Discussion Assignment 2

Katherine Wolf
Introduction to Causal Inference (PH252D)
March 8, 2020

1 Instructions

2 Background

3 Questions to be answered

1. **Specify your observed data.**

   (a) What notation do we use to refer to the distribution of the observed data?
   $\mathbb{P}_{\boldsymbol{O}}$, as in $\boldsymbol{O} = (\boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X}) \sim \mathbb{P}_{\boldsymbol{O}}$.

   (b) Specify the link between the SCM and the observed data.
   We link the the observed data to the SCM by assuming that we obtained them from a data-generating system described by our SCM, i.e., that each observation in the data represents a draw from the unknown probability distribution $\mathbb{P}_{\boldsymbol{U}}$ of the exogenous variables $\boldsymbol{U}$, i.e., we drew $\boldsymbol{U} = \boldsymbol{u}$, that we then plugged into our structural equations $\mathcal{F}$ to output a specific $\boldsymbol{X} = \boldsymbol{x}$, of which we observed the subset $\boldsymbol{O} = \boldsymbol{o}$.
   Thus we assume that the the structural equations $\mathcal{F}$ as applied to $\boldsymbol{U} \sim \mathbb{P}_{\boldsymbol{U}}$ will identify the distribution $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$ and, since $\boldsymbol{O} \subseteq \boldsymbol{X}$, the distribution of its observed subset $\boldsymbol{O} \sim \mathbb{P}_{\boldsymbol{O}}$. (Since here we assume that $\boldsymbol{X} = \boldsymbol{O}$, we can write $P_O(O = o) = \sum_u P_f(X = x|U = u)P(U = u) = \sum_u I(X(u) = x|U = u)P(U = u)$.)

   (c) What is the statistical model $\mathcal{M}$?
   The statistical model $\mathcal{M}$ consists of

   - The endogenous variables $\boldsymbol{X} = (W, A, Y)$;
   - The exogenous variables $\boldsymbol{U} = (U_W, U_A, U_Y) \sim \mathbb{P}_U$; and
   - The structural equations $\mathcal{F}$:

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

   The structural equations in directed acyclic graph (DAG) format (Figure 1):

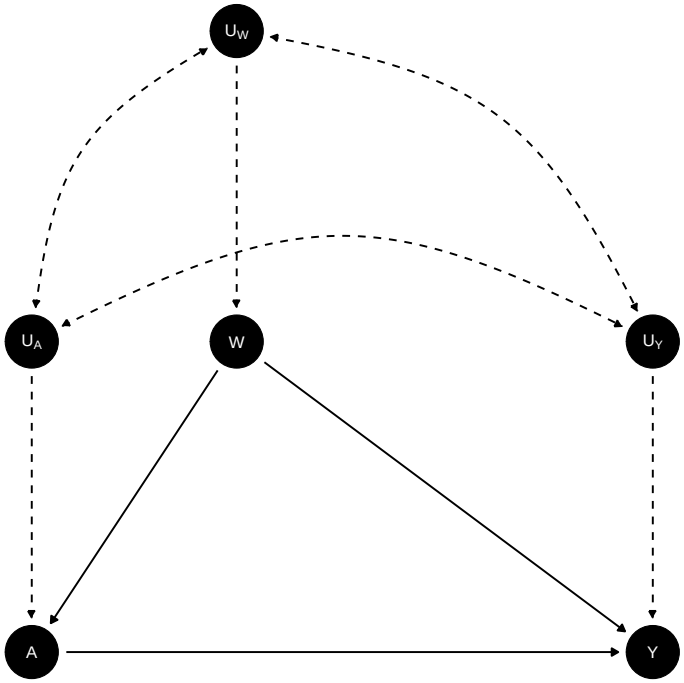
   Does the SCM place any restrictions on $\mathcal{M}$?
   There are no independence assumptions, i.e., no restrictions on the distribution of the exogenous variables $\mathbb{P}_{\boldsymbol{U}}$. Nor are there any exclusion restrictions aside from the ordering of the recursive, time-ordered SCM.

