

STATISTICS & ML WITH R

**Mapping causal & predictive
approaches**

2024

M. Chiara Mimmi & Luisa M. Mimmi

MODULE 4 – LECTURE OUTLINE

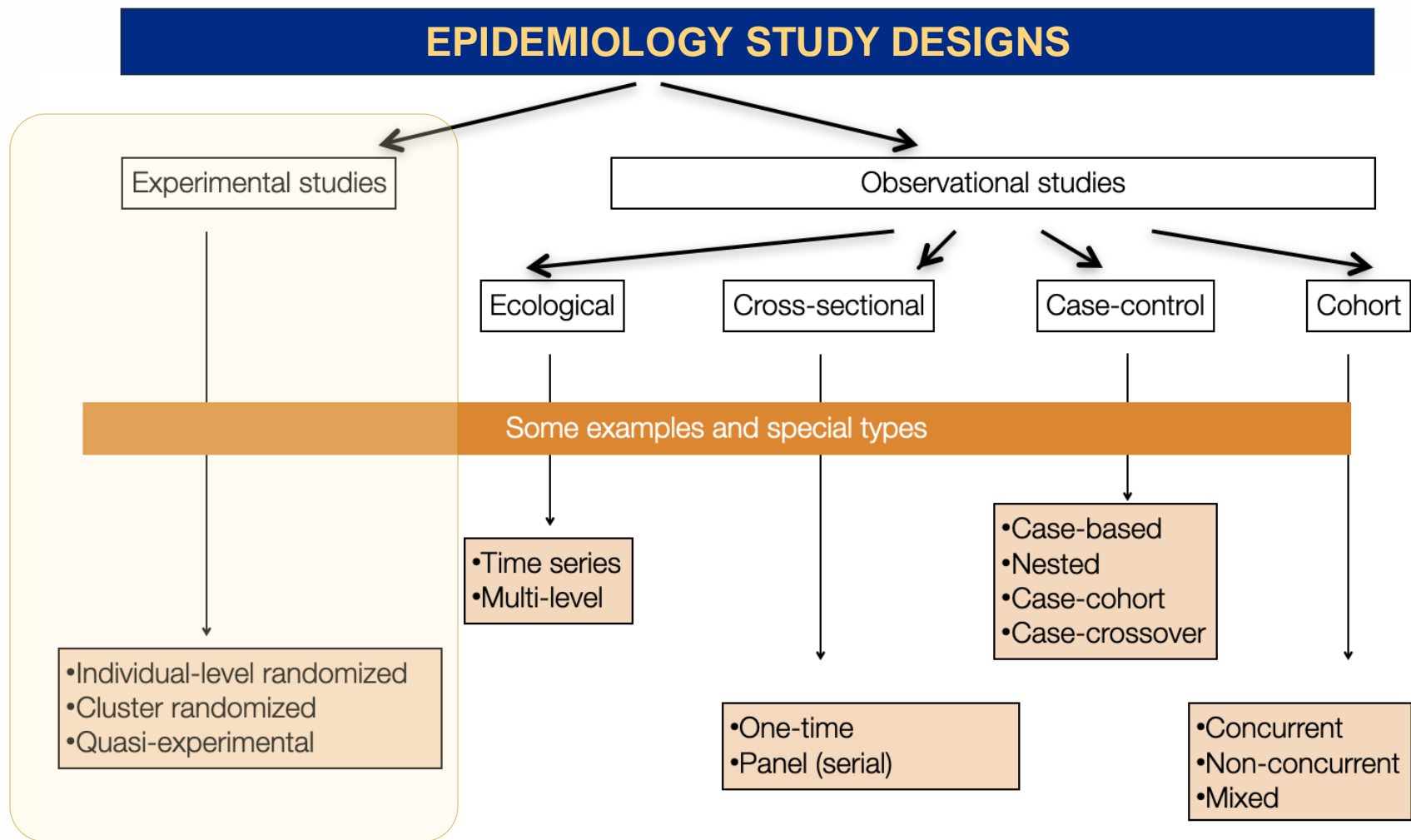
Mapping causal approaches

- Recall the essential features of experimental study designs
 - Learning the vocabulary of causal analysis
- Get a visual intuition of causal pathways, including challenging elements:
 - Collider variables
 - Confounder variables
 - Mediator variables
- Discuss the correct causal model to capture the association among exposure, outcome and other covariates, (including *challenging ones*)
- Define causal outcomes and choosing the appropriate “estimands”:
 - ATE, ATT, or ATU?
- Devise statistical methods to estimate ATE, ATT, and ATU based on research question and (sub)population of interest

