Machine learning and causal inference (causal inference with known structure)

Robert Castelo

robert.castelo@upf.edu

@robertclab

Dept. of Medicine and Life Sciences
Universitat Pompeu Fabra

Barcelona

Barcelona School of Economics
Data Science Methodology Program
Winter Term 2024-25

Motivation

- In observational studies we cannot assume ignorability, $X \perp\!\!\!\perp Y(x)$, because of *confounding* (treatment and potential outcomes share a common cause).

- If we knew the factors (covariates) Z that drive the confounding phenomenon, we could assume ignorability **conditional on covariates** Z:

$$X \perp\!\!\!\perp \{Y(0), Y(1)\} | Z\,,$$

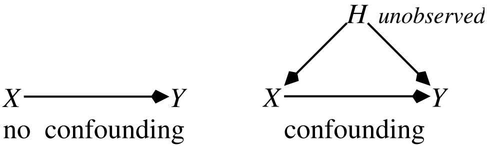
  also known in the literature as the*"conditional exchangeability"* or *"no unmeasured confounders"* assumption.

- We can use those confounder covariates Z to either obtain an estimate of the average treatment effect (ATE) adjusted for Z, or to calculate propensity scores $e(Z)$ and use them to mimic a randomized experiment by adjusting the ATE with $e(Z)$ as a covariate, or using a propensity score matching (PSM) algorithm.

- Key question: how do we choose Z? .. using graphical approaches to causality.

Motivation

- Using DAGs:



- Directed edges represent direct causal relationships and we assume that we can represent all potential confounders in the graph.

- Markov equivalence in DAGs complicates causal interpretations.

$$X \longrightarrow Y \qquad\qquad X \longleftarrow Y$$

$$p_{XY}(x,y) = p(x) \cdot p(y|x) \qquad p_{XY}(x,y) = p(y) \cdot p(x|y)$$

- Factorizing a joint probability distribution according to a graph is a necessary but not sufficient condition for a causal interpretation.

Observations and interventions

- Observational probability distribution P of an outcome $Y = y$ given a treatment $X = x$:

$$P(Y = y | X = x)\,,$$

by which we get the probability that $Y = y$ conditional on finding $X = x$.

- Interventional probability distribution P of an outcome Y given a treatment $X = x$ with the *do* operator (Pearl, 2009, pg. 70)[1]:

$$P(Y(x)|X = x) \equiv P(Y = y|do(X = x)) \equiv P(Y = y|do(x))\,,$$

by which we get the probability that $Y = y$ when we intervene to make $X = x$.

- Goal: find covariates Z using graphical criteria that allow us to estimate an interventional distribution from observational data (not always possible).

[1] Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Inverventions and modularity

- Given a DAG $\mathcal{D} = (V, E)$ and a probability distribution P Markov over $\mathcal{D}$, a corresponding joint probability function factorizes as:

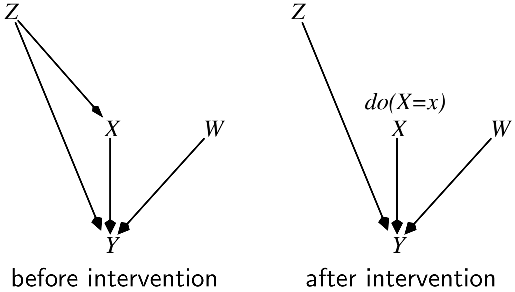
$$p_{X_V}(x_1, \ldots, x_p) = \prod_{i=1}^{p} p(x_i | x_{\mathrm{pa}(i)}) \,.$$

- When we *intervene* in one variable with $do(X = x)$ we will assume that such intervention is **modular**: changing the causal mechanism in one of variables doesn't change the causal mechanism in other variables.

- If we intervene on a subset of variables $X_I \subseteq X_V$, the modularity assumption implies:
  1. If $i \notin I$ then $P(x_i | x_{\mathrm{pa}(i)})$ remains unchanged.
  2. If $i \in I$ then $P(x_i | x_{\mathrm{pa}(i)}) = 1$ if $do(X_i = x_i)$ and $P(x_i | x_{\mathrm{pa}(i)}) = 0$ for every other $x_i' \neq x_i$.

