Discussion Assignment 2

Introduction to Causal Inference (Biostat683)

Assigned: October 4, 2021
Group presentation: In class on October 18, 2021. Group slides must be uploaded to GoogleDrive by 2pm on Oct 18. Please make sure to include your names and group number on the slides. One presentation per group
Individual write-ups: NO individual-level write-ups required. BUT you must work with your group on the slides and make them as detailed as possible. If 2+ group members disagree on a given question, you can include multiple slides to represent differing responses and their justifications.
Study: If you were assigned Study 1 in assignment #1, then you will complete answers for Study 2 in this assignment. Likewise, if you were assigned Study 2 in assignment #1, then you will complete answers for Study 1 in this assignment.

1 Instructions

For the previous two studies, **think** through the following questions. **Using the following background (Steps 0-2 of the Causal Roadmap), please provide brief *written* answers to the questions for your assigned study.** Use the notation developed in class. You are encouraged to discuss as a group, but please submit your own written responses. **As a group, be prepared to *present* your answers in class** (October 18, 2021).

2 Background (Steps 0-2 of the roadmap) for Studies #1-2

2.1 Study #1: Physical activity and mortality in the elderly

- **Step 0:** What is the effect of free-living energy expenditure on the seven-year survival among older active adults.
- **Step 1:** Consider the following structural causal model $\mathcal{M}^*$. Other causal models from assignment 1 may also be correct, but to simplify discussion we will work with the following.
  - Endogenous variables: $X = (W, A, Y)$
  - Background variables: $U = (U_W, U_A, U_Y) \sim \mathbb{P}_U$
  - Structural equations F:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, A, U_Y) \end{aligned}$$

    where W={smoking, comorbidities, body fat}, A={energy expenditure} and Y={survival}. By collapsing the baseline variables into a single node W, we lose the information that smoking was measured first. However, there are no consequences for our statistical estimation problem.
  - There are no independence assumptions on $\mathbb{P}_U$.
- **Step 2:** Suppose we are interested in the average treatment effect (i.e., the causal risk difference).
  - The target causal parameter is the difference in the counterfactual probability of survival if all elderly adults had a high energy expenditure ($A = 1$) and the counterfactual probability of survival if all elderly adults had a low energy expenditure ($A = 0$)

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0)$$

where the counterfactual outcome Y_a is the seven-year survival for an individual if, possibly contrary to fact, they had energy expenditure $A = a$ for this two week period.

2.2 Study #2: Effect of male circumcision on risk of HIV acquisition

- **Step 0:** What is the effect of male circumcision on HIV acquisition after two years? More specifically, what is the causal effect of male circumcision on the probability of becoming infected with HIV over two years in this rural Kenyan population?
- **Step 1:** Consider the following structural causal model $\mathcal{M}^*$. Other causal models from assignment 1 may also be correct, but to simplify discussion we will work with the following.
  - Endogenous variables: $X = (W, A, Z, Y)$
  - Background variables: $U = (U_W, U_A, U_Z, U_Y) \sim \mathbb{P}_U$
  - Structural equations F:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Z &= f_Z(W, A, U_Z) \\ Y &= f_Y(W, A, Z, U_Y) \end{aligned}$$

    where W={tribe, religion}, A=male circumcision, Z={sexual behavior, STI}, and Y=HIV status
  - There are no independence assumptions on $\mathbb{P}_U$.
- **Step 2:** Suppose we are interested in the average treatment effect (i.e., the causal risk difference).
  - The target causal parameter is the difference in the counterfactual risk of HIV acquisition if all males were circumcised ($A = 1$) and the counterfactual risk of HIV acquisition if all males were not circumcised ($A = 0$)

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0)$$

    where the counterfactual outcome Y_a is the two-year HIV status for an individual, if possibly contrary to fact, they had circumcision status $A = a$.

3 Questions to be answered

1. *Step 3: Observed data & link to causal model.*
   (a) Specify the observed data.
   (b) What notation do we use to refer to the distribution of the observed data?
   (c) What is the link between the structural causal model (SCM) and the observed data?
   (d) What is the statistical model $\mathcal{M}$? Does the SCM place any restrictions on $\mathcal{M}$?

