Causal Inference: What if

Derek Li

1 A Definition of Causal Effect

1.1 Individual Causal Effects

Definition 1.1 (Causal Effect for an Individual)**.** The treatment A has a causal effect on an individual's outcome Y if $Y^{a=1} \neq Y^{a=0}$ for the individual.

Note 1. The variables $Y^{a=1}$ and $Y^{a=0}$ are referred to as *potential outcomes* or as *counterfactual outcomes*.

Note 2. When i refers to a specific individual, Y_i^a is not a r.v. because we are assuming that individual counterfactual outcomes are *deterministic*.

Property 1.1 (Consistency)**.** $Y = Y^A$ where Y^A denotes the counterfactual Y^a evaluated at the value a corresponding to the individual's observed treatment A.

Note. For each individual, one of the counterfactual outcomes - the one that corresponds to the treatment value that the individual actually received - is actually factual.

1.2 Average Causal Effects

Definition 1.2 ((Average) Causal Effect)**.** An (average) causal effect of treatment A on outcome Y is present if $P(Y^{a=1} = 1) \neq P(Y^{a=0} = 1)$ or $\mathbb{E}[Y^{a=1}] \neq \mathbb{E}[Y^{a=0}]$ in te population of interest.

1. When more than two actions are possible, the particular contrast of interest needs to be specified.

2. Absence of an (average) causal effect does not imply absence of individual effects.

3. If the causal effect in the population is null, we say that the *causal null hypothesis* is true. If there is no causal effect for any individual in the population, i.e., $Y^{a=1} = Y^{a=0}$ for all individuals, we say that the *sharp causal null hypothesis* is true. The sharp causal null hypothesis implies the causal null hypothesis.

1.3 Measures of Causal Effect

1. We can represent the causal null by

(i) Causal risk difference: $P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = 0$

(ii) Risk ratio: $\dfrac{P(Y^{a=1} = 1)}{P(Y^{a=0} = 1)} = 1$

(iii) Odds ratio: $\dfrac{P(Y^{a=1} = 1)/P(Y^{a=1} = 0)}{P(Y^{a=0} = 1)/P(Y^{a=0} = 0)} = 1$

Note. The causal risk difference in the population is a measure of the average individual causal effect. The causal risk ratio in the population is not the average of any individual causal effects.

1.4 Random Variability

1. Sources of random error: Sampling variability and nondeterministic counterfactuals.

1.5 Causation versus Association

1. We can represent independence ($A \perp Y$) by
(i) Associational risk difference: $P(Y = 1|A = 1) - P(Y = 1|A = 0) = 0$
(ii) Risk ratio: $\dfrac{P(Y = 1|A = 1)}{P(Y = 1|A = 0)} = 1$
(iii) Odds ratio: $\dfrac{P(Y = 1|A = 1)/P(Y = 0|A = 1)}{P(Y = 1|A = 0)/P(Y = 0|A = 0)} = 1$

2. Treatment A and outcome Y are dependent or associated when $P(Y = 1|A = 1) \neq P(Y = 1|A = 0)$.

3. Association is defined by a different risk in two disjoint subsets of the population determined by the individuals' actual treatment value ($A = 1$ or $A = 0$), whereas causation is defined by a different risk in the same population under two different treatment values ($a = 1$ or $a = 0$).

2 Randomized Experiments

2.1 Randomization

1. Exchangeability/Exogeneity: When group membership is randomized, which particular group received the treatment is irrelevant for the value of $P(Y = 1|A = 1)$ and $P(Y = 1|A = 0)$.

2. Exchangeability implies $P(Y^a = 1|A = 1) = P(Y^a = 1|A = 0) = P(Y^a = 1)$. It means that the counterfactual outcome and the actual treatment are independent, or $Y^a \perp A, \forall a$.

3. In ideal randomized experiments, ***association is causation***.

4. Full exchangeability: Let $\mathcal{A} = \{a, a', a'', \cdots\}$ denote the set of all treatment values present in the population, and $Y^{\mathcal{A}} = \left\{Y^a, Y^{a'}, Y^{a''}, \cdots\right\}$ the set of all counterfactual outcomes. Randomization makes $Y^{\mathcal{A}} \perp A$.