2. **Using the backdoor criterion, assess identifiability of $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$.**

   The back-door criterion states that a set of variables $\boldsymbol{Z}$ satisfies it with respect to the variable for the exposure of interest A and the variable for the outcome of interest Y if $\boldsymbol{Z}$ blocks all unblocked back-door (i.e., with an arrow going into A) paths from A to Y without creating any new non-causal associations between A and Y.

   Another way to state it is that $\boldsymbol{Z}$ satisfies the back-door criterion with respect to (A, Y) if

Figure 1: Structural causal model (SCM) $\mathcal{M}^F$ represented in directed acyclic graph (DAG) format.



- $\boldsymbol{Z}$ blocks all paths from A to Y with an arrow into A, and
- No node in $\boldsymbol{Z}$ is a descendant of A.

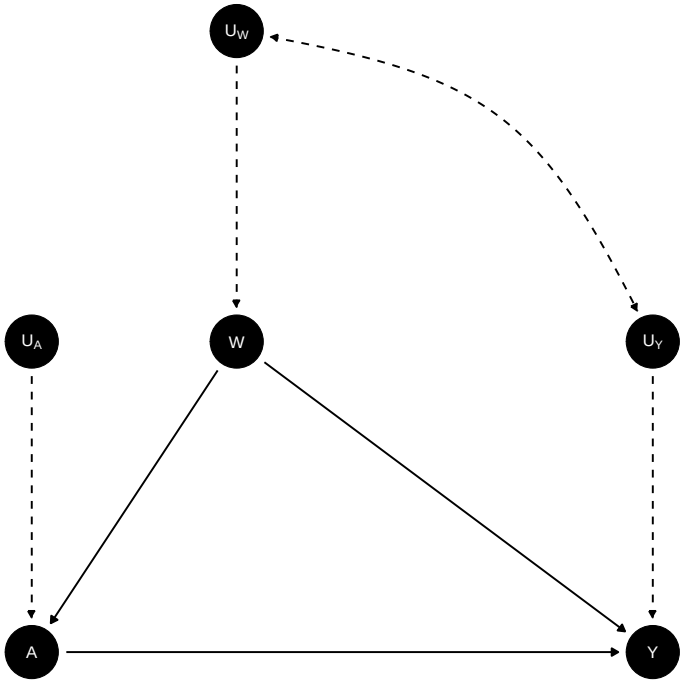
If such a set $\boldsymbol{Z}$ exists, then the back-door criterion holds, and thus we can identify the effect of A on Y, specifically the target causal parameter $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$, as a parameter of the observed data distribution $\Psi(\mathbb{P}_O)$ by the G-computation formula.

Unfortunately, in the SCM $\mathcal{M}^F$, $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$ is not identified, as the only measured variable available for satisfaction of the back-door criterion is $\boldsymbol{W}$ (or the empty set), and neither $\boldsymbol{W}$ nor the empty set meets the back-door criterion due to the remaining open back-door path between A and Y through U_A and U_Y (including the one through U_A, U_W, and U_Y.)

(a) If not identified, under what assumptions would it be?

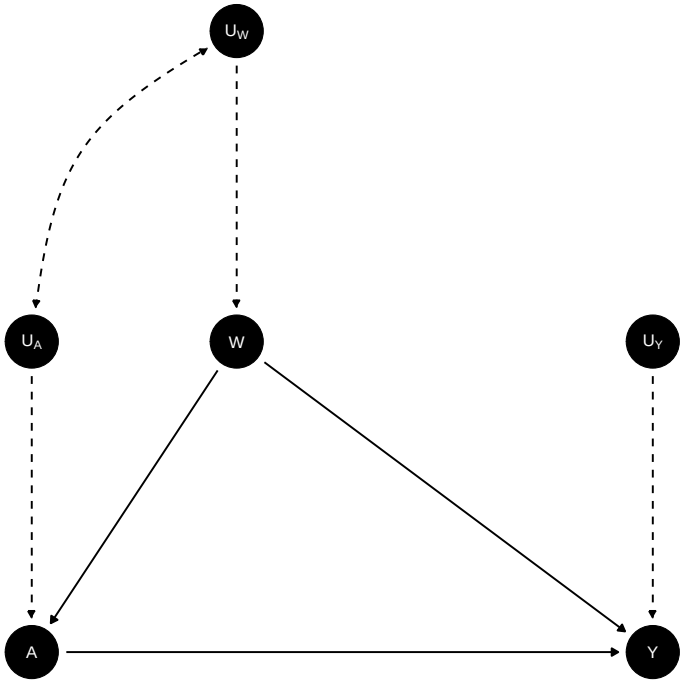
The following DAGs show three possible working structural causal models, $\mathcal{M}_1^{F^*}$, $\mathcal{M}_2^{F^*}$, and $\mathcal{M}_3^{F^*}$, that identify $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$.