Inverventions and modularity

- The modularity assumption has the following graphical counterpart:



- Modularity implies that the intervention only affects the incoming edges of the intervened variable, while the rest of the structure remains intact.
- Note that removing all incoming edges into X is also akin to a randomized experiment where X is the treatment variable.

Truncated factorization

- Altering the graph structure after an intervention has also a consequence in the factorization:

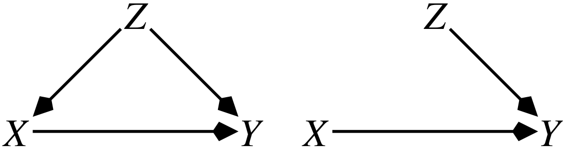
$$p_{X_V}(x_1, \dots, x_p) = \prod_{i=1}^{p} p(x_i | x_{\mathrm{pa}(i)}) \,.$$

- If we intervene on a subset of variables $X_I \subseteq X_V$ with $do(X_I = x_I)$, the resulting factorization is *truncated*:

$$p_{X_V}(x_1, \dots, x_p | do(X_I = x_I)) = \begin{cases} \prod_{i \notin I} p(x_i | x_{\mathrm{pa}(i)}) & \text{if } do(X_I = x_I) \\ 0 & \text{if } X_I = x'_I \text{ s.t. } x'_I \neq x_I \end{cases}$$

Truncated factorization: identification of causal effects

- Consider the DAG below before and after the intervention $do(X = x)$ with the goal of identifying $p(y|do(x))$.



- Step 1: Write the DAG factorization before intervention:

$$p(x, y, z) = p(z)p(x|z)p(y|x, z)\,.$$

- Step 2: Write the truncated factorization after intervention:

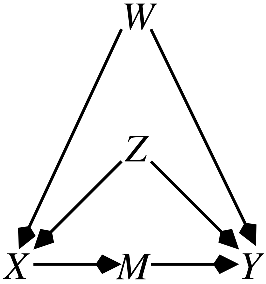
$$p(y, z|do(x)) = p(z)p(y|x, z)\,.$$

- Step 3: Marginalize over confounder Z:

$$p(y|do(x)) = \sum_{z} p(y|x, z)p(z)\,.$$

- Under what conditions is the structure of a DAG sufficient for estimating a causal effect from observational data?

- **Backdoor paths:** Given two variables X and Y between which we want to estimate the ATE, a *backdoor path* in a DAG $\mathcal{D} = (V, E)$ is a non-directed path from X to Y in $\mathcal{D}$ with no descendants of X, except for Y.



$X \leftarrow Z \rightarrow Y$ and $X \leftarrow W \rightarrow Y$ are backdoor paths; path $X \rightarrow M \rightarrow Y$ is **not** a backdoor path, because is a directed path from X to Y.

Backdoor criterion and backdoor adjustment

- **Backdoor criterion:** Given a DAG $\mathcal{D} = (V, E)$ and two variables X and Y between which we want to estimate the ATE, a subset of variables $Z \subseteq V \backslash \{X, Y\}$ satisfies the *backdoor criterion* w.r.t. X and Y in $\mathcal{D}$ if (Pearl, 2009, pg. 101):

  (i) no vertex in Z is a descendant of X; and

  (ii) Z blocks every path between X and Y that contains an arrow into X.

- A subset of vertices Z that meet the backdoor criterion is also known as a *sufficient adjustment set* and can be used to assume ignorability conditional on Z, also known as the *"no unmeasured confounders"* assumption.

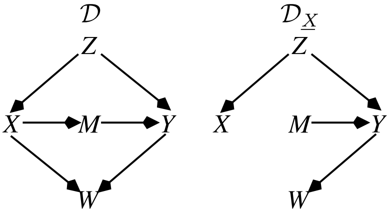
- The *"no unmeasured confounders"* assumption could be also rephrased as *"no unblockable backdoor paths"* assumption.

Backdoor criterion and backdoor adjustment

- **Backdoor adjustment:** Given a DAG $\mathcal{D} = (V, E)$ and a subset Z that satisfies the backdoor criterion w.r.t. X and Y, assuming modularity, we can identify the ATE as follows:

$$p(y|do(x)) = \sum_z p(y|x,z)p(z)\,.$$

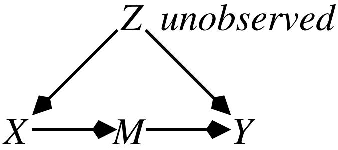
- The backdoor criterion can also be formulated in terms of d-separation, by first defining $\mathcal{D}_{\underline{X}}$ as $\mathcal{D}$ with all outgoing edges from X removed,



second, noting that Z blocks all backdoor paths from X to Y in $\mathcal{D}$, and third, verifying that Z d-separates X from Y in $\mathcal{D}_{\underline{X}}$, i.e., $X \perp_{\mathcal{D}_{\underline{X}}} Y | Z$.