From *observational* to *experimental* studies

- “**OBSERVATIONAL STUDIES**” on variables of interest and their relationships have *no controlled assignment of the treatment*
 - We may find **CORRELATION / ASSOCIATION**, but it DOES NOT IMPLY CAUSATION! Why?
 - ... **hidden variables** may affect the relationship between the **explanatory variable** and the **response variable**
 - ...*but often used (implicitly or not) to estimate causal effect of an exposure!*
- “**EXPERIMENTAL STUDIES**” seek to uncover **CAUSATION**, so they are *designed to provoke a response*
 - Researchers **assign the treatment** to an **experimental unit** (or **subject**) and observing its **effect**
 - These studies use some *ad hoc* **design principles** and **controlled independent variables**

Experimental and non-experimental study designs...



Source: <https://bookdown.org/jbrophy115/bookdown-clinepi/design.html>

Different goals of statistical modeling (part 1/2)

1. **ASSOCIATION/CORRELATION** → observational studies

- aimed at **summarizing or representing the data structure**, without an underlying causal theory
- may help **form hypotheses** for explanatory and predictive modeling

2. **CAUSAL EXPLANATION** → experimental studies

- aimed at **testing “explanatory connection”** between treatment and outcome variables
- prevalent in “**causal theory-heavy**” fields (economics, **psychology**, environmental science, etc.)

- **Note:**

- ✓ The **same modeling approach** (e.g., fitting a regression model) can be used for **different goals**
- ✓ While they shouldn't be confused, **explanatory power** and **predictive accuracy** are complementary goals: e.g., in bioinformatics (which has little theory and abundance of data), predictive models are pivotal in generating avenues for causal theory.

3. **EMPIRICAL PREDICTION** → algorithmic machine learning and data-mining modeling

A framework for CAUSAL ANALYSIS

Key terminology

The conceptual framework for causal analysis (1/3)

- **Fundamental vocabulary:**
 - **Intervention** decisions and actions that change the behaviors or situation of people/firms/other subjects (drug, vaccine, program participation)
 - **TREATMENT** = commonly used in experimental studies when researchers directly “assigns” the **causal variable**
 - **EXPOSURE** = commonly used observational studies when participants “naturally” experience the **causal variable**
 - **Subjects** = those that may be affected (at least in principle), in fact are
 - TREATED subjects
 - UNTREATED subjects
 - **Outcome** = variable(s) that may be affected by the intervention
 - can be caused by exposure either directly or through an intermediate process
 - **Causation** = causal processes that lead to the development of outcomes

The conceptual framework for causal analysis (2/3)

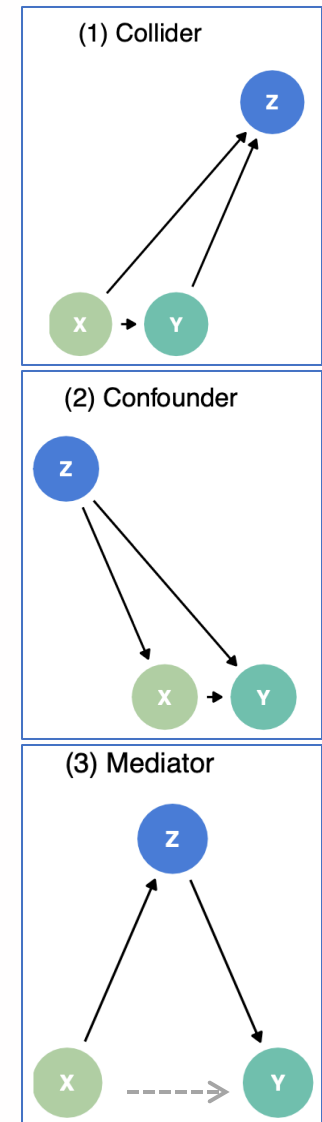
- Fundamental vocabulary (“tricky ones” 😊):
 - **Bias** = systematic error that can occur at different stages of the study: *data collection, analysis or interpretation* of the causal relationship exposure-outcome.
 - **Selection bias** = both the exposure and the outcome affect whether an individual is included in the sampled population
 - **Sampling bias** = some members of the intended population are less likely to be included than others
 - **Attrition bias** = participants who drop out of a study systematically differ from those who remain
 - **Non-response bias** = participants who refuse to participate in the study systematically differ from those who take part
 - **Recall bias** = a systematic difference in the ability of participant groups to accurately recall information
 - **Information bias** = there is misclassification or inaccurate measurement (e.g., patients underreporting smoking habits)
 - ...