Figure 2: Working structural causal model (SCM) $\mathcal{M}_1^{F^*}$ represented in directed acyclic graph (DAG) format.



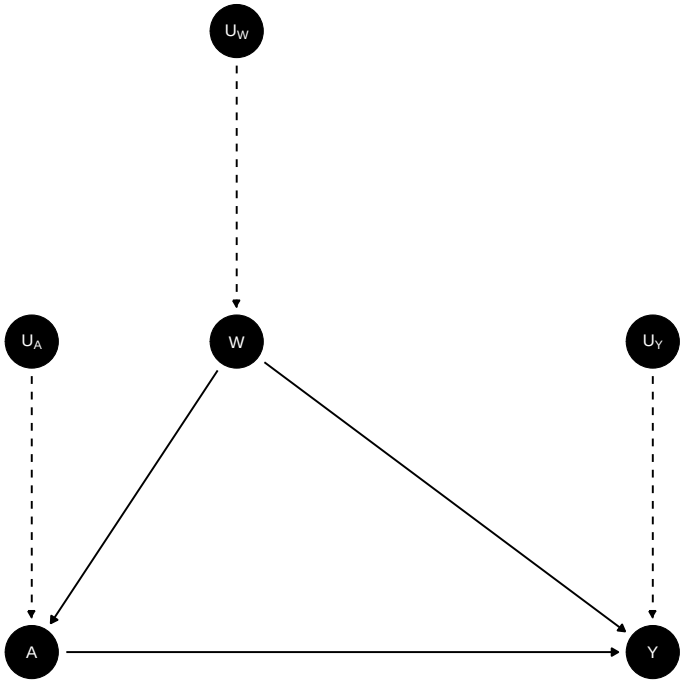
- $\mathcal{M}_1^{F^*}$ assumptions (see Figure 2):
  - U_A is indepdendent of U_Y, i.e., the unmeasured factors U_A that influence energy expenditure A do not affect and are not affected by the unmeasured factors U_Y influencing seven-year survival Y.
  - U_A is indepdendent of U_W, i.e., the unmeasured factors U_A that influence energy expenditure A do not affect and are not affected by the unmeasured factors U_W influencing the covariates like smoking, comorbidities, and body fat W.

Figure 3: Working structural causal model (SCM) $\mathcal{M}_2^{F^*}$ represented in directed acyclic graph (DAG) format.



- $\mathcal{M}_2^{F^*}$ assumptions (see Figure 3):
  - U_A is indepdendent of U_Y, i.e., the unmeasured factors U_A that influence energy expenditure A do not affect and are not affected by the unmeasured factors U_Y influencing seven-year survival Y.
  - U_W is indepdendent of U_Y, i.e., the unmeasured factors U_W that influence the covariates smoking, comorbidities, and body fat, W, do not affect and are not affected by the unmeasured factors U_Y influencing seven-year survival Y.

Figure 4: Working structural causal model (SCM) $\mathcal{M}_3^{F^*}$ represented in directed acyclic graph (DAG) format.



- $\mathcal{M}_3^{F^*}$ assumptions (see Figure 4):
  All unmeasured factors are independent, i.e.,
  - U_A is indepdendent of U_Y, i.e., the unmeasured factors U_A that influence energy expenditure A do not affect and are not affected by the unmeasured factors U_Y influencing seven-year survival Y.
  - U_A is indepdendent of U_W, i.e., the unmeasured factors U_A that influence energy expenditure A do not affect and are not affected by the unmeasured factors U_W influencing the covariates like smoking, comorbidities, and body fat W.
  - U_W is indepdendent of U_Y, i.e., the unmeasured factors U_W that influence the covariates smoking, comorbidities, and body fat, W, do not affect and are not affected by the unmeasured factors U_Y influencing seven-year survival Y.

