Practical 1

Part 1: Causal research questions

Which of the following are causal research questions? Think of what (if any) decision problems underly the question and how a target trial might look like.

When your answer is no, think about if/how you can turn the question into a causal one. (Multiple answers are possible.)

(a) What is currently the average difference in yearly earnings between men and women working as academics in higher education institutions in the UK? *Not a causal question, just descriptive; this is not about 'recommending' sex change to increase earnings, and we would not turn this into a causal question (though sex is randomly allocated at birth and hence a natural experiment).*

(b) What is the probability that a new bank customer (with given age, gender, profession, post code, credit history) will default on a loan? *Not a causal question, just individualised prediction for particular customer; this is not about recommending what to do to change the likelihood of defaulting. The predictors are difficult to intervene on, so that it is hard to turn this into a question of 'what could a customer do / change to be less likely to default'.*

(c) What difference regarding job chances will it make to a long-term unemployed if they do versus do not participate in a new training programme? *This is a causal question: we are targeting the probability of finding a job under a decision to participate in the new programme vs. deciding not to participate. A target trial would randomise unemployed people to the training programme.*

(d) Is an obese 7-year old child at a greater risk of developing asthma than a non-obese 7-year old child? *The wording suggests this is not a causal question, just descriptive: are obese children at a higher risk than non-obese ones. A causal quesion would be: if all obese children were (through what intervention?) not obese, would they be less likely to develop asthma? Or: if an intervention prevents any 7-year old child from obesity, would the incidence of asthma be reduced? Note that there may be a violation of causal consistency here, as it is unclear what intervention would prevent obesity and if this is reflected by those who happen to not be obese in the observational setting.*

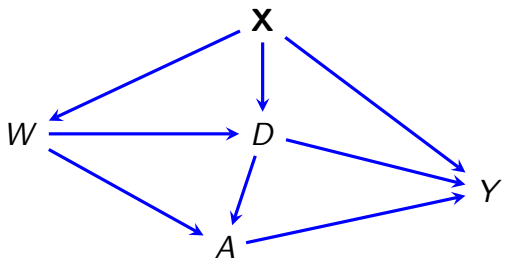
Part 1: Causal notation

Write down in terms of counterfactuals how you would formalise the following causal effect parameters.

(a) The effect of treatment on the treated, i.e. what would be the mean difference in outcome for those who naturally receive (or take) treatment if they were instead prevented from receiving it? $E[Y(1) - Y(0)|A = 1]$.

(b) Assume Y is binary where $Y = 1$ denotes 'diseased'; let $A = 1$ denote the presence of a risk factor for that disease ($A = 0$ is absence). The population attributable fraction is defined as the proportion of diseased that could be prevented if the risk factor could be eliminated from the population. Write this out formally with potential outcomes. $[P(Y = 1) - P(Y(0) = 1)]/P(Y = 1)$.

Part 2: Reasoning with DAGs and SWIGs

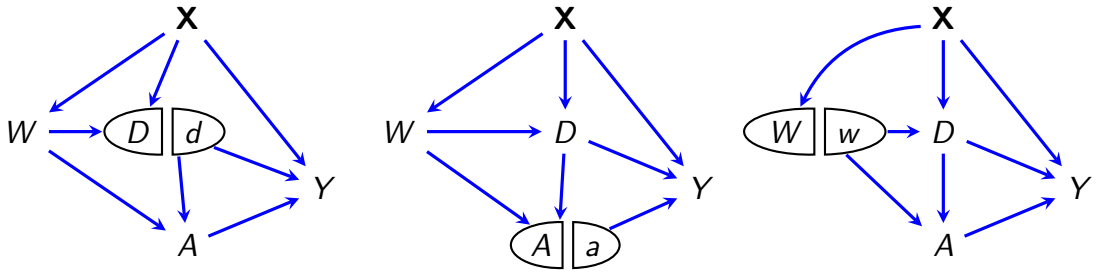


Consider the causal DAG above (Staplin et al, 2016). Here the outcome Y represents end-stage renal disease (ESRD), W is historical smoking, A is current smoking, D is prior disease, and $\mathbf{X}$ is a set of baseline variables.

[Many of the solutions can be found using d-separation. You may wish to try using `DAGitty` for some of these.]

(a) (i) With / (ii) without data on 'early smoking' (W) are there any testable implications of the causal model? *(i) Current smoking (A) should be conditionally independent of the baseline covariates ($\mathbf{X}$), after we condition on earlier smoking and prior disease (W, D). Similarly, ESRD (Y) should be independent of earlier smoking (W) given the other three nodes. (ii) There are no testable implications if W is unobserved.*

(b) We want to identify the causal effect of (i) prior disease (D); (ii) current smoking (A); or (iii) earlier smoking (W) on ESRD (Y). Draw the SWIGs relevant to an intervention on each of these variables.