5. Mean exchangeability: For a dichotomous outcome and treatment, $Y^a \perp A$ can be written as $P(Y^a = 1|A = 1) = P(Y^a = 1|A = 0)$ or $\mathbb{E}[Y^a|A = 1] = \mathbb{E}[Y^a|A = 0]$ for all a. For a continuous outcome, exchangeability $Y^a \perp A$ implies mean exchangeability $\mathbb{E}[Y^a|A = a'] = \mathbb{E}[Y^a]$, but mean exchangeability does not imply exchangeability because distributional parameters other than the mean may not be independent of treatment.

6. Crossover experiment: Let $Y_{i1}^{a_0 a_1}$ be the deterministic counterfactual outcome at $t = 1$ for individual i if treated with a_1 at $t = 1$ and a_0 at $t = 0$. Let $Y_{i0}^{a_0}$ be defined similarly for $t = 0$. The individual causal effect $Y_{it}^{a_t=1} - Y_{it}^{a_t=0}$ can be identified if three conditions hold:
(i) No carryover effect of treatment: $Y_{it=1}^{a_0,a_1} = Y_{it=1}^{a_1}$

Note 1. $Y_{it=1}^{a_1}$ means the deterministic counterfactual outcome at $t = 1$ for i if treated with a_1 at $t = 1$.

Note 2. The last treatment does not affect the next treatment. This condition implies that the outcome $Y_{it}^{a_t}$ has an abrupt onset that completely resolves by the next time period.

(ii) The individual causal effect does not depend on time: $Y_{it}^{a_t=1} - Y_{it}^{a_t=0} = \alpha_i$ for $t = 0, 1$

(iii) The counterfactual outcome under no treatment does not depend on time: $Y_{it}^{a_t=0} = \beta_i$ for $t = 0, 1$

Under these conditions, if the individual is treated at time 1 ($A_{i1} = 1$) but not time 0 ($A_{i0} = 0$), then by consistency, $Y_{i1} - Y_{i0}$ is the individual causal effect because

$$Y_{i1} - Y_{i0} = Y_{i1}^{a_1=1} - Y_{i0}^{a_0=0} = Y_{i1}^{a_1=1} - Y_{i1}^{a_1=0} + Y_{i1}^{a_1=0} - Y_{i0}^{a_0=0} = \alpha_i + \beta_i - \beta_i = \alpha_i$$

Similarly if $A_{i1} = 0$ and $A_{i0} = 1, Y_{i0} - Y_{i1} = \alpha_i$ is the individual level causal effect.

2.2 Conditional Randomization

1. A marginally randomized experiment is expected to result in exchangeability of the treated and the untreated; a conditionally randomized experiment will not generally result in exchangeability of the treated and the untreated because, by design, each group may have a different proportion of individuals.

2. Conditional randomization does not guarantee unconditional/marginal exchangeability $Y^a \perp A$, but it guarantees conditional exchangeability $Y^a \perp A|L$ within levels of L.

3. Randomization produces either marginal exchangeability or conditional exchangeability.

4. Two methods to compute causal risk ratio in a conditionally randomized experiment:

(1) Stratification: The average causal effect in each of subsets or strata of the population. Because association is causation within each subset, the stratum-specific causal risk ratio among people in critical condition is equal to the stratum-specific associational risk ratio among people in critical condition

$$\frac{P(Y^{a=1} = 1|L = 1)}{P(Y^{a=0} = 1|L = 1)} = \frac{P(Y = 1|L = 1, A = 1)}{P(Y = 1|L = 1, A = 0)}$$

and analogously for $L = 0$.

Note. If the stratum-specific causal risk ration in the subset $L = 1$ differs from the causal risk ratio in $L = 0$, the effect of treatment is modified by L, or there is *effect modification* by L.

(2) We can compute the average causal effect $\frac{P(Y^{a=1} = 1)}{P(Y^{a=0} = 1)}$ in the entire population.