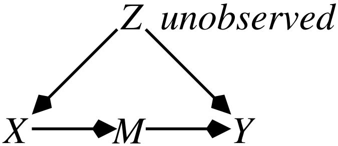
Frontdoor criterion and frontdoor adjustment

- Consider a DAG $\mathcal{D} = (V, E)$ with two variables X and Y between which we want to estimate the ATE, an unobserved confounder variable Z and a variable M that mediates the causal association between X and Y.



- Even though the confounder Z is not observed, we can still estimate the ATE between X and Y using M in three steps:

1. Identify the ATE of X on M.
2. Identify the ATE of M on Y.
3. Combine the previous steps to identify the ATE of X on Y.



- Step 1 (ATE of X on M): Since there are no backdoor paths from X to M,

$$p(m|do(x)) = p(m|x)\,.$$

- Step 2 (ATE of M on Y): We can block the backdoor path from M to Y by using X as backdoor adjustment set:

$$p(y|do(m)) = \sum_x p(y|m,x)p(x)\,.$$

- Step 3 (ATE of X on Y):

$$P(y|do(x)) = \sum_m p(m|do(x))p(y|do(m)) = \sum_m p(m|x) \sum_{x'} p(y|m,x')p(x')\,.$$

Frontdoor criterion and frontdoor adjustment

- **Frontdoor criterion:** Given a DAG $\mathcal{D} = (V, E)$ and two variables X and Y between which we want to estimate the ATE, a subset of variables $M \subseteq V \backslash \{X, Y\}$ satisfies the *frontdoor criterion* w.r.t. X and Y in $\mathcal{D}$ if (Pearl, 2009, pg. 81):

  (i) M completely mediates the effect of X on Y, i.e., all directed paths from X to Y go through M.

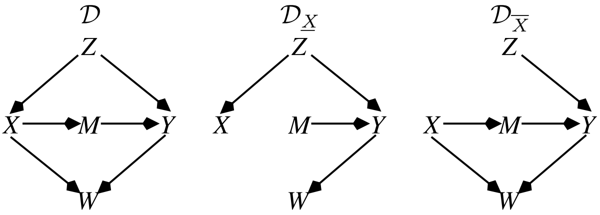
  (ii) There is no unblocked backdoor path from X to M.

  (iii) All backdoor paths from M to Y are blocked by X.

- The frontdoor criterion assumes that there are no confounders where we can apply the backdoor criterion.

Pearl's *do*-calculus

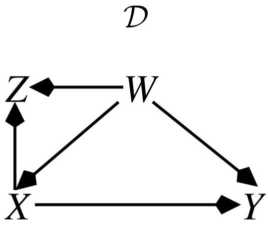
- Pearl (1995, 2009 pg. 85)[2] developed the so-called *do*-calculus to decide if a specific causal effect is identifiable in a DAG. It can be also seen as a rule system for replacing do-interventions with ordinary conditional probabilities.
- Intuitively, the rules of *do*-calculus aim to identify variables that we can safely ignore when estimating the ATE.
- Given a DAG $\mathcal{D} = (V, E)$, consider the following two transformations of $\mathcal{D}$, denoted by $\mathcal{D}_{\underline{X}}$ and $\mathcal{D}_{\overline{X}}$:

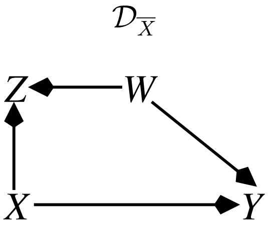


original DAG | X-outgoing edges removed | X-incoming edges removed

[2] Pearl, J. (1995) Causal diagrams for empirical research. *Biometrika*, 84:669-710. https://doi.org/10.1093/biomet/82.4.669.