Check out this cool list of all types of bias:
<https://quantifyinghealth.com/list-of-biases/>

The conceptual framework for causal analysis (3/3)

- Fundamental vocabulary (“tricky ones” 😊):

- Collider** = variable that is **influenced by treatment and outcome** (like a “common effect”)
 - EXAMPLE:** sleepiness (Z), with shift work (X) and apnea (Y)
 - Conditioning on or controlling for a collider in the causal model can create a distortion (*“collider bias”*)
- Confounder** = variable that **affect both treatment and outcome** (“apparent” cause), but it is **not in the causal pathway**
 - EXAMPLE:** smoking (Z), with exercise (X) and lung cancer (Y)
 - Most confounder variables involve some *kind of selection* (e.g., self-selection) that can be addressed stratifying subjects by it
- Mediator** = is a variable that is **in the causal pathway** and “explains” why **treatment affects outcome** (like a “mechanisms”)
 - EXAMPLE:** immune function (Z), with exercise (X) and lung cancer (Y)
 - Conditioning on or controlling for a mediator can be done to assess what *part of the effect* they play



Estimands, Estimators, Estimates

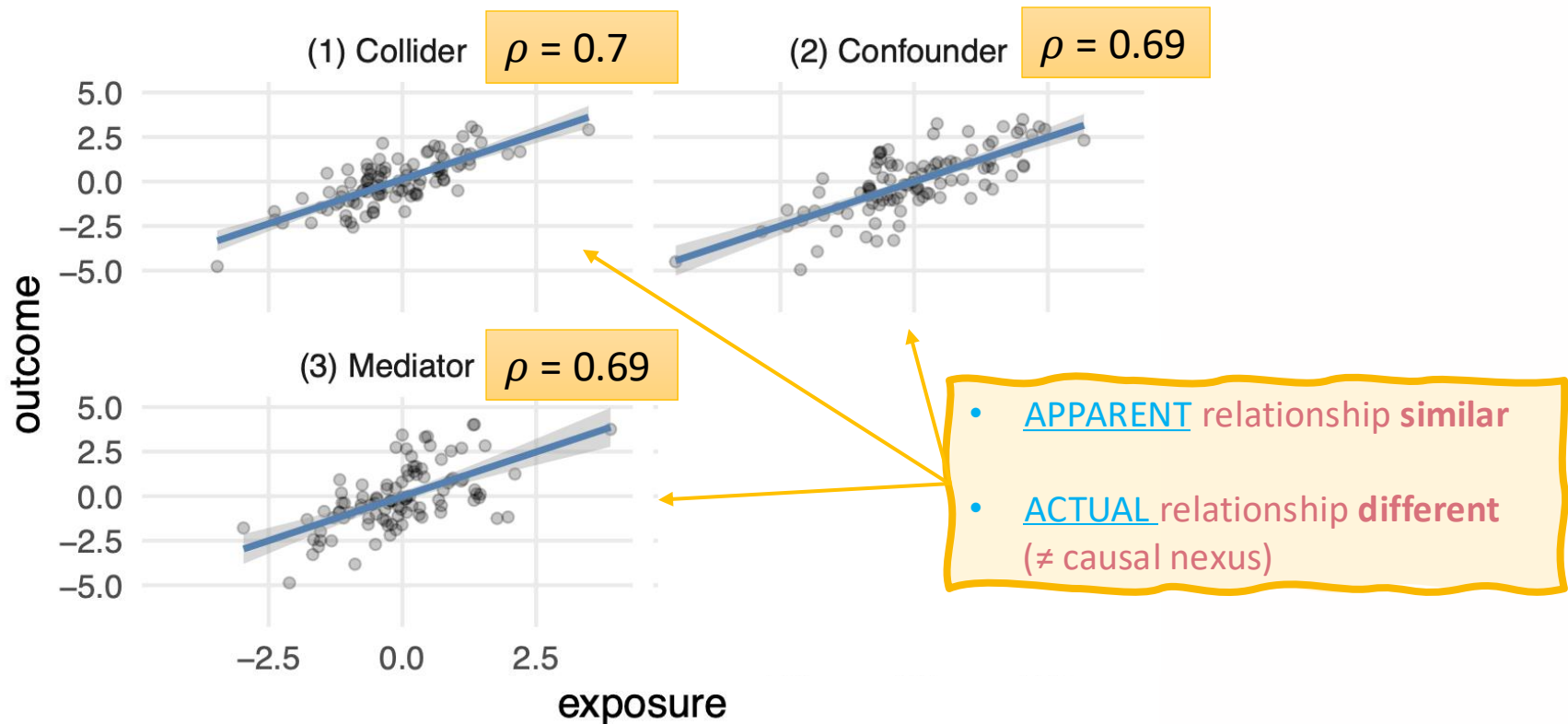
- The **estimand** is the *target outcome* of interest about the causal effect of a treatment in a population
 - **EXAMPLEs**: **ATE** (Average Treatment Effect), **ATT** (Average Treatment Effect on the Treated), or **ATU** (Average Treatment Effect on the Untreated)
- The **estimator** is the *statistical method* (“recipe”) by which we approximate this *estimand* using data
 - **EXAMPLEs**: **difference-in-means** for ATE in a randomized controlled trial (RCT), or **propensity scores matching** (PSM) for ATT within observational data.
- The **estimate** is the *numerical value we get* when we plug our data into the estimator
 - **EXAMPLE**: we calculate an **ATE = 3.5 units** (e.g., a treatment improves test scores by 3.5 points on average in the entire population)