Solution: STUDY 1:

(a) The observed data consist of n i.i.d. copies of the random variable $O = (W, A, Y)$, where W consists of smoking, comorbidities and body fat; A denotes the exposure (energy expenditure), and Y denotes the outcome (seven-year survival). Here, the observed data O are the same as the endogenous factors X, but this is not always the case. We may have included unmeasured variables in X. For example, socioeconomic status might be an unmeasured variable, included as endogenous X node.

(b) The random variable O has distribution $\mathbb{P}_0$:

$$O = (W, A, Y) \sim \mathbb{P}_0$$

(c) We assume the observed data were generating by sampling n times from a data generating system contained in (described by) the structural causal model $\mathcal{M}^*$, resulting in n i.i.d copies of the random variable O. This provides a link between the causal model $\mathcal{M}^*$ and the statistical model $\mathcal{M}$.

(d) The statistical model $\mathcal{M}$ is the set of possible distributions for the observed data. We have not placed any restrictions on the statistical model, which is thereby non-parametric.

Solution: STUDY 2:

(a) The observed data consist of n i.i.d. copies of the random variable $O = (W, A, Z, Y)$, where W consists of tribe and religion; A denotes the exposure (male circumcision); Z consists of sexual behavior and STIs, and Y denotes the outcome (HIV acquisition). Here, the observed data O are the same as the endogenous factors X, but this is not always the case. We may have included unmeasured variables in X.

(b) The random variable O has distribution $\mathbb{P}_0$:

$$O = (W, A, Z, Y) \sim \mathbb{P}_0$$

(c) We assume the observed data were generating by sampling n times from a data generating system contained in (described by) the structural causal model $\mathcal{M}^*$. This provides a link between the causal model $\mathcal{M}^*$ and the statistical model $\mathcal{M}$.

(d) The statistical model $\mathcal{M}$ is the set of possible distributions for the observed data. We have not placed any restrictions on the statistical model, which is thereby non-parametric.

2. *Steps 4-5: Identifiability & Committing to a statistical estimand*

   (a) Using the backdoor criterion, assess identifiability of $\Psi^*(\mathbb{P}^*)$.

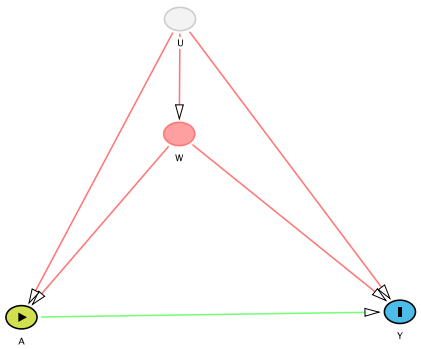
   (b) If not identified, under what assumptions would it be? Are some of these sets of additional assumptions more plausible than others? Are there additional measurements you could make so that the needed identifiability assumptions are more plausible?

   (c) Specify the target parameter of the observed data distribution (i.e., the statistical estimand).

   (d) What is the relevant positivity assumption? Are you concerned about violations of the positivity assumption in your study?

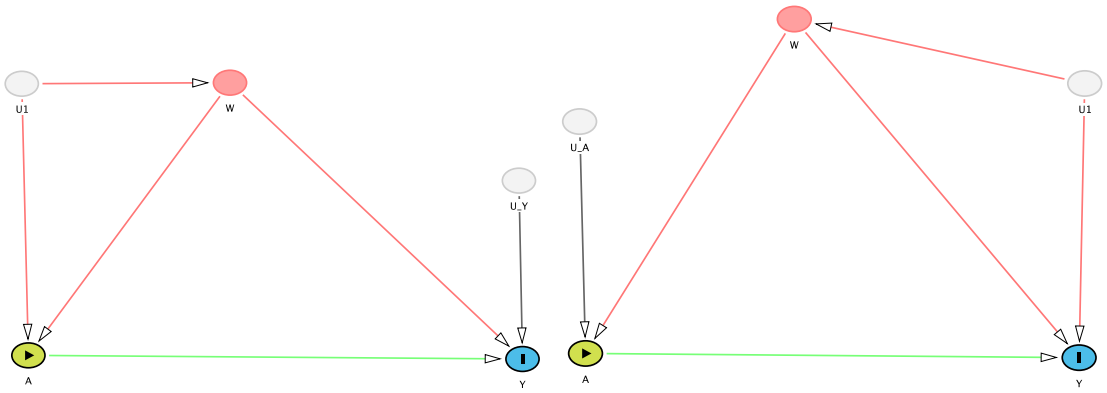
Solution: STUDY 1:

(a) No. There are several unblocked back-door paths from the outcome Y to exposure A. See Figure 1.



Solution Fig. 1: Causal graph for the energy and mortality study.

(b) We would need to place independence assumptions on the distribution of unmeasured factors $\mathbb{P}_U$. Specifically, we need $U_A \perp\!\!\!\perp U_Y$ and either $U_A \perp\!\!\!\perp U_W$ or $U_W \perp\!\!\!\perp U_Y$. See Figure 2.
- It is hard to say whether one set of independence assumptions is more plausible than the other. Both energy expenditure A and survival Y are affected by a tremendous number of factors, many of which we do not understand. For example, complex social and genetic factors are expected to have large impacts on both A and Y.
- That said, doing our best possible job to measure determinants of survival Y in this population would increase the plausibility of these assumptions. Suppose we measure all of the determinants of survival Y that might possibly affect energy expenditure A. These determinants could be included in W as long as they occur before the exposure A or are known not to be affected by it. Then we would have $U_A \perp\!\!\!\perp U_Y$; $U_W \perp\!\!\!\perp U_Y$.



Solution Fig. 2: Causal graph for study 1 where the back-door criterion would hold conditional on the covariates W

(c) While the causal risk difference is not identified under our causal model $\mathcal{M}^*$, we still have an interesting statistical estimand, given by the G-computation formula:

$$\begin{aligned}\Psi(\mathbb{P}_0) &= \mathbb{E}_0\Big[\mathbb{E}_0(Y|A=1,W) - \mathbb{E}_0(Y|A=0,W)\Big] \\ &= \sum_w \big[\mathbb{E}_0(Y|A=1,W=w) - \mathbb{E}_0(Y|A=0,W=w)\big]\mathbb{P}_0(W=w)\end{aligned}$$

where the summation generalizes to the integral for continuous covariates. Formally, the parameter Ψ is a mapping from the statistical model $\mathcal{M}$ to the parameter space $\Psi : \mathcal{M} \mapsto \mathbb{R}$. In other words, Ψ is a function with input a distribution in $\mathcal{M}$ and output a value in the parameter space (here, a real number bounded between -1 and 1).

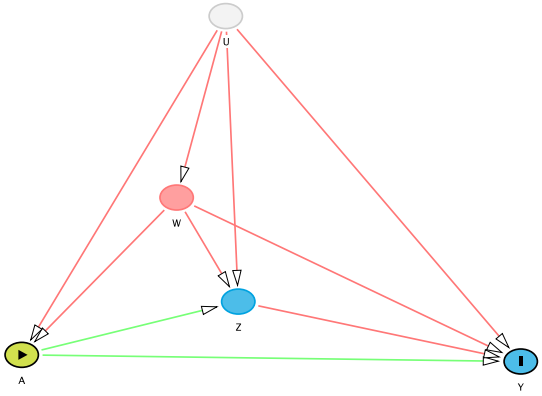
(d) Each exposure of interest must occur with positive probability for each possible value of the covariates:

$$\min_{a\in\mathcal{A}} \mathbb{P}_0(A=a|W=w) > 0, \text{ for all } w \text{ for which } \mathbb{P}_0(W=w) > 0$$

There are definitely concerns about positivity violations in this study. For example, suppose $A=1$ denotes high energy expenditures (e.g., vigorous physical activity for at least two hours each day). This amount of exercise might be contraindicated for people who recently have suffered from a heart attack: $\mathbb{P}_0(A=1|\text{recent heart attack}) = 0$. Likewise, the positivity assumption might be "practically violated" for a given sample $\mathbb{P}_n$ of $\mathbb{P}_0$. In the sample, there might not be any smokers with high energy expenditures: $\mathbb{P}_n(A=1|\text{smoker}) = 0$.