Pearl's *do*-calculus



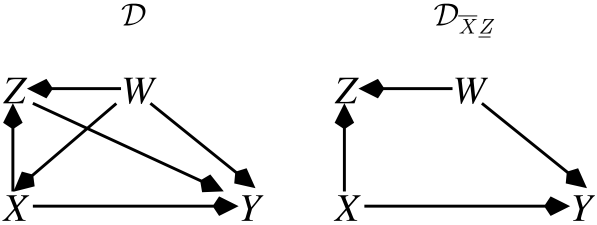


- Rule 1: *Ignoring observations*

$$p(y|do(x), z, w) = p(y|do(x), w) \text{ if } Y \perp_{\mathcal{G}_{\overline{X}}} Z|X, W\,.$$

- Intuition for rule 1: if we remove the intervention $do(x)$, then $p(y|z, w) = p(y|w)$ if $Y \perp_{\mathcal{G}} Z|W$, i.e., the Markov property for DAGs. Rule 1 can be interpreted as a generalization of d-separation for interventional distributions.

Pearl's *do*-calculus



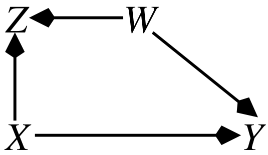
- Rule 2: *Replace interventions by observations*

$$p(y|do(x), do(z), w) = p(y|do(x), z, w) \text{ if } Y \perp_{\mathcal{G}_{\overline{X},\underline{Z}}} Z | X, W .$$

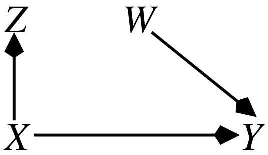
- Intuition for rule 2: if we remove the intervention $do(x)$, then $p(y|do(z), w) = p(y|z, w)$ if $Y \perp_{\mathcal{G}_{\underline{Z}}} Z | W$, i.e., the backdoor adjustment in terms of d-separation.

Pearl's *do*-calculus

$\mathcal{D}$



$\mathcal{D}_{\overline{X}\,\overline{Z(W)}}$



- Rule 3: *Ignoring interventions*

$$p(y|do(x), do(z), w) = p(y|do(x), w) \text{ if } Y \perp_{\mathcal{G}_{\overline{X},\overline{Z(W)}}} Z|X, W\,.$$

  where $Z(W)$ refers to variables in Z that are not ancestors of W in $\mathcal{G}_{\overline{X}}$.

- Intuition for rule 3: if we remove the intervention $do(x)$, then $p(y|do(z), w) = p(y|w)$ if $Y \perp_{\mathcal{G}_{\overline{Z(W)}}} Z|W$, i.e., we can ignore an intervention $do(z)$ when it does not influence the outcome Y through any path.

Backdoor, frontdoor and *do*-calculus

- Backdoor and frontdoor criteria are *sufficient* for estimating ATEs, but are not *necessary*, i.e., if they are not satisfied, it does not mean we could not estimate the ATE (identifiability).
- Backdoor and frontdoor criteria can be derived using the rules of *do*-calculus, but when they apply, they are easier to use than *do*-calculus.
- *do*-calculus is *complete*, i.e., necessary and sufficient, for estimating all identifiable ATEs (Shpitser and Pearl[3], 2006; Huang and Valtorta[4], 2006).
- The proof of completeness is constructive and provides polynomial-time algorithms for the identification of (identifiable) ATEs.
- There are necessary and sufficient graphical criteria for establishing the identifiability of (some) ATEs (Tian and Pearl, 2002)[5].

---

[3] Shpitser, I. and Pearl, J. (2006) Identification of conditional interventional distributions. *UAI 2006*, pp. 437-444.

[4] Huang, Y. and Valtorta, M. (2006) Pearl's calculus of intervention is complete. *UAI 2006*, pp. 217-224.

[5] Tian, J. and Pearl, J. (2002) A general identification condition for causal effects. *AAAI 2002*, pp. 567-573.

Graphical and non-graphical approaches to causality

- Neyman-Rubin's framework of potential outcomes *does not use* graphs.
- Pearl's framework of *do*-calculus *does use* graphs.
- Are these two frameworks related to each other?

no confounding	confounding
$X \perp\!\!\!\perp Y(0)$ & $X \perp\!\!\!\perp Y(1)$	$X \not\perp\!\!\!\perp Y(0)$ or $X \not\perp\!\!\!\perp Y(1)$

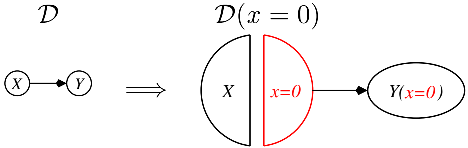
- Elephant in the room:

Variables $Y(0)$ and $Y(1)$ do not appear in these graphs !!