Are some of these sets of additional assumptions more plausible than others?

The fewer assumptions made, the more plausible the scenario, making working SCMS $\mathcal{M}_1^{F^*}$ and $\mathcal{M}_2^{F^*}$ more plausible than $\mathcal{M}_3^{F^*}$.

All the working SCMs require the assumption that the unmeasured factors influencing energy expenditure be independent of the unmeasured factors influencing seven-year survival, i.e., that U_A be independent of U_Y, or, alternately stated, that all factors influencing both be measured and included in W.

The factors actually included in W, per the original study description from the first discussion assignment, are

- Smoking
- Body fat
- Medical conditions (self-reported and confirmed by treatment and/or medication) including
  - Hypertension
  - Coronary heart disease
  - Myocardial infarction
  - Stroke
  - Lung disease
  - Diabetes
  - Hip or knee osteoarthritis
  - Osteoporosis
  - Cancer
  - Depression

This assumption is likely not justified, as lots of other unmeasured social and environmental factors also likely influence both energy expenditure and survival but were not included in W. For example, neighborhood gun violence could affect both energy expenditure, by reducing both participant walking in the neighborhood, and all-cause seven-year survival directly. Another example might be the absence of air pollution, which would both make a person more likely to exercise and less likely to die from air pollution exposure. Social support, too, might influence energy expenditure (e.g., by inspiring travel to meet friends) and is well documented as a factor that increases survival.

That said, if a trial, for example, were to randomly select participants and assign one group activities that increase energy expenditure and not the other, and were able to influence participants somehow not to alter their activities otherwise, that might make this assumption more plausible. Barring actual random assignment of the exposure, the study could also include neighborhood location data in W to try to control for neighborhood-level effects as well as racial/ethnic data in W as a first step to attempt to control for impacts arising from racism and ethnocentrism that affect both energy expenditure and all-cause seven-year survival.

The working SCMs then also require one of two additional assumptions:

- $\mathcal{M}_1^{F^*}$: The unmeasured factors influencing the covariates listed above neither influence or are influenced by the unmeasured factors influencing the exposure of energy expenditure, i.e., that U_W is independent of U_A, or
- $\mathcal{M}_2^{F^*}$: The unmeasured factors influencing the covariates listed above neither influence or are influenced by the unmeasured factors influencing the outcome of seven-year survival, i.e., that U_W is independent of U_Y.

Unfortunately, neither of these assumptions is justifed either. In $\mathcal{M}_1^{F^*}$, innumerable unmeasured factors likely affect both energy expenditure A and the presence of various covariates in W. For example, an anxiety disorder might increase both energy expenditure and the incidence of various diseases included in W. Household wealth might also affect both energy expenditure, as a person with more might have more ability to join a gym, less need to work two jobs and more time for recreational exercise, etc., and might also have a lower likelihood of the diseases in W due to access to a healthier diet, access to preventive healthcare, etc.

In $\mathcal{M}_2^{F^*}$, various factors likely affect both the values of the covariates in W and seven-year survival. For example, a genetic mutation affecting cell repair might both affect cancer risk and might also affect all-cause seven-year mortality directly. Age might also affect both, even with the existing limitation of the study to participants ages 70 to 79.

Are there additional measurements you could make so that the needed identifiability assumptions are more plausible?

Additional measurements that would make the needed identifiability assumptions more plausible and that would probably be relatively easy to get include neighborhood of residence (which could be linked to area-level factors like air pollution exposure), age, race/ethnicity, and some wealth metric, for the reasons outlined above.