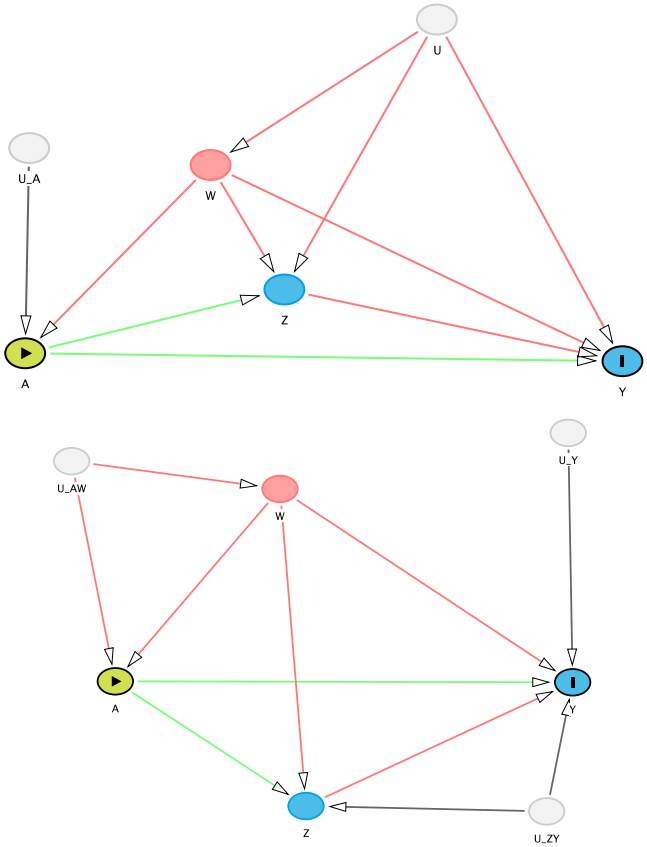
Solution: STUDY 2:

(a) No. There are several unblocked backdoor paths from outcome Y to exposure A. See Figure 3.



Solution Fig. 3: Causal graph for Study 2.

(b) We would need to place independence assumptions on the distribution of unmeasured factors $\mathbb{P}_U$. Specifically, we would need $U_A \perp\!\!\!\perp U_Y$, $U_A \perp\!\!\!\perp U_Z$, and either (a) $U_A \perp\!\!\!\perp U_W$ or (b) $U_W \perp\!\!\!\perp U_Y$ and $U_W \perp\!\!\!\perp U_Z$. See Figure 4.



Solution Fig. 4: Causal graphs for Study 2 where the back door criterion would hold conditional on W.

- Note: We need the additional assumption $U_A \perp\!\!\!\perp U_Z$ to avoid unmeasured common causes of A and Y (through Z) not blocked by W.

- This example makes it clear that we should be concerned with measuring factors that affect both circumcision and sexual behavior/STIs (to improve the plausibility that $U_A \perp\!\!\!\perp U_Z$) as well as factors that affect HIV acquisition Y directly. (If we have unmeasured factors that affect both A and Z, we have no way to close the backdoor path to Y; even though we measure Z, we cannot condition on it.)

- Further, this is an example where it might be easier to satisfy the weaker independence assumptions (listed above) rather than the stronger assumption that all of the background factors are independent.

(c) While the causal risk difference is not identified under our causal model $\mathcal{M}^*$, we still have an interesting statistical estimand, given by the G-computation formula:

$$\begin{aligned}\Psi(\mathbb{P}_0) &= \mathbb{E}_0\Big[\mathbb{E}_0(Y|A=1,W) - \mathbb{E}_0(Y|A=0,W)\Big] \\ &= \sum_w \big[\mathbb{E}_0(Y|A=1,W=w) - \mathbb{E}_0(Y|A=0,W=w)\big]\mathbb{P}_0(W=w)\end{aligned}$$

Formally, the parameter Ψ is a mapping from the statistical model $\mathcal{M}$ to the parameter space $\Psi : \mathcal{M} \mapsto \mathbb{R}$. In other words, Ψ is a function with input as a distribution in $\mathcal{M}$ and output a value in the parameter space (e.g., a number).