Single-world intervention graphs (SWIGs)

- Single world intervention graphs (SWIGs), introduced by Richardson and Robins (2013)[6], provide an explicit way to connect the non-graphical framework of potential outcomes with graphical approaches to causality.

- The node splitting operation, split the treatment variable into a random piece and a fixed piece representing the intervention, e.g., setting $x = 0$:

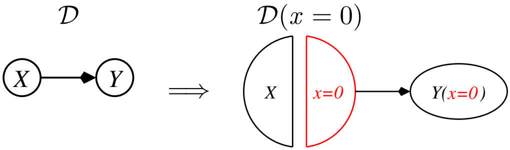


- Given a DAG $\mathcal{D}$, after the node splitting operation we obtain a SWIG based on $\mathcal{D}$ denoted by $\mathcal{D}(x = 0)$.

[6] Richardson, T.S. and Robins, J.M. (2013) Single world intervention graphs (SWIGs): a unification of counterfactual and graphical approaches to causality. *Univ. Washington Center Stat. Soc. Sci. Working Papers*, 128. https://csss.uw.edu/files/working-papers/2013/wp128.pdf.

Single-world intervention graphs (SWIGs)

- Setting $x = 0$:



- After node splitting, we can read $X \perp\!\!\!\perp Y(x = 0)$.

- This is the corresponding factorization:

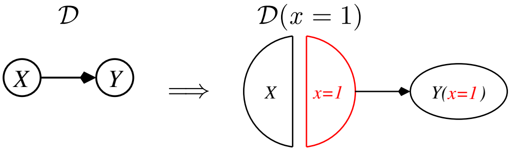
$$p(X = x, Y(x = 0) = y) = p(X = x)p(Y(x = 0) = y)\,,$$

  where

$$p(Y(x = 0) = y) = p(Y = y | X = 0)\,.$$

Single-world intervention graphs (SWIGs)

- Setting $x = 1$:



- After node splitting, we can read $X \perp\!\!\!\perp Y(x = 1)$.

- This is the corresponding factorization:

$$p(X = x, Y(x = 1) = y) = p(X = x)p(Y(x = 1) = y)\,,$$

where

$$p(Y(x = 1) = y) = p(Y = y|X = 1)\,.$$

Single-world intervention graphs (SWIGs)

- The SWIG $\mathcal{D}(x=0)$ is associated with the distribution $P(X, Y(x=0))$.
- The SWIG $\mathcal{D}(x=1)$ is associated with the distribution $P(X, Y(x=1))$.
- Under no confounding, these marginals are identifed from $P(X, Y)$.
- However, the distribution $P(X, Y(x=0), Y(x=1))$ is not identified, because $Y(x=0)$ and $Y(x=1)$ are **never** on the same SWIG.
- Although we have weak ignorability:

$$X \perp\!\!\!\perp Y(x=0) \;\text{ and }\; X \perp\!\!\!\perp Y(x=1)\,,$$

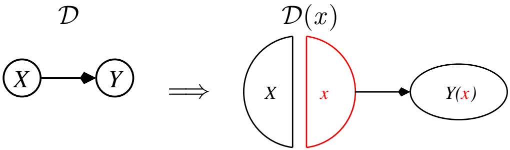
  we do **not** assume strong ignorability:

$$X \perp\!\!\!\perp Y(x=0), Y(x=1)\,.$$

- Constructing a single graph continaing both $Y(x=0)$ and $Y(x=1)$ is impossible, hence the name *Single-World Interention Graphs (SWIGs)*.

Single-world intervention graphs (SWIGs)

- Represent both graphs using a *template*:



- Formally, the template is a *graph valued function*:
  - Takes as input a specific value x^*.
  - Returns as output a SWIG $\mathcal{D}(x^*)$.

