Introduction to Causal Inference

UC DAVIS HEALTH

CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Ezra Morrison, Ph.D.

Assistant Professor of Biostatistics

UC Davis School of Medicine

The UC Davis CTSC receives support from the NIH National Center for Advancing Translational Sciences (award TR001860).

Learning Objectives

1. Define causes and effects
2. Understand how causal inference is used in medical research
3. Define confounding and understand how it makes causal inference difficult
4. Understand how to select study design and analysis methods to answer causal questions

Potential outcomes

- A potential outcome is the outcome that an individual would experience if we intervene to give them a particular treatment or exposure.

- Denoted $Y(x)$; may or may not be the outcome that actually occurs, Y

Example: phototherapy for neonatal jaundice

- Example: treating neonatal jaundice (excess bilirubin) with light exposure ("phototherapy")
  - Outcome (Y): 1 = condition worsens within 48 hours; 0 = not.
  - Treatment (X): 1 = phototherapy; 0 = watchful waiting

    $Y(1)$: if we choose phototherapy, will the jaundice worsen?

    $Y(0)$: if we choose watchful waiting, will the jaundice worsen?

- Data set: 20,731 newborns at 12 NorCal Kaiser hospitals between 1995-2004, with bilirubin levels within 3mg/dL of the guideline threshold for phototherapy
  - (Newman et al, Pediatrics 2009; https://doi.org/10.1542/peds.2008-1635)
- Analysis: Vittinghoff et al, *Regression Methods in Biostatistics* 2e, 2012, Springer
  - https://link.springer.com/book/10.1007/978-1-4614-1353-0

Defining causes

- "x causes y" if:
  - y would occur if we did x, and
  - y would not occur if we did some alternative to x
- If a given infant would recover with phototherapy and not with watchful waiting, then phototherapy causes recovery for that infant.
- $Y(x) = y$ and $Y(x') \neq y$ for some $x' \neq x$
- Necessary cause: y would not occur for any alternative to x.
- Sufficient cause: y would occur if we did x, no matter what else we also did.

Defining effects

- The effect of an intervention on an individual is a comparison between the potential outcomes for that intervention and some alternative: $Y(1)$ versus $Y(0)$.
  - Difference in potential outcomes: $Y(1) - Y(0)$
  - Potential outcomes ratio: $\frac{Y(1)}{Y(0)}$
  - Relative difference in potential outcomes: $\frac{Y(1)-Y(0)}{Y(0)}$

Average effects

- $\mathrm{E}[Y(x)]$: Average potential outcome of treatment x for a population of individuals
- $\mathrm{E}[Y(x) - Y(x')]$: Average Treatment Effect (ATE) or Average Causal Effect (ACE)
- $\mathrm{E}[Y(x)|Z = z]$: Average potential outcome of treatment x in subpopulation $Z = z$
- $\mathrm{E}[Y(1) - Y(0)|X = 1]$: “Average Treatment effect among the Treated” (ATT)

- For binary outcomes with $Y = 1$ denoting the adverse event:
  - Potential risk: $P(Y(x) = 1) = \mathrm{E}[Y(x)]$
  - Causal risk difference: $P(Y(x) = 1) - P(Y(x') = 1) = \mathrm{E}[Y(x) - Y(x')]$
  - Causal risk ratio: $P(Y(x) = 1) \,/\, P(Y(x^*) = 1)$
  - Causal odds ratio:

$$\frac{P(Y(x) = 1)/P(Y(x) = 0)}{P(Y(x') = 1)/P(Y(x') = 0)}$$

Calculating effects

- Suppose we have data on 10 individuals (e.g., newborns with jaundice)

- We would like to estimate the average potential outcomes and average causal effect:
  - $\hat{E}[Y(1)] = \frac{1}{n}\left(Y_1(1) + Y_2(1) + \cdots + Y_{10}(1)\right)$
  - $\hat{E}[Y(1) - Y(0)] = \hat{E}[Y(1)] - \hat{E}[Y(0)]$

- What do we know about $Y(1)$ and $Y(0)$?

X	Y	$Y(1)$	$Y(0)$
0	1	?	?
0	1	?	?
0	0	?	?
0	1	?	?
0	1	?	?
1	0	?	?
1	1	?	?
1	1	?	?
1	0	?	?
1	1	?	?