Visualizing causal maps

Using “DAGs” in guiding statistical modeling

Typical challenges in estimating causal effects: a.k.a. “correlation does not imply causation”

- Consider 3 distinct datasets: while their statistical summaries and visualizations are very similar, the **true causal nexus differs!**
- Deciding the** correct model requires knowledge of the data-generating mechanism (i.e. the random assignment to exposure/not exposure in experiments)

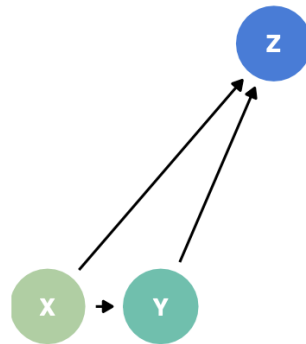


Source: Barrett, M., McGowan, L. D., & Gerke, T. (2024). *Causal Inference in R*. Retrieved from <https://www.r-causal.org/>

Typical challenges in estimating causal effects: visual intuition

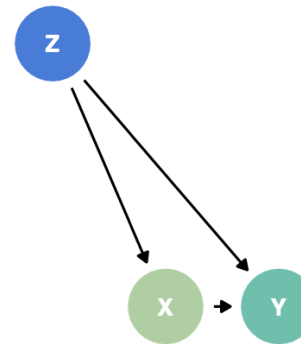
- Directed Acyclic Graphs (DAGs) can offer visual intuition of the causal nexus at play in the 3 datasets. Failure to adjust models to these situation leads to **BIAS**
 - X** is some exposure of interest, **Y** an outcome, and **Z** a known, measured factor

(1) Collider



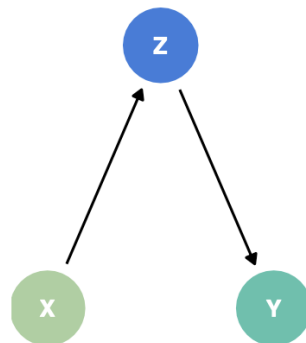
- (1) a “**COLLIDER**”, common effect (that invertedly connects). E.g.:
- X = sodium intake
 - Y = systolic blood pressure
 - Z = urinary protein excretion

