Session 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 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?

b) 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?

c) Is an obese 7-year old child at a greater risk of developing asthma than a non-obese 7-year old child?

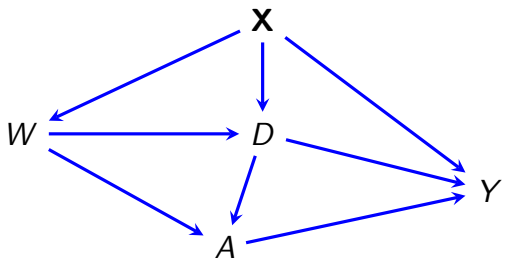
Part 1: Causal notation

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?

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.

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?

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?

d) [*] Which sets are 'optimal' to adjust for? *[You will be able to do this after the next lecture.]*

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.

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.

Part 2: SWIGs for Causal Inference

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.

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

b) Write down the distributions that are identifiable from the description of the study.

c) Explain, with reference to d-separation, why we cannot use $\mathbb{E}[Y \mid A = a, S = 1]$ to estimate $\mathbb{E}Y(a)$.

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

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

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

g) Comment on a condition necessary for the quantity of interest to be identifiable.