2.3 Standardization

1. The marginal counterfactual risk $P(Y^a = 1)$ is the weighted average of the stratum-specific risks $P(Y^a = 1|L = 0)$ and $P(Y^a = 1|L = 1)$ with weights equal to the proportion of individuals in the population, i.e.,

$$P(Y^a = 1) = \sum_l P(Y^a = 1|L = l)P(L = l)$$

By conditional exchangeability,

$$P(Y^a = 1) = \sum_l P(Y = 1|L = l, A = a)P(L = l)$$

where a counterfactual quantity can be expressed as function of the distribution of the observed data so that the counterfactual quantity is *identifiable*. This method is called *standardization*.

2.4 Inverse Probability Weighting

1. Inverse probability weight:

$$W^A = \frac{1}{f(A|L)}$$

Note. $f(A|L)$ is the probability density function. For discrete variables A and L, $f(a|l)$ is the conditional probability $P(A = a|L = l)$.

2. Adjustment for L: Both standardization and IP weighting simulate what would have been observed if the variables L had not been used to decide the probability of treatment.

Property 2.1 (Equivalence of IP Weighting and Standardization)**.** Under positivity, IP weighted mean equals to standardized mean.

Proof. Assume A is discrete with finite number of values and that $f(a|l)$ is positive for all l s.t. $P(L = l)$ is nonzero. The standardized mean for treatment level a is defined as

$$\sum_l \mathbb{E}[Y|A = a, L = l]P(L = l)$$

and the IP weighted mean of Y for treatment level a is defined as

$$\mathbb{E}\left[\frac{I(A = a)Y}{f(A|L)}\right]$$

By definition,

$$\begin{aligned}
\mathbb{E}\left[\frac{I(A = a)Y}{f(A|L)}\right] &= \sum_l \sum_{a'} \sum_y \frac{I(a' = a)y}{f(a|l)} P(Y = y|A = a', L = l)P(A = a'|L = l)P(L = l) \\
&= \sum_l \sum_y \frac{y}{f(a|l)} P(Y = y|A = a, L = l)P(A = a|L = l)P(L = l) \\
&= \sum_l \frac{1}{f(a|l)} \mathbb{E}[Y|A = a, L = l]f(a|l)P(L = l) \\
&= \sum_l \mathbb{E}[Y|A = a, L = l]P(L = l)
\end{aligned}$$

□

Property 2.2. Assume conditional exchangeability, then both the IP weighted and the standardized are equal to the counterfactual mean $\mathbb{E}[Y^a]$.

3 Observational Studies

3.1 Identifiability Conditions

1. An observational study can be conceptualized as a conditionally randomized experiment if the following conditions hold:

(1) Consistency: The values of treatment under comparison correspond to well-defined interventions that correspond to the versions of treatment in the data.

Note. Ill-defined treatments complicate the interpretation of causal effect estimates, but so do sufficiently well-defined treatments that are absent in the data.

(2) Exchangeability: The conditional probability of receiving every value of treatment, though not decided by the investigators, depends only on measured covariates L.

Note. Conditional exchangeability $Y^a \perp A|L$ will not hold if there exist unmeasured independent predictors U of the outcome s.t. the probability of receiving treatment A depends on U within strata of L. Even if it held, we cannot empirically verity it. Hence, when analyzing an observational study under conditional exchangeability, our expert knowledge guides us correctly to collect enough data so that the assumption is at least approximately true.

(3) Positivity: The probability of receiving every value of treatment conditional on L is greater than zero.

Note 1. There is positivity if $P(A = a|L = l) > 0$ for all l with $P(L = l) \neq 0$ in the population of interest.

Note 2. Positivity is only required for the variables L that are required for exchangeability.

Note 3. The positivity condition is sometimes referred to as the experimental treatment assumption.

2. Causal inference from observational data requires two elements: data and identifiability conditions.

3.2 Target Trial

1. If the emulation is successful, there is no difference between the observational estimates and the numerical results that the target trial would have yielded.

2. Excess fraction: The excess fraction or attributable fraction is defined as

$$\frac{P(Y = 1) - P(Y^{a=0} = 1)}{P(Y = 1)}$$

4 Graphical Representation of Causal Effects

4.1 Causal Diagrams

1. The presence of an arrow pointing from a variable V to another variable W indicates that there is a direct causal effect for at least one individual. The lack of an arrow means that V has no direct causal effect on W for any individual in the population.

