Transparent causal inference for observational epidemiology

Lorenzo Fabbri

22/01/2025 (Colicino group meeting)

Who am I?

- Final-year* PhD student at the Barcelona Institute for Global Health (ISGlobal).
  - Google Scholar: https://shorturl.at/IJCU8
  - CV: https://shorturl.at/N3YeP
  - Code: https://github.com/lorenzoFabbri

*Looking for post-doctoral position?

Today’s talk

Today’s talk

- *Prenatal and childhood exposure to mixtures of non-persistent endocrine disrupting chemicals (EDC) and adolescence neurodevelopment: a triangulation study*.

Vrijheid, M., Basagaña, X., Gonzalez, J.R., Jaddoe, V.W., Jensen, G., Keun, H.C., McEachan, R.R., Porcel, J., Siroux, V., Swertz, M.A. and Thomsen, C., 2021. Advancing tools for human early lifecourse exposome research and translation (ATHLETE): Project overview. *Environmental Epidemiology*, *5*(5), p.e166.

Today’s talk

- *Prenatal and childhood exposure to **mixtures** of **non-persistent endocrine disrupting chemicals** (EDC) and adolescence neurodevelopment: a triangulation study*.

Today’s talk

- *Prenatal and childhood exposure to mixtures of non-persistent endocrine disrupting chemicals (EDC) and adolescence **neurodevelopment**: a triangulation study*.

Today’s talk

- ***Prenatal*** *and* ***childhood*** *exposure to mixtures of non-persistent endocrine disrupting chemicals (EDC) and* ***adolescence*** *neurodevelopment: a* ***triangulation*** *study*.

Approaching the question

Some notation

- I will make use of the following notation:
  - A will be our (vector of) exposure(s).
  - W will be our vector of *a priori* selected confounders.
  - C is an indicator variable for censoring ($C = 1$ indicating censored).
  - Y will be our outcome.
  - The potential outcomes will be indicated with $Y^{\{a\}}$.

Approaching the question

1. Start with a well-defined causal **question**, and define the causal **estimand** and **model** (e.g., DAG):

$$\Psi^* = E\left[Y^{\{a=a^*,c=0\}}\right] - E\left[Y^{\{a=a^{nc},c=0\}}\right]$$

Dang, L.E., Gruber, S., Lee, H., Dahabreh, I.J., Stuart, E.A., Williamson, B.D., Wyss, R., Díaz, I., Ghosh, D., Kıcıman, E. and Alemayehu, D., 2023. A causal roadmap for generating high-quality real-world evidence. *Journal of Clinical and Translational Science*, *7*(1), p.e212.

Intermezzo: causal policies / shift functions

- Binary exposures (presence/absence, treated/untreated): $Y^{\{a=1\}} - Y^{\{a=0\}}$.
- Categorical exposures: $Y^{\{a=a_5\}} - Y^{\{a=a_0\}}$.
- Continuous exposures: ?
  - $Y^{\{a=avg(a)\}} - Y^{\{a=nc\}}$.
  - $Y^{\{a=75^{th}\}} - Y^{\{a=25^{th}\}}$.
  - MTP: if $a \geq 0.5$ and $bw < 70$ then $a^g = a * 0.1$, else $a^g = a$.
  - …
- A *zoo* of interventions: static, dynamic, MTPs, deterministic, stochastic…

Hoffman, K.L., Salazar-Barreto, D., Williams, N.T., Rudolph, K.E. and Díaz, I., 2024. Studying Continuous, Time-varying, and/or Complex Exposures Using Longitudinal Modified Treatment Policies. *Epidemiology*, *35*(5), pp.667-675.

Approaching the question

1. Start with a well-defined causal **question**, and define the causal **estimand** and **model** (e.g., DAG).
2. Consider the **observed** data and **identifiability** conditions. Define the statistical estimand:

$$\Psi = E_W[E[Y|W, A = a^*, C = 0] - E[Y|W, A = a^{nc}, C = 0]]$$

Approaching the question

3. Choose the *best* statistical **estimator**:

$$Y \sim \underbrace{A_1 + A_2 + \ldots + A_p}_{\text{Components of mixture } i} + W$$

How to get the *best* estimator?

How to get the *best* estimator

- To summarize:
  1. Define your causal estimand.
  2. *Derive* the associated statistical estimand.
  3. Estimate the so-called *nuisance functions*.
  4. Estimate your effect of interest.

How to get the *best* estimator: g-formula

Remember: $E[Y|W, A = a^*, C = 0]$

$$\Psi^g = \int_A \int_W E\,[Y|A = a, W = w] f^g(A|W)\, f(W)\, dadw$$

Díaz, I., Williams, N., Hoffman, K.L. and Schenck, E.J., 2023. Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, *118*(542), pp.846-857.