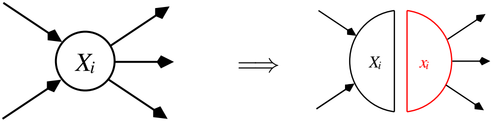
- Each *instantiation* of the template represent a different margin:
  - SWIG $\mathcal{D}(x = 0)$ represents $P(X, Y(x = 0))$.
  - SWIG $\mathcal{D}(x = 1)$ represents $P(X, Y(x = 1))$.

Single-world intervention graphs (SWIGs)

- Let $\mathcal{D} = (V, E)$ be a DAG and $I \subseteq V$ a subset whose associated variables X_I are going to be intervened. Let x_I represent the values set in the intervention, i.e., $X_I = x_I$.

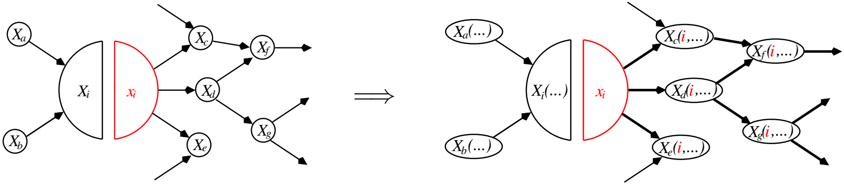
- We form the SWIG template $\mathcal{D}(x_I)$ in two steps.

- Step 1 (node splitting): for every intervened variable $X_i = x_i$, split the node X_i into a random part X_i and a fixed part x_i:



where the random half inherits all incoming edges into X_i and the fixed half inherits all outgoing edges from X_i.

- Step 2 (node relabeling): for every fixed node, label every of its descendants with the value set in the intervention:



- Given a DAG $\mathcal{D} = (V, E)$ and a joint probability distribution $P(X_V)$ Markov over $\mathcal{D}$, and given a SWIG $\mathcal{D}(x_I)$ obtained by intervening with $X_I = x_I$ and $I \subseteq V$, we may consider a *counterfactual distribution* $P(X_V(X_I = x_I))$, Markov over $\mathcal{D}(X_I)$.

Concluding remarks

- DAGs can have a causal interpretation under certain assumptions (no unmeasured confounders, modularity) and help identifying covariates to assume ignorability.

- Backdoor and frontdoor criteria can used to derive an expression of the causal effect in terms of observed probabilities, if the effect is identifiable.

- Pearl's *do*-calculus allows one to replace interventions with observed conditional probabilities, but it doesn't play well with potential outcomes and counterfactuals. An implementation of *do*-calculus and other extensions is available in the R packages `causaleffect` (Tikka and Karvanen, 2017)[7] and `dosearch` (Tikka, Hyttinen and Karvanen, 2021)[8].

[7] Tikka, S. and Karvanen, J. Identifying causal effects with the R package causaleffect, *J. Stat. Soft.*, 76:1-30. https://doi.org/10.18637/jss.v076.i12.

[8] Tikka, S., Hyttinen, A. and Karvanen, J. Causal effect identification from multiple incomplete data sources: a general search-based approach, *J. Stat. Soft.*, 99:1-40. https://doi.org/10.18637/jss.v099.i05.

Concluding remarks

- SWIGs are an attempt to bring together the potential outcomes and graphical frameworks (Shpitser et al.[9], 2022; Robins et al.[10], 2022).
- SWIGs can generalise *do*-calculus with the so-called *po*-calculus (Malinsky et al., 2019)[11].
- SWIGs have been employed to provide a visual representation of estimands in clinical trials[12].

---

[9] Shpitser, I., Richardson, T.S. and Robins, J.M. Multivariate counterfactual systems and causal graphical models, *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 813-852, 2022. https://doi.org/10.1145/3501714.3501757

[10] Robins, J.M., Richardson, T.S., and Shpitser, I. An interventionist approach to mediation analysis, *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 713-764, 2022. https://doi.org/10.1145/3501714.3501754

[11] Malinksy, D., Shpitser, I., and Richardson, T.S. A potential outcomes calculus for identifying conditional path-specific effects, In *Proc. AISTATS, PMLR*, 89:3080-3088, 2019. https://proceedings.mlr.press/v89/malinsky19b.html

[12] Ocampo, A. and Bather, J.R. Single-world intervention graphs for defining, identifying, and communicating estimands in clinical trials, *Statistics in Medicine*, 42:3892-3902, 2023. https://doi.org/10.1002/sim.9833