In practice, I would probably use $\mathcal{M}_1^{F^*}$ as the working SCM, which assumes no association between U_A and U_W (nor between U_A and U_Y), and try to measure and include the factors listed in the prior paragraph in W to make both of those assumptions more plausible. If nothing else, I would choose $\mathcal{M}_1^{F^*}$ because a truly randomized trial forces the independence of $U_A \perp\!\!\!\perp U_W$ and $U_A \perp\!\!\!\perp U_Y$, so a later researcher seeking to confirm conclusions from this study using a randomized trial to somehow assign energy expenditure could use a more similar SCM in confirming or challenging my results. Moreover, I can check assumptions that $U_A \perp\!\!\!\perp U_W$ more easily at the beginning of the study, should I gain access to additional detail to move factors from U_W to W, then I can check assumptions that $U_W \perp\!\!\!\perp U_Y$, as I would have access to the exposure data for A, whereas the outcomes included in Y will not have happened yet, and by the time they have, it might be too late to measure some of the factors in U_W that affected them.

(b) What notation do we use to denote the original SCM, augmented with additional assumptions needed for identifiability?

$\mathcal{M}^{F^*}$. add to this

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

Under the working SCM $\mathcal{M}^{F^*}$, the target causal parameter of the observed data distribution, i.e., the statistical estimand, is

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

4. What is the relevant positivity assumption? Are you concerned about violations of the positivity assumption in your study?

The relevant positivity assumption is $\min_{a\in\mathcal{A}} P_O(A=a|W=w) > 0$ for all w for which $P_O(W=w) > 0$, i.e., that there must be a positive probability of each exposure level of A within each possible stratum of each variable in W. Given the sheer number (12) of factors represented by W, even without measuring the additional factors that I think are necessary to include in W, and even if each factor in the existing W is binary, that leaves 2^12 or 4,096 different possible combinations of covariate values in W. Thus the assumption that each combination of covariate values actually present in W holds for at least one treated and one untreated participant seems highly likely to be violated. Thus I am very concerned.

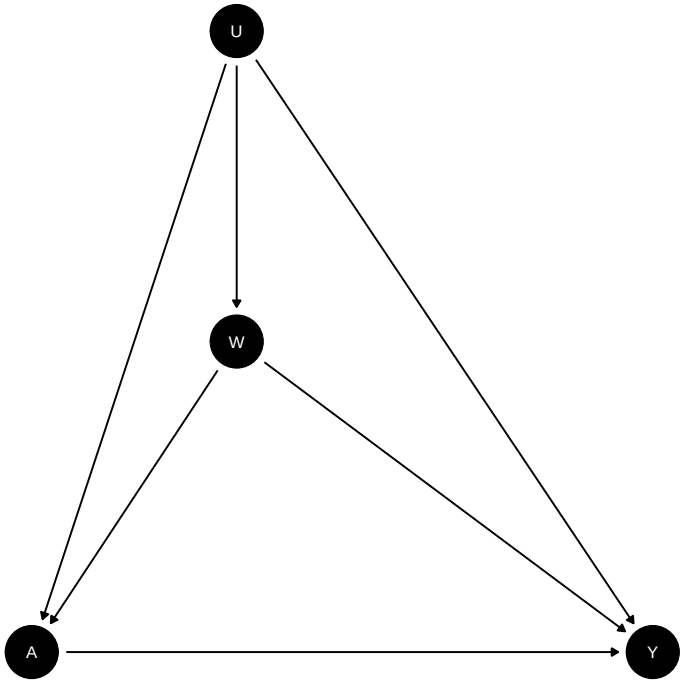
4 Study-specific questions

The investigators assume no unmeasured common causes of (W, A, Y). Is this necessary? Is this sufficient?

I'm not exactly sure if this is saying that the investigators made the weaker assumption that no single unmeasured cause caused all three, or that the investigators made the stronger assumption that none of the three share any unmeasured cause with another of the three, so I'll answer for both possibilities.

The weaker assumption that no single unmeasured cause exists that causes all three factors represented in the variable set (W, A, Y) is necessary but not sufficient. The case in which such an unmeasured cause does exist that affects all three is a subset of the scenario represented in the structural causal model of Figure 1, where all the unmeasured causes are associated with all other unmeasured causes. (Those associations could represent identity, i.e., that a single random variable representing the same cause is an element in all three random variables U_W, U_A, and U_Y). The minimal situation, in which only one unmeasured cause U exists but is a cause of W, A, and Y, we can represent as

Figure 5: Structural causal model where only one unmeasured cause U exists but causes all three of W, A, and Y.



