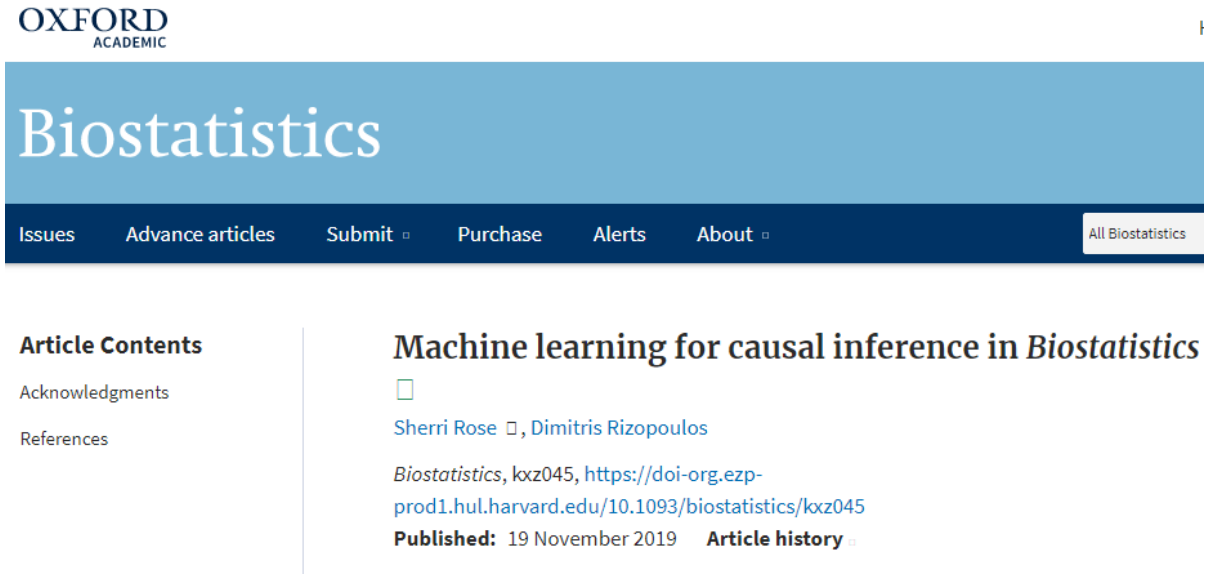
[8] Blum et al., EHP 2020

Summary and final discussion

- Multiple exposures do not act independently of each other and should be jointly evaluated to correctly estimate the causal effects of risk factors on health outcomes
- The more complex the setting, the more important the emphasis on pre-analytical aspects
- To evaluate the causal effects of high dimensional exposures on a given health outcome, adopt a pluralistic approach with several techniques that can confirm your results
- Machine learning approaches can be used to screen the set of exposures and look for interactions, but struggle in distinguishing confounders, exposures, and mediators
- A 2-step (or more) approach, where a set of analyses is used to build a final causal model, could represent a solid approach in most situations

Summary and final discussion (2)

- When using complex methodologies, do not forget that interpretable results are key to provide relevant public health information
- Integrating machine learning algorithms with causal thinking is a very hot topic across different fields, and several resources are available.



Acknowledgments

- Harvard T.H. Chan School of Public Health
  - Ran Rotem
  - Marc Weisskopf
  - Brent Coull
  - Paige Williams

- Utrecht University
  - Eugenio Traini
  - Roel Vermeulen

Additional references

- Bellavia A et al. Approaches for incorporating environmental mixtures as mediators in mediation analysis. Environment international. (2019). Feb 1;123:368-74.
- Bind MA. Causal modeling in environmental health. Annual review of public health. (2019). Apr 1;40:23-43.
- Dominici, F., Peng, R. D., Barr, C. D. & Bell, M. L. Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. Epidemiology (Cambridge, Mass.) 21, 187 (2010).
- HEI Health Effect Institute. Strategic Plan for Understanding the Health Effects of Air Pollution. 2020–2025. Available online: https://www.healtheffects.org/sites/default/files/First-Draft-HEI-StrategicPlan2020-2025.pdf. (2019).
- Renzetti S, Gennings C, Curtin PC. gWQS: an R package for linear and generalized weighted quantile sum (WQS) regression. Journal of Statistical Software. (2019).
- Scholkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y. Toward causal representation learning. Proceedings of the IEEE. 2021 Feb 26;109(5):612-34. https://arxiv.org/abs/2102.11107 (2021).
- Vermeulen, Roel, et al. "The exposome and health: Where chemistry meets biology." Science 367.6476 (2020): 392-396.
- Williams, J. R. & Crespi, C. M. Causal inference for multiple continuous exposures via the multivariate generalized propensity score. arXiv preprint arXiv:2008.13767. (2020).
- Zanobetti, A., Austin, E., Coull, B. A., Schwartz, J. & Koutrakis, P. Health effects of multi-pollutant profiles. Environ Int 71, 13–19 (2014).