Basic assumptions for causal inference

- Q1: Are the observed treatments the same as the potential interventions we are interested in?
  - How long is phototherapy applied?
  - How bright is the light?
  - **"Consistency assumption":** If $X = x$, then $Y(x) = Y$

- Q2: does treating one individual affect any other individuals?
  - Vaccinating one individual can protect others
  - Educating one individual can affect others
  - **"Non-interference assumption"**

- Consistency + Non-interference = **"Stable Treatment Value Assumption" (SUTVA)**

Basic assumptions for causal inference

- If we assume consistency and non-interference, we can fill in half of the potential outcomes:

X	Y	$Y(1)$	$Y(0)$
0	1	?	?
0	1	?	?
0	0	?	?
0	1	?	?
0	1	?	?
1	0	?	?
1	1	?	?
1	1	?	?
1	0	?	?
1	1	?	?

Basic assumptions for causal inference

- If we assume consistency and non-interference, we can fill in half of the potential outcomes:

- For $X = 0$, $Y(0) = Y$

X	Y	$Y(1)$	$Y(0)$
0	1	?	1
0	1	?	1
0	0	?	0
0	1	?	1
0	1	?	1
1	0	?	?
1	1	?	?
1	1	?	?
1	0	?	?
1	1	?	?

Basic assumptions for causal inference

- If we assume consistency and non-interference, we can fill in half of the potential outcomes:
- For $X = 0$, $Y(0) = Y$
- For $X = 1$, $Y(1) = Y$

X	Y	$Y(1)$	$Y(0)$
0	1	?	1
0	1	?	1
0	0	?	0
0	1	?	1
0	1	?	1
1	0	0	?
1	1	1	?
1	1	1	?
1	0	0	?
1	1	1	?

The Fundamental Problem of Causal Inference

- Even assuming consistency and non-interference:
  - We are still missing half of the potential outcomes
  - No rows are complete

- If we want to estimate average potential outcomes and risk differences, we need to decide what to do about the missing potential outcomes.

X	Y	$Y(1)$	$Y(0)$
0	1	?	1
0	1	?	1
0	0	?	0
0	1	?	1
0	1	?	1
1	0	0	?
1	1	1	?
1	1	1	?
1	0	0	?
1	1	1	?

Analysis 1: Assume treatment is randomized

- We could assume observed treatments are completely random, or at least, assume that the observed treatments are **independent** of the potential outcomes (i.e., $Y(x) \perp\!\!\!\perp X$) (an independence assumption). Then:
- $\mathrm{E}[\mathrm{Y}(1)|X=0] = \mathrm{E}[\mathrm{Y}(1)] = \mathrm{E}[\mathrm{Y}(1)|X=1] = \mathrm{E}[Y|X=1]$
- $\hat{\mathrm{E}}[Y|X=1] = \frac{1}{5}(0+1+0+1+1) = \frac{3}{5}$
- $\hat{\mathrm{E}}[Y(1)] = \frac{1}{10}\left[\left(5 \times \frac{3}{5}\right) + 3\right] = \frac{3}{5} = 60\%$

X	Y	$Y(1)$	$Y(0)$
0	1	3/5	1
0	1	3/5	1
0	0	3/5	0
0	1	3/5	1
0	1	3/5	1
1	0	0	?
1	1	1	?
1	1	1	?
1	0	0	?
1	1	1	?

Analysis 1: Assume treatment is randomized

- Similarly:
  $$E[Y(0)|X=1] = E[Y(0)] = E[Y(0)|X=0] = E[Y|X=0]$$
- $\hat{E}[Y|X=0] = \frac{1}{5}(1+1+0+1+1) = \frac{4}{5}$
- $\hat{E}[Y(0)] = \frac{1}{10}\left[4 + \left(5 \times \frac{4}{5}\right)\right] = \frac{4}{5} = 80\%$
- $\hat{E}[Y(1) - Y(0)] = \hat{E}[Y(1)] - \hat{E}[Y(0)]$
  $$= \frac{3}{5} - \frac{4}{5} = -\frac{1}{5} = -20\%$$
- Given our assumptions, we would estimate that treatment 1 (phototherapy) reduces the risk of worsening jaundice by 20 percentage points.