Because U is unmeasured, we cannot block the path through U from A to Y by controlling for it, so we cannot satisfy the backdoor criterion. Thus we cannot use the G-computation formula to identify the effect of energy expenditure A on all-cause seven-year survival Y with the data we have measured. We would need to measure the cause U affecting all three. Thus, the assumption that no such U exist is necessary. That said, though, the nonexistence of such a U is not sufficient. Take this hypothetical case, where a common cause of A and Y exists that is not a cause of W:

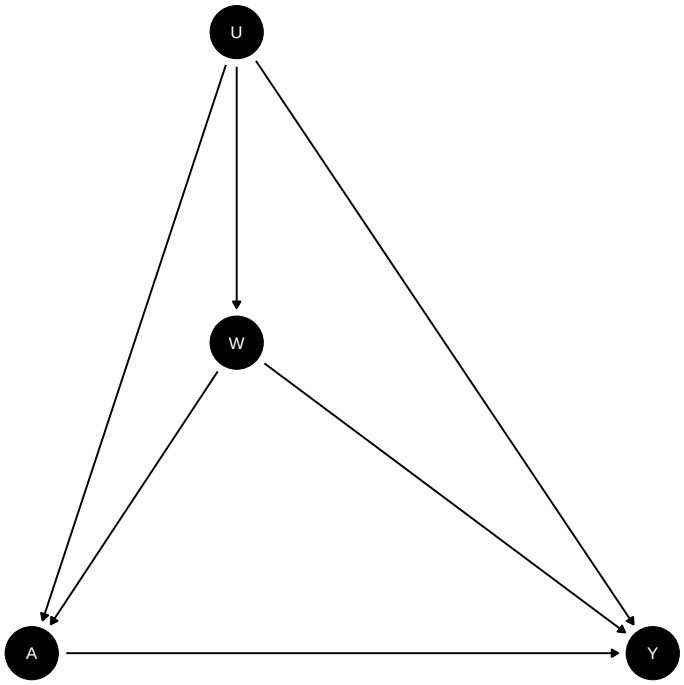
The stronger assumption that none of the three factors represented in the variable set (W, A, Y) shares a common cause with another variable in the set is sufficient but not necessary. This case is represented by $\mathcal{M}_3^{F*}$ in Figure 4. In this case, W blocks all paths with an arrow into A without including any descendants of A, so it satisfies the backdoor criterion, and we can use the G-computation formula to identify the target causal parameter $\Psi^F(P_{U,X})$ as a parameter of the observed data distribution $\Psi(\mathbb{P}_O)$. Thus the assumption is sufficient. The working SCMs $\mathcal{M}_1^{F*}$ and $\mathcal{M}_2^{F*}$ in figures 2 and 3, however, show that it is not necessary, as the target causal parameter $\Psi^F(P_{U,X})$ is identifiable as a parameter of the observed data distribution $\Psi(\mathbb{P}_O)$ despite the potential existence of a common cause of W and Y in $\mathcal{M}_1^{F*}$ (Figure 2) and the potential existence of a common cause of W and A in $\mathcal{M}_2^{F*}$ (Figure 3).

5 Questions for GSIs

Hi, GSIs!
Maya's new slides for Lecture 5 appear to have a blank slide at the end labeled "Positivity for Continuous W" but no actual content on that slide.

Figure 6: Structural causal model where only one unmeasured cause U exists and causes A and Y but not W.

```
## Error in dagify(Y ~ W + A + U, A ~ W + U, W, exposure = "A", outcome = "Y"): object 'W' not found
```



Because U is unmeasured, we cannot block the path through U from A to Y by controlling for it, so we cannot satisfy the backdoor criterion. Thus we cannot use the G-computation formula to identify the effect of energy expenditure A on all-cause seven-year survival Y with the data we have measured. We would need to measure the cause U affecting both A and Y. Thus, the assumption that no U exists that affects all three of W, A, and Y is not sufficient to identify the target casual parameter as a parameter of the observed data distribution, since we've shown by counterexample a situation in which no such U exists but we still cannot satisfy the backdoor criterion or use the G-computation formula to identify the effect of energy expenditure A on all-cause seven-year survival Y with the data we have measured.