Causal Inference from Case-control Studies

The effect of specific interventions can most unambiguously be investigated in randomised controlled trials (RCTs). However, there are many reasons to use observationaldata and in particular case-control studies in practice, such as because the disease under investigation is very rare.

Potential structural bias regarding causal inference:

- Case-control studies are necessarily observational, so **confounding** is likely to be present.
- Case-control studies are retrospective with sampling being conditional on disease status which means there is also a threat of **selection bias**.
- A consequence of the retrospective sampling is that methods which depend on, or are sensitive to, the marginal distribution of the outcome cannot be used without some modification, since the required information is not generally available. This is potentially relevant to methods of adjusting for confounding as well as to the identifiability of typical causal effect measures, such as the average causal effect (ACE). An effect measure that is not sensitive to the retrospective sampling is the odds ratio (OR) but its interpretation can be complicated by its noncollapsibility

Set-ups:

D: a binary outcome, with D=1 denoting the cases and D=0 the controls.

X: exposure. Exposure refers to the specific factor or variable that is hypothesized to influence the outcome. It is the main variable of interest in the study.

C: covariates. Covariates are other variables that might influence the outcome but are not the primary variables of interest.

S: sampling indicator, with S=1 means the unit is present in the sample

to represent the data from a case-control study: $P(X, C, D|S = 1)$, with the sampling proportion always known: $\tilde{p}_s = P(D = 1|S = 1)$.

Directed acyclic graphs (DAGs)

- express the (structural) assumptions we are willing to make about the joint distribution within the population
- represents conditional independence constraints on the joint probability distri- bution over the variables shown as nodes in the graph
- it is the absence of edges that encodes assumptions; the presence of an edge means merely that we allow the possibility of a dependence, not that we claim there necessarily is one
- Some independencies cannot be seen graphically; for instance it is common to match such that $D \perp C|S = 1$, but this is not represented by the graph

FIGURE 6.1
Directed acyclic graphs representing sampling under (a) an unmatched case-control study; (b) a matched case control study with matching variables C.

(a): sampling only depends on the outcome D alone: $S \perp (X, C)|D$

(b): selection bias is ruled out: $S \perp X|(C, D)$ ---- could be the assumption for case-control study: the sampling does not, directly or indirectly, depend on both the exposure and the disease status

do(·)-operator: $P(D|do(X = x)) = P(D_x)$: the distribution of D if X is set to x by an intervention (D_x: the disease status when the exposure or treatment X is set to x by an intervention). This can be estimated from a RCT

Counterfactuals: can not observe D_1^i and D_0^i for a given unit i.

Causal null hypothesis:

$$H_0^{\emptyset} : P(D|do(X = x)) = P(D|do(X = x')), \forall x \neq x'. \tag{1}$$

If the causal null hypothesis is rejected, we say that *X has a causal effect on D*.

Average causal effect (ACE): $E(D|do(X = x)) - E(D|do(X = x'))$

Causal risk difference: ACE when D is binary

Conditional causal null hypothesis (within C = c):

$$H_0^{C=c} : P(D = 1|do(X = x), C = c) = P(D = 1|do(X = x'), C = c), \forall x \neq x'. \tag{2}$$

If the conditional causal null hypothesis is rejected, we say that *X has a causal effect on D within a subgroup* $C = c$.
$H_0^C = \cap_{c \in \mathcal{C}} H_0^{C=c}$

- C needs to be pre-exposure so that C can not be affected by X

  (1) people with C = c are randomised to different values of X, $P(D|C; do(X))$ can be estimated from an RCT

  (2) $P(C|do(X)) = P(C)$

- unfaithfulness: $H_0^C \Rightarrow H_0^{\emptyset}; H_0^{\emptyset} \nRightarrow H_0^C$ there could be zero overal causal effect, but nonzero causal effect within some subgroups defined by C=c.

Reason of randomization can make causal inference ---- Exchangeability of two priori groups due to randomization:

At the time of randomisation, those units with X = 0 are entirely comparable to units with X = 1 regarding the distribution of any measured or unmeasured factors such as age, lifestyle, socioeconomic background.

$\Rightarrow$ any subsequent differences between the groups must be due to the different values of X

$\Rightarrow P(D|do(X)) = P(D|X)$

However, in observational studies exchangeability cannot be guaranteed, as units observed to have X = 0 may well have very different attributes than those observed to have X = 1 $\Rightarrow$ Confounding might exist

Confounding

the variables impact both X and D $\Rightarrow P(D|do(X)) \neq P(D|X)$

Solution for confounding: seek to take suitable covariates into account to compensate for the lack of exchangeability: Conditional exchangeability:

A set C of pre-exposure covariates is sufficient to adjust for confounding with respect to the effect of X on D if

$$P(D = d|C = c, do(X = x)) = P(D = d|C = c, X = x) \tag{3}$$

once we condition on C, observing X = x is as if X had been set to x by an intervention, at least as far as the resulting distribution of D is concerned

Alternative of sufficiency to adjust for confounding: $D_x \perp X|C$,

When C is sufficient to adjust for confounding, $H_0^C \iff D \perp X|C$

This help us to get the intervention distribution by standardisation:

$$P(D = d|do(X = x)) = \sum_c P(D = d|C = c, X = x)P(C = c) \tag{4}$$

How to find sufficient C to adjust for confounding? ---- back-door criterion: C blocks all *back-door* paths from X to D

*__Back-Door Path__: A path from the exposure (X) to the outcome (D) that starts with an arrow pointing into X.