X	Y	$Y(1)$	$Y(0)$
0	1	3/5	1
0	1	3/5	1
0	0	3/5	0
0	1	3/5	1
0	1	3/5	1
1	0	0	4/5
1	1	1	4/5
1	1	1	4/5
1	0	0	4/5
1	1	1	4/5

Analyzing the phototherapy data, assuming observed treatment is completely random

	Condition Worsened ($Y = 1$)	Condition Stabilized or Improving ($Y = 0$)	All
Phototherapy ($X = 1$)	15 (0.3%)	4569 (99.7%)	4584 (22%)
Watchful Waiting ($X = 0$)	113 (0.7%)	16,034 (99.3%)	16,147 (78%)
All	128 (0.6%)	20,603 (99.4%)	20,731

Under our assumptions (consistency, no interference, observed treatment is independent of the potential outcomes):

- Estimated Causal Risk (of jaundice worsening if we choose phototherapy) = $\widehat{\mathrm{P}}[Y(1) = 1] = 0.3\%$
- Estimated Causal Risk (of jaundice worsening if we choose waiting) = $\widehat{\mathrm{P}}[Y(0) = 1] = 0.7\%$
- Estimated Causal Risk Difference $= 0.3\% - 0.7\% = -0.4\%$

That is, we estimate that giving phototherapy to all cases would reduce the event rate by 0.4%

Not-completely-random treatment assignment

- Maybe, the pattern of observed treatments is not completely random

- Maybe, the infants who received phototherapy have different characteristics than those who were treated with watchful waiting

Gestational Age and Phototherapy

	Watchful Waiting ($X = 0$)	**Phototherapy ($X = 1$)**	**All**
Gestational Age $\leq$ 37 weeks ($Z = 0$)	4240 (69%)	1900 (31%)	6140 (30%)
Gestational Age $>$ 37 weeks ($Z = 1$)	11,907 (82%)	2684 (18%)	14,591 (70%)
All	16,147 (78%)	4584 (22%)	20,731

Non-random treatment assignment

- We know X is not independent of Z
- We're not sure if $Y(x) \perp\!\!\!\perp X$

- Suppose Z indicates the gestational age of the infant, categorized:
  - $Z = 0$ if gest. age ≤ 37 weeks
  - $Z = 1$ if gest. age > 37 weeks

- 2 of 5 infants who received phototherapy had gest. age > 37 weeks, versus 3 of 5 of infants who did not receive phototherapy

- Does it still make sense to just average the observed outcomes from all the phototherapy infants together?

Z	X	Y	$Y(1)$	$Y(0)$
0	0	1	?	1
0	0	1	?	1
1	0	0	?	0
1	0	1	?	1
1	0	1	?	1
0	1	0	0	?
0	1	1	1	?
0	1	1	1	?
1	1	0	0	?
1	1	1	1	?

Analysis 2: Stratification

- Maybe we are willing to assume that the observed treatment is being randomly chosen, conditional on gestational age Z; (mathematically: $Y(x) \perp\!\!\!\perp X|Z$). This is called a "conditional independence" assumption (or "conditional exchangeability" or "ignorability")
- Then:

$$E[Y(1)|Z=0,X=0] = E[Y(1)|Z=0,X=1] \\ = E[Y|Z=0,X=1]$$

Now, we have a basis for imputing the missing outcomes again:

Z	X	Y	$Y(1)$	$Y(0)$
0	0	1	2/3	1
0	0	1	2/3	1
1	0	0	?	0
1	0	1	?	1
1	0	1	?	1
0	1	0	0	?
0	1	1	1	?
0	1	1	1	?
1	1	0	0	?
1	1	1	1	?

Analysis 2: Stratification

- Maybe we are willing to assume that the observed treatment is being randomly chosen, conditional on gestational age Z; (mathematically: $Y(x) \perp\!\!\!\perp X|Z$). This is called a "conditional independence" assumption (or "conditional exchangeability" or "ignorability")
- Then:

$$E[Y(1)|Z=1,X=0] = E[Y(1)|Z=1,X=1] \\ = E[Y|Z=1,X=1]$$

Now, we have a basis for imputing the missing outcomes again:

Z	X	Y	$Y(1)$	$Y(0)$
0	0	1	2/3	1
0	0	1	2/3	1
1	0	0	1/2	0
1	0	1	1/2	1
1	0	1	1/2	1
0	1	0	0	?
0	1	1	1	?
0	1	1	1	?
1	1	0	0	?
1	1	1	1	?