(c) What should we adjust for and not adjust for, respectively? *(i) We must adjust for $\mathbf{X}$ and W, but cannot adjust for A as it mediates the effect. (ii) We must adjust for prior disease (D), and also either baseline covariates or earlier smoking or both ($\mathbf{X}, W$). (iii) Again we must adjust for $\mathbf{X}$, but cannot adjust for either D or A.*

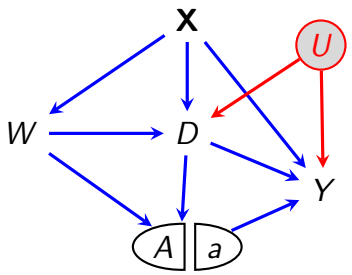
(d) [*] Which sets are 'optimal' to adjust for? *[Hopefully you will be able to do this after the next two lectures.]*

(i) We have no choice here, as there is only one valid adjustment set.
(ii) Most efficient is to adjust for prior disease and baseline covariates only ($D, \mathbf{X}$).
(iii) Again, there is only one valid adjustment set.

(e) Specifically: can the causal effect of current smoking on ESRD be identified without data on 'early smoking'? Convince yourself that your chosen adjustment set blocks all back-door paths. *Yes. As noted above, we can adjust for the other two variables*

and indeed this is the most efficient adjustment set.

(f) Add a node for your SWIG in scenario (ii) that represents a further, unobservable, variable that directly affects prior disease (D) and ESRD (Y). Can the effect of current smoking on ESRD still be identified from the measured data? Explain your conclusion.



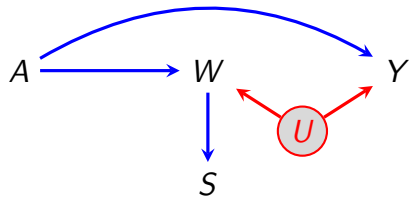
Only if we have data on earlier smoking. We still need to adjust for prior disease, but if we do this and do not adjust for earlier smoking then the path $A \leftarrow W \rightarrow D \leftarrow U \rightarrow Y$ *is opened by the collider at prior disease.*

Part 2: SWIGs for Causal Inference

*[This example is taken from Breskin et al. (*Epidemiology*, 2018).]*

Consider a randomized trial for the effect of a vaccine (A) on a disease (Y). The vaccine causes pain (W), and participants who experience pain are more likely to drop out of the study ($S = 0$), and for these individuals Y is not observed. Patients who have poor underlying health (U) are both more likely to experience the pain and more likely to have the disease.

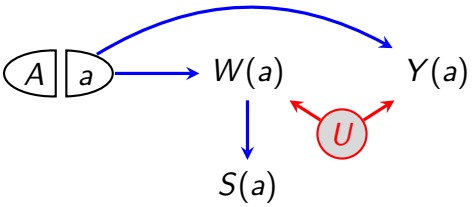
(i) Draw a DAG that represents the causal structure described.



(ii) Write down the distributions that are identifiable from the description of the study.
These are $P(A, W)$ *and* $P(Y \mid A, W, S = 1)$.

(iii) Explain, with reference to d-separation, why we cannot use $\mathbb{E}[Y \mid A = a, S = 1]$ to estimate $\mathbb{E}Y(a)$. *If we condition on* $S = 1$ *this opens up the spurious path* $A \to W \leftarrow U \to Y$.

(iv) Now turn the graph into a SWIG by splitting the node A, and minimally relabelling other vertices.



(v) Using d-separation, show that $S(a) \perp\!\!\!\perp Y(a) \mid W(a)$.

There are three paths, all of which are blocked by the non-collider $W(a)$. *These are* $Y(a) \leftarrow W(a) \to S(a)$, $Y(a) \leftarrow U \to W(a) \to S(a)$ *and* $Y(a) \leftarrow a \to W(a) \to S(a)$.

(vi) Hence, derive an identifying formula for $\mathbb{E}Y(a)$. *[Hint:* $\mathbb{E}Y(a) = \sum_w \mathbb{E}[Y(a) \mid W(a) = w] \cdot P(W(a) = w)$*]*

$$
\begin{aligned}
&\mathbb{E}Y(a) \\
&= \sum_w \mathbb{E}[Y(a) \mid W(a) = w] \cdot P(W(a) = w) && \textit{[by hint]} \\
&= \sum_w \mathbb{E}[Y(a) \mid W(a) = w, S(a) = 1] \cdot P(W(a) = w) && \textit{[by (v)]} \\
&= \sum_w \mathbb{E}[Y(a) \mid W(a) = w, S(a) = 1, A = a] \cdot P(W(a) = w \mid A = a) && \textit{[since } A \perp_d W(a), Y(a)\textit{]} \\
&= \sum_w \mathbb{E}[Y \mid W = w, S = 1, A = a] \cdot P(W = w \mid A = a) && \textit{[consistency]},
\end{aligned}
$$

and note that these factors are identifiable from (ii).