How to get the *best* estimator: g-formula

$$\Psi^g = \int_A \int_W E\left[Y|A = a, W = w\right] f^g(A|W)\, f(W)\, da dw$$

Díaz, I., Williams, N., Hoffman, K.L. and Schenck, E.J., 2023. Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, *118*(542), pp.846-857.

How to get the *best* estimator: EIF

$$\Psi^g = \int_A \int_W E\left[Y|A=a, W=w\right] f^g(A|W)\, f(W)\, dadw$$

Kennedy, E.H., 2024. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pp.207-236.

How to get the *best* estimator: EIF

$$\Psi^g = \int_A \int_W E\,[Y|A=a, W=w] f^g(A|W)\, f(W)\, dadw$$

$$IF_1 = \int_A \int_W \frac{I(A=a, W=w)}{\big(f(A=a, W=w)\big)} \{Y - E[Y|A,W]\} f^g(A|W) f(W) dadw$$
$$= \frac{f^g(A|W)}{f(A|W)} \{Y - E[Y|A,W]\}.$$

Kennedy, E.H., 2024. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pp.207-236.

How to get the *best* estimator: EIF

$$\Psi^g = \int_A \int_W E\,[Y|A = a, W = w] f^g(A|W)\; f(W)\; dadw$$

$$IF_2 = \int_A \int_W E[Y|A,W] f^g(A|W)\{I(W = w) - f(W = w)\} dadw$$

$$= \int_A E[Y|A,W] f^g(A|W) da \; - \Psi$$

$$= \int_A E[Y|A^*,W) f(A|W) f^+(A|W) da \; - \Psi$$

Kennedy, E.H., 2024. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pp.207-236.

How to get the *best* estimator: EIF

$$\Psi^g = \int_A \int_W E\,[Y|A = a, W = w] f^g(A|W)\, f(W)\, dadw$$

$$EIF = \frac{f^g(A|W)}{f(A|W)}\{Y - E[Y|A, W]\} + E[Y|A^g, W] - \Psi$$

- One possible way to use the EIF to estimate the parameter of interest is by **plugging in** estimates of the individual components and averaging over the sample (AIPW).

Kennedy, E.H., 2024. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pp.207-236.

How to get the *best* estimator: nuisance functions

- $\Psi^{g}_{AIPW} = avg\left(\widehat{EIF}\right)$
$= \frac{1}{n}\sum_{i} \frac{f^{g}(A_i|W_i)}{f(A_i|W_i)}\{Y_i - E[Y_i|A_i, W_i]\} + E[Y_i|A_i^{g}, W_i]$
- We combine 2 nuisance functions (double-robustness): the propensity score and the outcome model.
- For my project, probably using lmtp R package: https://github.com/nt-williams/lmtp. Works also with multiple exposures. Implements parametric g-computation, IPW, TML, and SDR estimators. Static, dynamic, and MT policies.

Why that?

- Literature on mixtures (environmental epidemiology): quantile g-computation, BKMR, BWQS…
  - Parametric models.
  - No causal “background”.
  - Discretization of exposures that are continuous in nature.

Triangulation and open questions

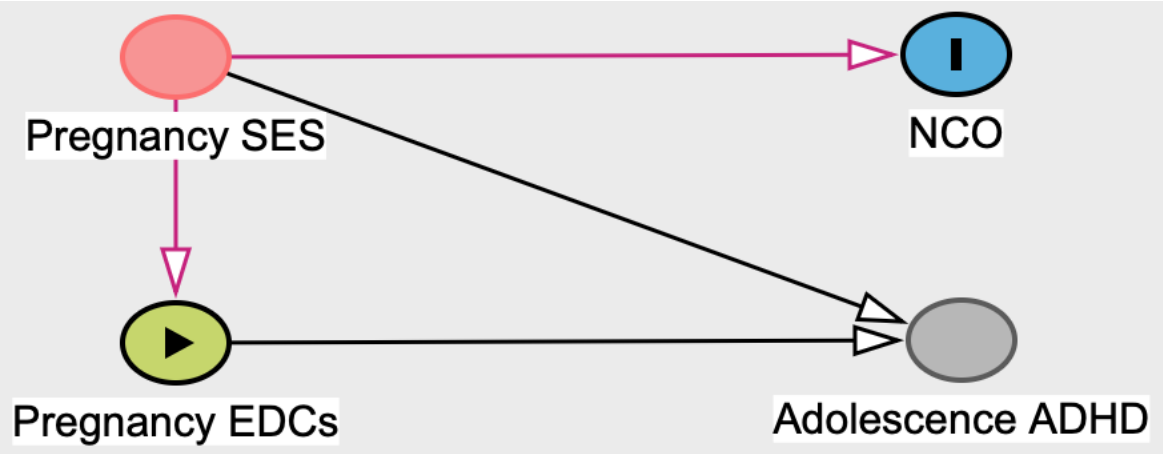
Triangulation and open questions

- Observational studies are subject to **bias** (many).
- We can strengthen our conclusions by combining different data sources and methods:
  - Negative controls (exposure and/or outcome).
  - Cross-cohort comparisons.
  - Genetic variants (MR), although different estimand.
  - …

Lawlor, D.A., Tilling, K. and Davey Smith, G., 2016. Triangulation in aetiological epidemiology. *International journal of epidemiology*, *45*(6), pp.1866-1886.