(d) Each exposure of interest must occur with positive probability for each possible value of the covariates:

$$\min_{a\in\mathcal{A}} \mathbb{P}_0(A=a|W=w) > 0, \text{ for all } w \text{ for which } \mathbb{P}_0(W=w) > 0$$

There may be concern about positivity violations in this study. For example, suppose all males are circumcised at birth in certain tribes or religious groups: $\mathbb{P}_0(A=0|W=w)=0$ for some w. Likewise, the positivity assumption might be "practically violated" for a given sample $\mathbb{P}_n$ of $\mathbb{P}_0$. In the sample, there might not be any males who are circumcised and in a particular tribe: $\mathbb{P}_n(A=1|W=w)=0$

3. *Study-specific question for Study 1:* Suppose the investigators assume no unmeasured common causes of (W, A, Y) is this necessary for identifiability? Is it sufficient?

4. *Study-specific question for Study 2:*The study investigators adjust for (condition on) W={tribe, religion} and Z={sexual behavior, STI}. Under what causal structure would W would satisfy the back door criterion, but (W, Z) would not? Under what causal structure would (W, Z) satisfy the back door criterion, but W alone would not?

Solution: STUDY 1:

The assumption that all the background factors U are independent ($U_Y \perp\!\!\!\perp U_W$; $U_A \perp\!\!\!\perp U_W$; and $U_A \perp\!\!\!\perp U_Y$) is sufficient (in combination with the positivity assumption), but not necessary for identifiability. In other words, we could weaken the independence assumptions as follows. If we have $U_A \perp\!\!\!\perp U_Y$ and $U_A \perp\!\!\!\perp U_W$, then we do not need $U_Y \perp\!\!\!\perp U_W$. Similarly, if we have $U_A \perp\!\!\!\perp U_Y$ and $U_Y \perp\!\!\!\perp U_W$, then we do not need $U_A \perp\!\!\!\perp U_W$.

When might this be helpful? Suppose we understand and have measured all the proximal factors that go into determining the outcome Y. This tells us it is not necessary to understand and measure all distal factors that affect Y only through W and also affect A. For example, if we think that unmeasured factors like SES only affect survival through the measured factors like health care access, smoking habits and comorbidities, then we may have identifiability despite the fact that we did not measure SES and it is a common cause of energy expenditure and also survival. That said, with an outcome like survival, which is affected by a lot of things, this is maybe not too helpful. (It is a bit of a stretch to say we have measured all of the proximal determinants of survival. For example, SES probably does have various effects on survival via pathways not captured by measured W.) However, the sufficiency of a weaker identifiability assumption might be more helpful with an outcome like HIV infection, where the proximal causes are better understood.

Solution: STUDY 2:

- *Under what causal structure would W satisfy the back door criterion, but (W, Z) not?*
  - Endogenous variables: $X = (W, A, Z, Y)$
  - Background factors: $U = (U_W, U_A, U_Z, U_Y) \sim \mathbb{P}_U$, where $U_A \perp\!\!\!\perp U_Y$, $U_A \perp\!\!\!\perp U_Z$, and either (a) $U_A \perp\!\!\!\perp U_W$ or (b) $U_W \perp\!\!\!\perp U_Y$ and $U_W \perp\!\!\!\perp U_Z$ (i.e., the answer to Question 2).
  - Structural equations F:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Z &= f_Z(W, A, U_Z) \\ Y &= f_Y(W, A, Z, U_Y) \end{aligned}$$

  Here, Z is a descendant of A and thereby disqualified from satisfying the backdoor criterion (for a single timepoint intervention). If we condition on Z, it would block part of the effect of A on Y. One could argue that the effect of A through Z is not of interest. In other words, we are really interested in a direct effect: $\mathbb{E}^*(Y_{a=1,z} - Y_{a=0,z})$. If Z were just sexual behavior, it would be easier to argue for this parameter. Since Z is both sexual behavior and STIs, it is a little trickier.