2. Directed acyclic graph/DAG: We define a DAG G to be a graph whose nodes are r.v.s. $V = (V_1, \cdots, V_M)$ with directed edges and no directed cycles. We denote the parent of V_m as PA_m, i.e., the set of nodes from which there is a direct arrow into V_m. V_m is a descendant of V_j (V_j is an ancestor of V_m) if there is a sequence of nodes connected by edges between V_j and V_m s.t. following the direction indicated by the arrows, one can reach V_m by starting at V_j. We define the distribution of V to be Markov w.r.t. G if for each j, V_j is independent of its non-descendants conditional on its parents.

3. Causal DAG: A causal DAG is a DAG in which

(1) the lack of an arrow from V_j to V_m (i.e., V_j is not a parent of V_m) can be interpreted as the absence of a direct causal effect of V_j on V_m relative to the other variables on the graph;

(2) all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph;

(3) any variable is a cause of its descendants.

Causal DAG is of no practical use unless we make an assumption linking the causal structure represented by the DAG to the data obtained in a study - conditional on its direct causes, V_j is independent of any variable for which it is not a cause, i.e., conditional on its parents, V_j is independent of its non-descendants, i.e., the density $f(V)$ of V in DAG G satisfies the Markov factorization

$$f(v) = \prod_{j=1}^{M} f(v_j|pa_j)$$

Rule 4.1 (D-Separation)**.** (1) If there are no variables being conditioned on, a path is blocked iff two arrowheads on the path collide at some variable on the path: $L \to A \to Y$ is open, whereas $A \to Y \leftarrow L$ is blocked and we call Y a collider.
(2) Any path that contains a non-collider that has been conditioned on is blocked.
(3) A collider that has been conditioned on does not block a path.
(4) A collider that has a descendant that has been conditioned on does not block a path.

4. Faithfulness: For three disjoint sets A, B, C on a causal DAG, where C may be the empty set, A independent of B given C implies A is d-separated from B given C. When the causal diagram makes us expect a non-null association that does not actually exist in the data, the joint distribution of the data is not faithful to the causal DAG.

4.2 Positivity and Consistency in Causal Diagrams

1. Positivity is translated into graph as the condition that the arrows from : to A are not deterministic. Consistency is embedded in the notation because we only consider treatment nodes with relatively well-defined interventions.

2. Positivity is concerned with arrows into the treatment nodes, and consistency is concerned with arrows leaving the treatment nodes.

4.3 A Structural Classification of Bias

Definition 4.1. There is ***systematic bias*** when the data are insufficient to identify or compute the causal effect even with an infinite sample size.

Note 1. We refer to systematic bias as any structural association between treatment and outcome that does not arise from the causal effect of treatment on outcome in the population of interest.

Note 2. A crucial source of bias is the lack of exchangeability between the treated and the untreated.

1. For the average causal effect in the entire population, there is unconditional bias when

$$P(Y^{a=1} = 1) - P(Y^{a=0} = 1) \neq P(Y = 1|A = 1) - P(Y = 1|A = 0)$$

which is the case when unconditional exchangeability $Y^a \perp A$ does not hold.

Note. Absence of bias implies that the association measure in the population is a consistent estimate of the corresponding effect measure in the population.

2. Bias under the null: Lack of exchangeability results in bias when the null hypothesis of no causal effect of treatment on the outcome holds, i.e., even if the treatment had no causal effect on the outcome, treatment and outcome would be associated in the data.

Note. Any causal structure that results in bias under the null will cause bias under the alternative; the converse is not true.

3. For the average causal effects within levels of L, there is conditional bias when

$$P(Y^{a=1} = 1|L = l) - P(Y^{a=0} = 1|L = l) \neq P(Y = 1|L = l, A = 1) - P(Y = 1|L = l, A = 0)$$

for at least one stratum l, which is the case when conditional exchangeability $Y^a \perp A|L = l$ does not hold for all a and l.

4. Lack of exchangeability can result from:

(1) Common causes/Confound: When the treatment and outcome share a common cause, the association measure generally differs from the effect measure.

(2) Conditioning on common effects/Selection bias.

5 Confounding

5.1 Confounding and Exchangeability

Definition 5.1. A set of covariates L satisfies the ***backdoor criterion*** if all backdoor paths between treatment A and outcome Y are blocked by conditioning on L and L contains no variables that are descendants of A.