Analysis 2: Stratification

- Maybe we are willing to assume that the observed treatment is being randomly chosen, conditional on gestational age Z; (mathematically: $Y(x) \perp\!\!\!\perp X|Z$). This is called a "conditional independence" assumption (or "conditional exchangeability" or "ignorability")
- Then:

$$E[Y(0)|X = 1, Z = 0] = E[Y(0)|X = 0, Z = 0]$$
$$= E[Y|X = 0, Z = 0]$$

Now, we have a basis for imputing the missing outcomes again:

Z	X	Y	$Y(1)$	$Y(0)$
0	0	1	2/3	1
0	0	1	2/3	1
1	0	0	1/2	0
1	0	1	1/2	1
1	0	1	1/2	1
0	1	0	0	2/2
0	1	1	1	2/2
0	1	1	1	2/2
1	1	0	0	?
1	1	1	1	?

Analysis 2: Stratification

- Maybe we are willing to assume that the observed treatment is being randomly chosen, conditional on gestational age Z; (mathematically: $Y(x) \perp\!\!\!\perp X|Z$). This is called a "conditional independence" assumption (or "conditional exchangeability" or "ignorability")
- Then:

$$E[Y(0)|X=1,Z=1] = E[Y(0)|X=0,Z=1] \\ = E[Y|X=0,Z=1]$$

Now, we have a basis for imputing the missing outcomes again:

Z	X	Y	$Y(1)$	$Y(0)$
0	0	1	2/3	1
0	0	1	2/3	1
1	0	0	1/2	0
1	0	1	1/2	1
1	0	1	1/2	1
0	1	0	0	2/2
0	1	1	1	2/2
0	1	1	1	2/2
1	1	0	0	2/3
1	1	1	1	2/3

Analysis 2: Stratification

- Once we have imputed all of the $Y(1)$s and $Y(0)$s, we can estimate $\hat{\text{E}}[Y(0)]$ and $\hat{\text{E}}[Y(1)]$:

- $\hat{\text{E}}[Y(1)] = \frac{1}{10}\left[\left(\frac{2}{3} \times 2\right) + \left(\frac{1}{2} \times 3\right) + 2 + 1\right] = .58$

Z	X	Y	$Y(1)$	$Y(0)$
0	0	1	2/3	1
0	0	1	2/3	1
1	0	0	1/2	0
1	0	1	1/2	1
1	0	1	1/2	1
0	1	0	0	2/2
0	1	1	1	2/2
0	1	1	1	2/2
1	1	0	0	2/3
1	1	1	1	2/3

Analysis 2: Stratification

- Once we have imputed all of the $Y(1)$s and $Y(0)$s, we can estimate $\hat{E}[Y(0)]$ and $\hat{E}[Y(1)]$:

- $\hat{E}[Y(1)] = \frac{1}{10}\left[\left(\frac{2}{3} \times 2\right) + \left(\frac{1}{2} \times 3\right) + 2 + 1\right] = .58$
- $\hat{E}[Y(0)] = \frac{1}{10}\left[2 + 2 + \left(\frac{2}{2} \times 3\right) + \left(\frac{2}{3} \times 2\right)\right] = .83$
- $\hat{E}[Y(1) - Y(0)] = .58 - .83 = -.25$

- Compare with what we got from the unstratified analysis:

$$\hat{E}[Y(1) - Y(0)] = .6 - .8 = -.20$$

Z	X	Y	$Y(1)$	$Y(0)$
0	0	1	2/3	1
0	0	1	2/3	1
1	0	0	1/2	0
1	0	1	1/2	1
1	0	1	1/2	1
0	1	0	0	2/2
0	1	1	1	2/2
0	1	1	1	2/2
1	1	0	0	2/3
1	1	1	1	2/3

Gestational Age, Phototherapy, and Worsened Jaundice

		Condition Stabilized or Improving ($Y = 0$)	Condition Worsened ($Y = 1$)	All	All
Gestational Age ≤ 37 weeks ($Z = 0$)	Watchful Waiting ($X = 0$)	4154 (98.0%)	86 (2.0%)	4240 (69%)	6140 (30%)
	Phototherapy ($X = 1$)	1890 (99.5%)	10 (0.5%)	1900 (31%)	
Gestational Age > 37 weeks ($Z = 1$)	Watchful Waiting ($X = 0$)	11,880 (99.8%)	27 (0.2%)	11,907 (82%)	14,591 (70%)
	Phototherapy ($X = 1$)	2679 (99.8%)	5 (0.2%)	2684 (18%)	
All		20,603 (99.4%)	128 (0.6%)		20,731

Estimated causal risk of phototherapy $= 0.3\%$
Estimated causal risk of waiting $= 0.8\%$

Estimated Causal Risk Difference from Stratified Analysis = $-.5\%$
(Estimated Causal Risk Difference from Unstratified Analysis = $-.4\%$)

Analysis 3: Regression

- What if Z is a numeric variable, e.g., gestational age measured in weeks?
- Stratification likely won't work: there aren't any rows with $Z = 37$ and $X = 0$ that we can use to estimate $\mathrm{E}[Y|Z = 37, X = 0]$.
- We could categorize Z as we did before, but maybe we need Z in its continuous form to justify $Y \perp\!\!\!\perp X(x)|Z$.
- However, we can still estimate $\mathrm{E}[Y|Z = 37, X = 0]$ by fitting a regression model!