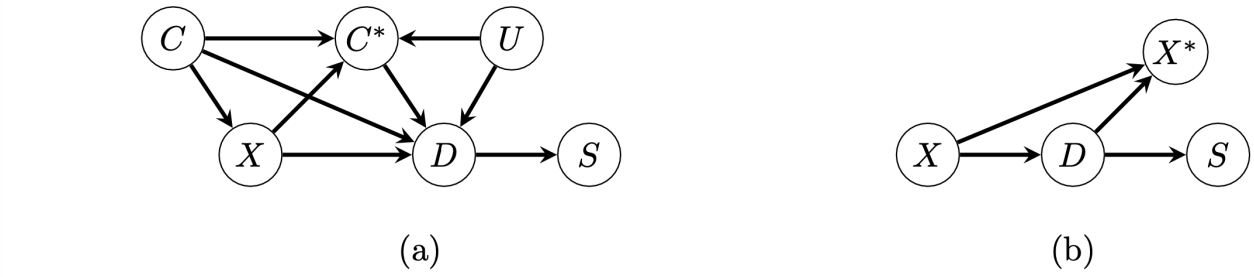
Selection Bias:

Selection bias occurs when there is a systematic difference in how participants are selected or included in the study, leading to a non-representative sample.

two quantities that are not associated in the population may very well be associated in the sample if they both affect the probability of being sampled

Ascertainment Bias

Ascertainment bias occurs when there is a systematic difference in how cases and controls are identified or diagnosed, leading to differential detection or reporting of exposure or outcome between the groups.

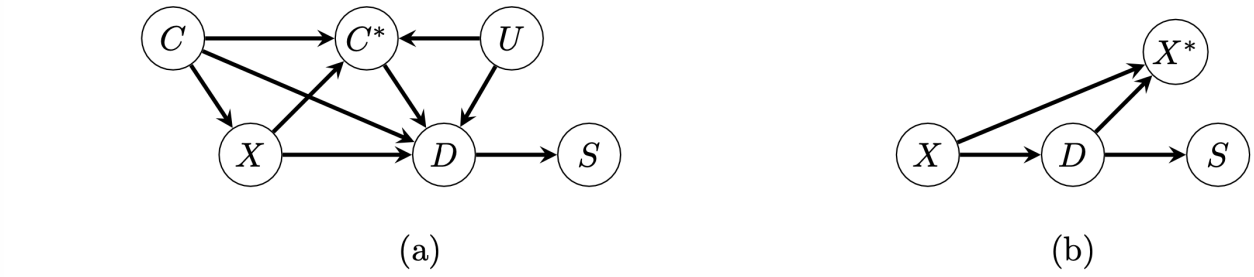


(a) (b)

FIGURE 6.3
Causal graphs representing forms of recall bias. In (a), some measured covariates C^* are actually recalled after treatment; in (b), we only obtain a measure of exposure after observing the outcome.

Recall Bias

Recall bias occurs when participants do not remember past events or exposures accurately, and this inaccuracy differs between cases and controls.



(a) (b)

FIGURE 6.3
Causal graphs representing forms of recall bias. In (a), some measured covariates C^* are actually recalled after treatment; in (b), we only obtain a measure of exposure after observing the outcome.

The key problem of causal inference:

Under what assumptions are certain aspects of the intervention distributions $P(D|do(X))$ or $P(D|C;do(X))$ uniquely cimputable/ **identified** from $P(X, C, D|S = 1)$

Odds ratio

a measure of association; in some assumptions can be used for causal inference

$$OR_{DX} = \frac{P(D=1|X=1)P(D=0|X=0)}{P(D=0|X=1)P(D=1|X=1)} \tag{5}$$

Properties:

1. Symmetric $OR_{D|X} = OR_{X|D}$
2. invariant to changes of marginal distribution of D and X
3. can be used to test independence: $OR_{D|X} = 1 \iff D \perp X$

Conditional odds ratio

$$OR_{DX}(C=c) = \frac{P(D=1|X=1,C=c)P(D=0|X=0,C=c)}{P(D=0|X=1,C=c)P(D=1|X=1,C=c)} \tag{6}$$

Properties:

1. Symmetric $OR_{D|X}(C=c) = OR_{X|D}(C=c)$
2. invariant to changes of marginal distribution of D | C and X | C
3. can be used to test independence: $OR_{D|X}(C=c) = 1 \iff D \perp X|C$

$^*OR_{DX}(C) = \{OR_{DX}(C=c), c \in \mathcal{C}\}$

Collapibility of odds ratios ----- the marginal association is the same as the conditional association

Conditional odds ratio $OR_{D|X}(C_1, C_2)$ is collapsible over C_2 if for $\forall c_2 \neq c_2'$ and all c_1 in the respective domains

$$OR_{DX}(C_1=c_1, C_2=c_2) = OR_{DX}(C_1=c_1, C_2=c_2') = OR_{DX}(C_1=c_1) \tag{7}$$

Conditional causal odds ratio

$$COR_{D|X}(C=c) = \frac{P(D=1|do(X=1),C=c)P(D=0|do(X=0),C=c)}{P(D=0|do(X=1),C=c)P(D=1|do(X=1),C=c)} \tag{8}$$

Properties:

1. $COR_{D|X}(C=c) = 1$: no causal effect of the exposure X on the outcome D
2. not symmetric : if X is causal for D, D is not causal for X : $COR_{D|X}(C=c) \neq 1$ then $COR_{X|D}(C=c) = 1$
3. $COR_{D|X}(C=c) = COR_{D|X}(C=c')$ for all $c \neq c'$ can not deduce that $COR_{D|X}(C) = COR_{D|X}$;
4. $COR_{D|X}(C) \equiv 1 \Rightarrow COR_{D|X} = 1$

To test the causal null hypothesis H_0^C and estimate the CORs we need COR to be able to collapse over S

Propositions

(i) $OR_{DX}(C, S)$ is collapsible over S $\iff$ $S \perp X|(C, D)$.

Further, if C is sufficient to adjust for confounding w.r.t. the effect of X on D, then the conditional OR is equal to the causal OR, i.e., $OR_{DX}(C) = COR_{D|X}(C)$.

(ii) If $S \perp X|(C, D)$, then

$$D \perp X|C \iff D \perp X|(C, S = 1) \tag{9}$$

Further, if C is sufficient to adjust for confounding w.r.t. the effect of X on D, then $D \perp X|(C, S = 1)$ is a test of the causal null hypothesis H_0^C.

(iii) In the particular case of binary D and continuous X, OR can be consistently estimated using logistic regression

For a causal interpretation of the results, need 2 assumptions:

1. the covariates C are sufficient to adjust for confounding
2. $S \perp X|(C, D)$, ruling out selection bias

semi parametric methods for adjusting for confounding

- Regression adjustment
- Standardisation
- Propensity Scores

Instrumental Variables ----- approach for unobserved confounding

"imperfectly mimicking" randomization of X when exposure itself cannot be randomized

A variable Z is an instrumental for the causal effect of an exposure X on an outcome D in the presence of unobserved confounding U under the following three core conditions:

- $Z \perp U$, the instrument is independent of the unobserved confounding;
- $Z \not\perp X$, the instrument and exposure are dependent;
- $D \perp Z|(X, U)$, the outcome is conditionally independent of the instrument.

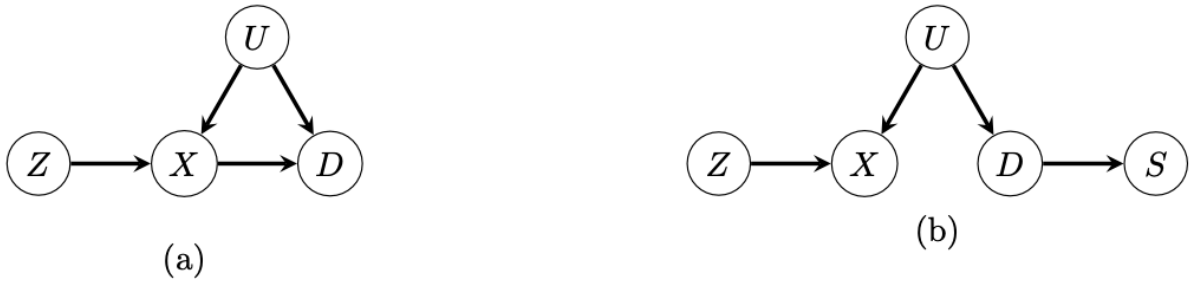


FIGURE 6.7
DAGs representing (a) the IV conditions (prospectively); (b) the IV conditions under the causal null, with a sampling indicator.

*U is the covariates such that if it were observed it would be sufficient to adjust for

We aim to observe data from $P(Z, X, D|S = 1)$,

and $S \perp (X, U, Z)|D \Rightarrow P(Z, X, D|S = 1) = P(Z, X|D)P(D|S = 1)$