Theorem 5.1. Assume an FFRCISTG model and faithfulness, conditional exchangeability $Y^a \perp A|L$ holds iff L satisfies the backdoor criterion.

1. Two settings in which the backdoor criterion is satisfied are:

(1) No common causes of treatment and outcome: No backdoor paths that need to be blocked, then the set of variables that satisfies the backdoor criterion is the empty set and there is no confounding.

(2) Common causes of treatment and outcome but a subset L of measured non-descendants of A suffices to block all backdoor paths.

2. The backdoor criterion does not answer questions regrading the magnitude or direction of confounding.

Note 1. We can predict the direction of confounding bias by using signed causal diagrams. But the rule may fail in more complex causal diagrams or when the variables are non dichotomous.

Note 2. For discrete confounders, the magnitude of the bias depends on prevalence of the confounder. If the confounders are unknown, we can only guess what the magnitude of the bias is - sensitivity analyses (i.e., repeating the analyses under several assumptions regarding the magnitude of the bias), which may help quantify the maximum bias that is reasonably expected.

5.2 Confounding and the Backdoor Criterion

1. Confounding: The presence of common causes of treatment and outcome creates an open backdoor path.

2. Selection bias: Conditioning on a common effect may open a previously blocked backdoor path.

5.3 Confounding and Confounders

1. Associational or statistical criteria are insufficient to characterize confounding. Change in estimates may occur for reasons other than confounding, including selection bias when adjusting for non-confounders and the use of noncollapsible effect measures. In contrast, we first identify the sources of confounding and then identify a sufficient set of adjustment variables - ***structural approach***.

Note. In structural approach, a particular variable in a sufficient set depends on the variables already included in the set. Given a causal DAG, confounding is an absolute concept whereas confounder is a relative one.

2. Surrogate confounders: We refer to variables that can be used to reduce confounding bias even though they are not on a backdoor path (and so could never completely eliminate confounding) as ***surrogate confounders***. A possible strategy to handle confounding is to measure as many surrogate confounders as possible and adjust for all of them.

5.4 Single-World Intervention Graph (SWIG)

1. A SWIG depicts the variables and causal relations that would be observed in a hypothetical world in which all individuals received treatment level a.

Note. A SWIG is a graph that represents a counterfactual world created by a single intervention; while the variables on a standard causal diagram represent the actual world.

2. Any variable that is a non-descendant of counterfactual variable need not be labeled as a counterfactual, because under the faithfulness assumption, treatment has no causal effect on its non-descendants for any individual.

Theorem 5.2. On the SWIG, Y^a is d-separated from A given L iff L is a non-descendant of A that blocks all backdoor paths from A to Y.

5.5 Confounding Adjustment

1. Methods that adjust for confounders L can be classified into two broad categories:

(1) G-methods: Standardization, IP weighting, and g-estimation. These methods exploit conditional exchangeability given L to estimate the causal effect of A on Y in the entire population or in any subset of the population.

(2) Stratification-based methods: Stratification (including restriction) and matching. These methods exploit conditional exchangeability given L to estimate the association between A and Y in subsets defined by L.

Note 1. In settings with time-varying treatments, and therefore time-varying confounders, g-methods are the methods of choice to adjust for confounding because stratification-based methods may result in selection bias.

Note 2. All the methods above require conditional exchangeability given L. (*But confounding can sometimes be handled by methods that do not require conditional exchangeability, e.g., difference-in-difference, instrumental variable estimation, the front door criterion, and others. These methods require alternative assumptions that are also unverifiable.*)

2. Difference-indifference and negative outcome control: Suppose that for each individual in the population, we have measured the value of the outcome right before treatment was available in the population.