$\boldsymbol{Z}$	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Y(1)}$	$\boldsymbol{Y(0)}$
36	0	1	?	1
35	0	1	?	1
38	0	0	?	0
40	0	1	?	1
39	0	1	?	1
35	1	0	0	?
37	1	1	1	?
36	1	1	1	?
40	1	0	0	?
38	1	1	1	?

Assumptions for Causal Regression Modeling

- Still need conditional independence:
  $$Y(x) \perp\!\!\!\perp X|Z$$
- Still need consistency and non-interference
- Need all treatment options to be possible for every possible value of Z:
  $$0 < P(X = 1|Z = z) < 1$$
  - Called “positivity assumption”; more of a practical requirement: if there some observations with $X = 1$ and $Z = 33$ but none with $X = 0$ and $Z \approx 33$, then how can we reliably predict $\mathrm{E}[Y|X = 0, Z = 33]$?
  - Will end up extrapolating, with extreme uncertainty (low precision).

$\boldsymbol{Z}$	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Y(1)}$	$\boldsymbol{Y(0)}$
36	0	1	?	1
35	0	1	?	1
38	0	0	?	0
40	0	1	?	1
39	0	1	?	1
35	1	0	0	?
37	1	1	1	?
36	1	1	1	?
40	1	0	0	?
38	1	1	1	?

Estimating Causal Effects with Regression Modeling

- If our assumptions hold, then: $\mathrm{E}[Y(1)|X=0,Z=z] = \mathrm{E}[Y|X=0,Z=z]$
- We can impute the missing potential outcomes:

$\boldsymbol{Z}$	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Y(1)}$	$\boldsymbol{Y(0)}$
36	0	1	$\hat{\mathrm{E}}[Y\|X=1,Z=36]$	1
35	0	1	?	1
38	0	0	?	0
40	0	1	?	1
39	0	1	?	1
35	1	0	0	?
37	1	1	1	?
36	1	1	1	?
40	1	0	0	?
38	1	1	1	?

Estimating Causal Effects with Regression Modeling

- If our assumptions hold, then:
  $\mathrm{E}[Y(1)|X=0,Z=z]=\mathrm{E}[Y|X=0,Z=z]$
- We can impute the missing potential outcomes:

$\boldsymbol{Z}$	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Y(1)}$	$\boldsymbol{Y(0)}$
36	0	1	$\hat{\mathrm{E}}[Y\|X=1,Z=36]$	1
35	0	1	$\hat{\mathrm{E}}[Y\|X=1,Z=35]$	1
38	0	0	?	0
40	0	1	?	1
39	0	1	?	1
35	1	0	0	?
37	1	1	1	?
36	1	1	1	?
40	1	0	0	?
38	1	1	1	?

Estimating Causal Effects with Regression Modeling

- If our assumptions hold, then: $\mathrm{E}[Y(1)|X=0,Z=z] = \mathrm{E}[Y|X=0,Z=z]$
- We can impute the missing potential outcomes
- Can also regress on more than one Z variable, to better justify $Y(x) \perp\!\!\!\perp X|Z_1,\ldots,Z_p$

$\boldsymbol{Z}$	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Y(1)}$	$\boldsymbol{Y(0)}$
36	0	1	$\hat{\mathrm{E}}[Y\|X=1,Z=36]$	1
35	0	1	$\hat{\mathrm{E}}[Y\|X=1,Z=35]$	1
38	0	0	?	0
40	0	1	?	1
39	0	1	?	1
35	1	0	0	?
37	1	1	1	?
36	1	1	1	?
40	1	0	0	?
38	1	1	1	?