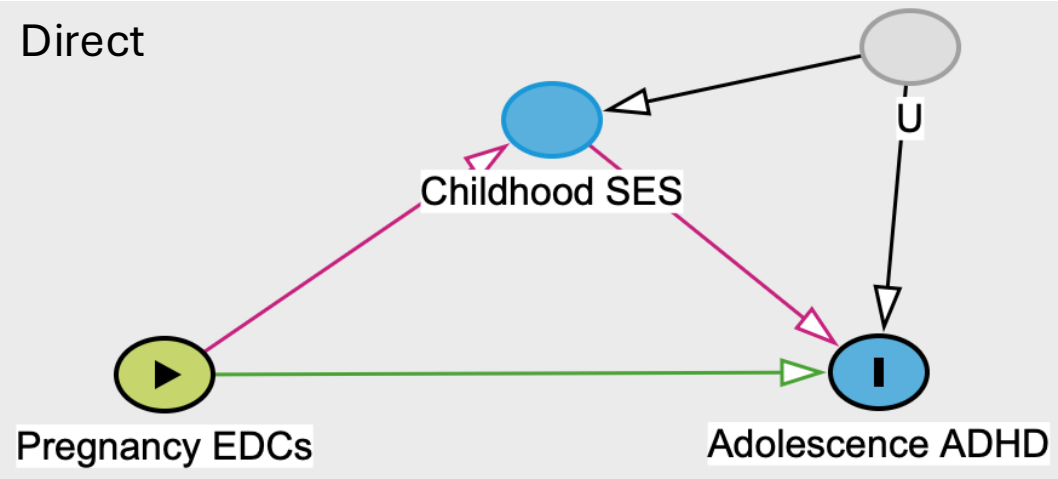
Triangulation and open questions

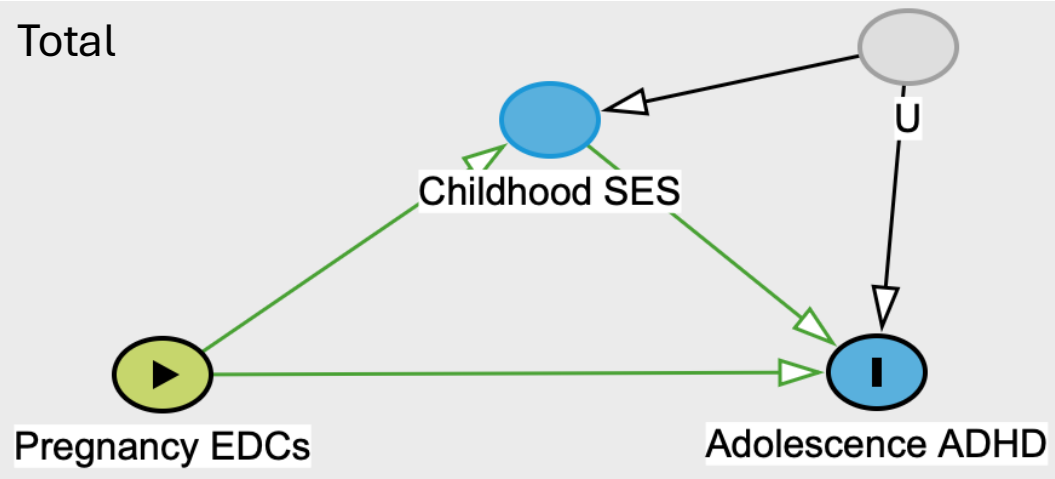
- Observational studies are subject to **bias** (many).
- We can strengthen our conclusions by combining different data sources and methods:
  - Negative control outcome, assuming “SES” is main culprit.



Triangulation and open questions

- **Direct or total effects**: "…when there is an unmeasured common cause of the intermediate and the outcome, associations adjusted for the intermediate are subject to bias."





- What if there is no arrow between EDCs and SES?

VanderWeele, T.J., Mumford, S.L. and Schisterman, E.F., 2012. Conditioning on intermediates in perinatal epidemiology. *Epidemiology*, *23*(1), pp.1-9.

Thank you for your attention!

You can contact me at lorenzo.fabbri@isglobal.org**.**