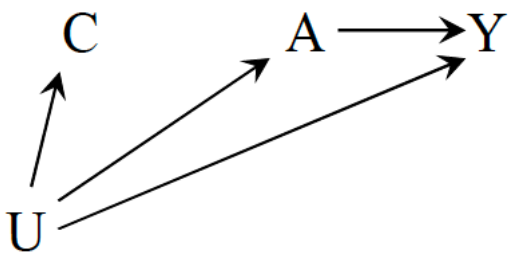


Figure 5.1: A - Aspirin, Y - Blood pressure, U - Unmeasured common causes of A and Y as history of heart disease, C - Pre-treatment outcome/Negative outcome control

As shown in Figure 5.1, though the causal effect of A on C is zero, $\mathbb{E}[C|A=1] - \mathbb{E}[C|A=0] \neq 0$ because of confounding by U. Under the assumption of additive equi-confounding

$$\mathbb{E}[Y^0|A=1] - \mathbb{E}[Y^0|A=0] = \mathbb{E}[C|A=1] - \mathbb{E}[C|A=0]$$

the effect is

$$\mathbb{E}[Y^1 - Y^0|A=1] = (\mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]) - (\mathbb{E}[C|A=1] - \mathbb{E}[C|A=0])$$

3. The front door criterion: As shown in Figure 5.2, we cannot directly use standardization nor IP weighting to compute the counterfactual risks because U is available.

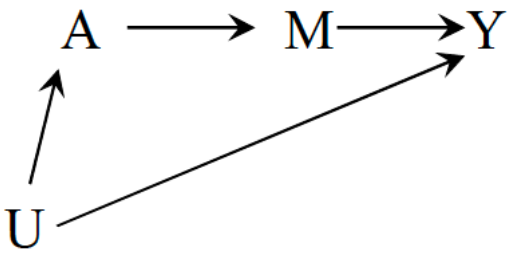


Figure 5.2: A and Y share an unmeasured cause U, M fully mediates the effect of A on Y, and M shares no unmeasured causes with A or Y.

We can compute $P(Y^a = 1)$ by front door formula

$$\sum_m P(M=m|A=a) \sum_{a'} P(Y=1|M=m, A=a')P(A=a')$$

6 Selection Bias

6.1 Adjustment for Selection Bias

1. We can correct selection bias by IP weighting, which is based on assigning a weight W^C to each selected individual ($C=0$) so that she accounts in the analysis not only for herself, but also for those with the same values of L and A, who were not selected ($C=1$). The IP weight W^C is the inverse of the probability of her selection $P(C=0|L,A)$.

2. The association measure in the pseudo-population (with IP weighting) equals the effect measure in the original population if the following three identifiability conditions are met:

(1) The average outcome in the uncensored individuals must equal the unobserved average outcome in the censored individuals with the same values of A and L.

(2) All conditional probabilities of being uncensored given the variables in L must be greater than zero.

Note. The positivity condition is not required for the probability of being censored ($C = 1$).

(3) Consistency (including well-defined interventions).

Note. The effect measure may be relatively well defined when censoring is the result of loss to follow up or non-response, but not when censoring is defined as the occurrence of a competing event.

3. Stratification could yield unbiased conditional effect measures within levels of L if conditioning on L is sufficient to block the backdoor path.

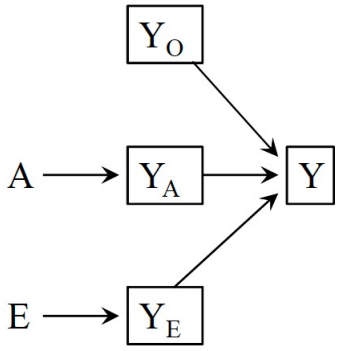
6.2 Selection without Bias

1. Conditioning on a collider induces an association between its causes, but this association could be restricted to certain levels of the common effect, i.e., it is possible that selection on a common effect does not result in selection bias when the analysis is restricted to a single level of the common effect.

2. Multiplicative Survival Model: When the conditional probability of survival given A and $E, P(Y = 0|E = e, A = a) = g(e)h(a)$, we say that a multiplicative survival model holds.

Note 1. Equivalently, the survival ratio $\dfrac{P(Y = 0|E = e, A = a)}{P(Y = 0|E = e, A = 0)}$ does not depend on e and is equal to $h(a)$.

Note 2. The data follow a multiplicative survival model when there is no interaction between A and E on the multiplicative scale as shown in Figure 6.2.



Note 3. If $P(Y = 0|E = e, A = a) = g(e)h(a)$, then $P(Y = 1|E = e, A = a) = 1 - g(e)h(a)$ does not follow a multiplicative mortality model, i.e., when A and E are conditionally independent given $Y = 0$, they will be conditionally dependent given $Y = 1$.