  However, the same set of assumptions is not sufficient to identify the direct effect of male circumcision A on HIV acquisition Y, not through the mediator of sexual behavior/STIs Z. Mediation is closely related to longitudinal (multiple time point) causal effects and identifiability. A sneak preview: if you think of (A, Z) as a joint intervention node, we need W to satisfy backdoor criterion w.r.t. effect of (A, Z) on Y. Sufficient conditions to identify an effect such as $\mathbb{E}^*(Y_{a=1,z} - Y_{a=0,z})$ are $U_A \perp\!\!\!\perp U_Y$, $U_Z \perp\!\!\!\perp U_Y$, and either (a) $U_A \perp\!\!\!\perp U_W$ and $U_Z \perp\!\!\!\perp U_W$ or (b) $U_W \perp\!\!\!\perp U_Y$.

- *Study-specific additional question: Under what causal structure would (W, Z) satisfy the back door criterion, but W alone would not?*

  – One option:
  - Endogenous variables: $X = (W, A, Z, Y)$
  - Background factors: $U = (U_W, U_A, U_Z, U_Y) \sim \mathbb{P}_U$, where $U_A \perp\!\!\!\perp U_Y$, $U_Y \perp\!\!\!\perp U_Z$, and either (a) $U_A \perp\!\!\!\perp U_W$ or (b) $U_W \perp\!\!\!\perp U_Y$.
  - Structural equations F:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Z &= f_Z(W, U_Z) \\ Y &= f_Y(W, A, Z, U_Y) \end{aligned}$$

  Here, we have made an additional exclusion restriction on f_Z: circumcision status A does not affect sexual risk behavior and acquisition of STIs Z. (The former may be plausible, but the latter is really not. In other words, we are not saying that this is a reasonable causal model for this study, but rather using it to understand what types of assumptions/underlying causal structure would be needed for the adjustment strategy used in the paper, which also adjusts for Z, to be warranted). We have also altered our independence assumptions. Now we can allow for unmeasured common causes of A and Z (e.g., culture or family values), and thus open a back door path $A \leftarrow U \rightarrow Z \rightarrow Y$. The concern that circumcision status may be associated with high risk behaviors was likely the motivation for the investigators' decision to adjust for these post-exposure risk factors in the initial study. However, the example illustrates the additional assumptions needed for this approach to work. In addition to the exclusion restriction, we now need to assume there are no unmeasured common causes of risk behaviors/STI acquisition and HIV acquisition: $U_Z \perp\!\!\!\perp U_Y$. This independence assumption is needed to avoid introducing new spurious associations between A and Y through conditioning on collider Z.
  For the modified SCM, the statistical estimand is:

$$\Psi(\mathbb{P}_0) = \sum_{w,z} \big[\mathbb{E}_0(Y|A=1, W=w, Z=z) - \mathbb{E}_0(Y|A=0, W=w, Z=z)\big]\mathbb{P}_0(W=w, Z=z)$$

We now rely on Z as well as W to satisfy the backdoor criteria. This is why we "commit" to a target parameter of the observed data distribution. The target parameter may be different depending on which set of identifiability assumptions you choose. (Of course, you can do more than one, and discuss the results of each in light of the plausibility of the identifiability assumptions that each requires...)

– Another option, if one is willing to change the causal ordering assumption:
- Endogenous variables: $X = (W, Z, A, Y)$
- Background factors: $U = (U_W, U_A, U_Z, U_Y) \sim \mathbb{P}_U$, where $U_A \perp\!\!\!\perp U_Y$, and either (a) $U_A \perp\!\!\!\perp (U_W, U_Z)$ or (b) $(U_W, U_Z) \perp\!\!\!\perp U_Y$.
- Structural equations F:

$$
\begin{aligned}
W &= f_W(U_W) \\
Z &= f_Z(W, U_Z) \\
A &= f_A(W, Z, U_A) \\
Y &= f_Y(W, A, Z, U_Y)
\end{aligned}
$$

This will give the same estimand as above.