Causal regression modeling of phototherapy and jaundice

- Vittinghoff et al (2012) performed logistic regression modeling on the jaundice data using the following predictor covariates: **treatment** (phototherapy vs. watchful waiting), **chromosomal sex**, **gestational age** (discretized into 6 categories), **birth weight** (numeric, linear term), **interaction between gestational age and birth weight**, **bilirubin level at time of treatment assignment** (relative to a guideline threshold for phototherapy treatment), **age at time of treatment assignment** (discretized into days), and **hospital** (treated as a clustering variable)
- Results:
  - $\widehat{\mathrm{P}}[Y(1) = 1] = 0.16\%;$
  - $\widehat{\mathrm{P}}[Y(1) = 0] = 0.96\%;$
  - Estimated risk difference $= \; -0.79\%$
- Compare: unadjusted analysis risk difference: -0.4%; risk difference stratifying on gestational age $\leq$ 37 weeks: -0.5%

Analysis 4: Matching (briefly)

- If a given observation has no exact counterparts (with opposite treatment), maybe we can use an approximate counterpart instead:

Z	X	Y	$Y(1)$	$Y(0)$
36	0	1	?	1
35	0	1	?	1
38	0	0	?	0
40	0	1	?	1
39	0	1	?	1
35	1	0	0	?
37	1	1	1	?
36	1	1	1	?
40	1	0	0	?
38	1	1	1	?

Analysis 4: Matching (briefly)

- If a given observation has no exact counterparts (with opposite treatment), maybe we can use an approximate counterpart instead.
- Maybe we just pick one of the closest matches and call it close enough?

Z	X	Y	$Y(1)$	$Y(0)$
36	0	1	?	1
35	0	1	?	1
38	0	0	?	0
40	0	1	?	1
39	0	1	?	1
35	1	0	0	?
37	1	1	1	?
36	1	1	1	?
40	1	0	0	?
38	1	1	1	?

Analysis 4: Matching (briefly)

- If a given observation has no exact counterparts (with opposite treatment), maybe we can use an approximate counterpart instead.
- Maybe we just pick one of the closest matches and call it close enough?
- Maybe we pick one “matching” counterpart for every observation?
- We might need to discard some observations without any close matches.
- There are many different methods for matching.

Z	X	Y	$Y(1)$	$Y(0)$
36	0	1	?	1
35	0	1	?	1
38	0	0	?	0
40	0	1	?	1
39	0	1	?	1
35	1	0	0	?
37	1	1	1	0
36	1	1	1	?
40	1	0	0	?
38	1	1	1	?

Analysis 5: Propensity scores (also briefly)

- If we need several Zs to justify the conditional independence assumption, stratification, regression, and matching can become very complicated.
- Maybe we can combine those Zs into a single variable that summarizes them and still provides conditional independence.
- If X is binary and the conditional independence assumption holds for $Z_1, \dots, Z_p$, then it also holds for $\pi\left(z_1, \dots, z_p\right) = P(X = 1|Z_1 = z_1, \dots, Z_p = z_p)$; that is, $Y(x) \perp\!\!\!\perp X \mid \pi(Z)$
- We can estimate $\hat{\pi}\left(z_1, \dots, z_p\right) = \widehat{\mathrm{P}}\left(X = 1\middle|Z_1 = z_1, \dots, Z_p = z_p\right)$ and use it with univariate stratification, regression, or matching.
- More on this topic in the third seminar in this series!

How do we know if we have the right covariates?

- The conditional independence assumption is crucial for all the methods we discussed today. How can we tell if it is plausible? Hard to even think about.
- Maybe we can make smaller, easier-to-understand, possibly even testable, assumptions, from which we could mathematically deduce whether a given set of covariates provides conditional independence.
- Next session: we draw flow-chart diagrams (called directed acyclic graphs, “DAGs”) to represent our assumptions about the data-generating process, and analyze these diagrams to determine which sets of covariates would produce conditional independence, given our assumptions.

Other causal inference topics to explore

- We haven't discussed how to compute standard errors or confidence intervals for our causal effect estimates.
  - There are various methods, but when in doubt, try the bootstrap: often conceptually simple, although computationally time-consuming.
- There are other common causal inference methods:
  - inverse-probability weighting (IPW)
  - g-estimation
  - Instrumental variables

Help is available

- My email: demorrison@ucdavis.edu
- CTSC and Cancer Center Biostatistics Office Hours
  - Every Tuesday from 12 – 2:00 currently via WebEx
  - 1st & 3rd Wednesday from 1:00 – 2:00 currently via WebEx
  - Sign-up through the CTSC Biostatistics Website
- EHS Biostatistics Office Hours
  - Upon request
- Request Biostatistics Consultations
  - CTSC
  - MIND IDDRC
  - Cancer Center Shared Resource
  - EHS Center

Thanks for attending!