(2) Confounder



- (2) a “**CONFOUNDER**”, common cause. E.g.:
- X = smoking
 - Y = lung cancer
 - Z = alcohol (consumers also tend to be smokers)

(3) Mediator

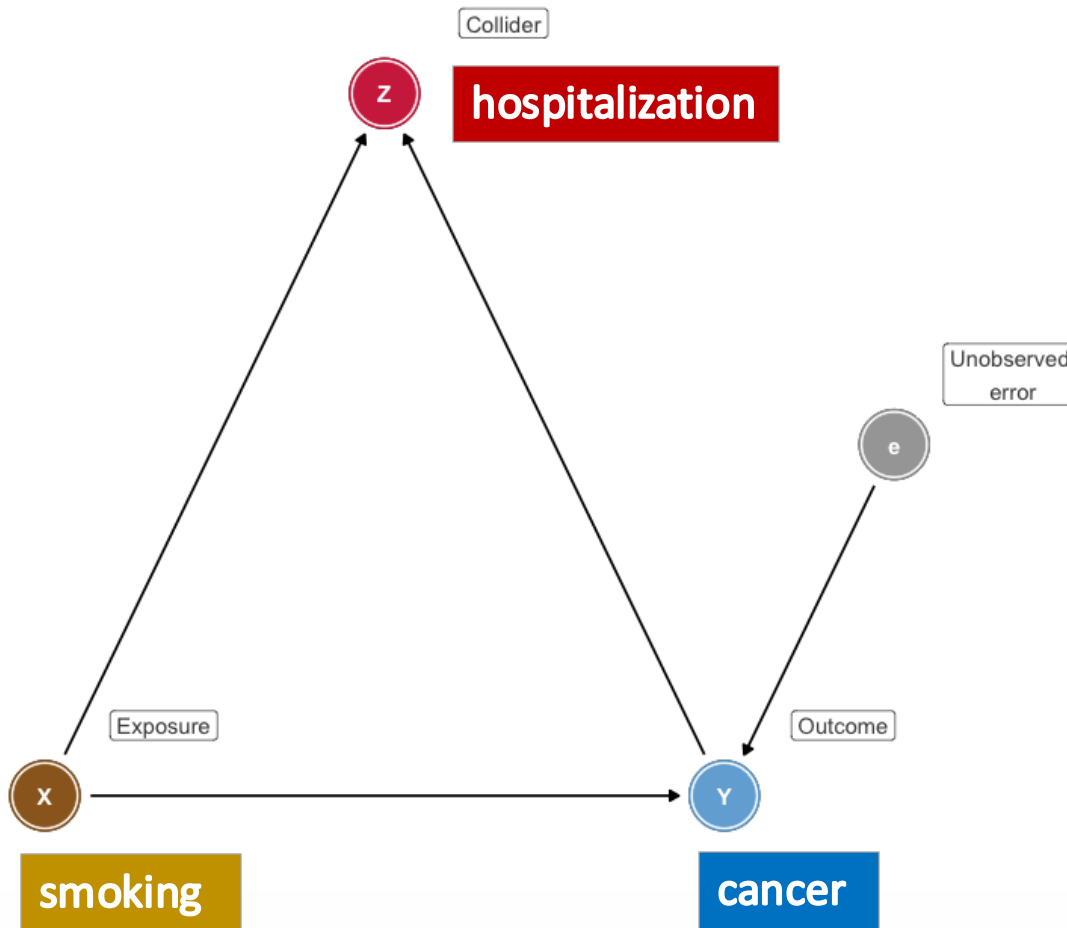


- (3) a “**MEDIATOR**” is caused by X and then it causes Y. E.g.:
- X = screen time
 - Y = obesity
 - Z = physical exercise

Source: Barrett, M., McGowan, L. D., & Gerke, T. (2024). *Causal Inference in R*. Retrieved from <https://www.r-causal.org/>

How to deal with **collider** (*common effect*) when modeling?

Causal map with COLLIDER (Z)

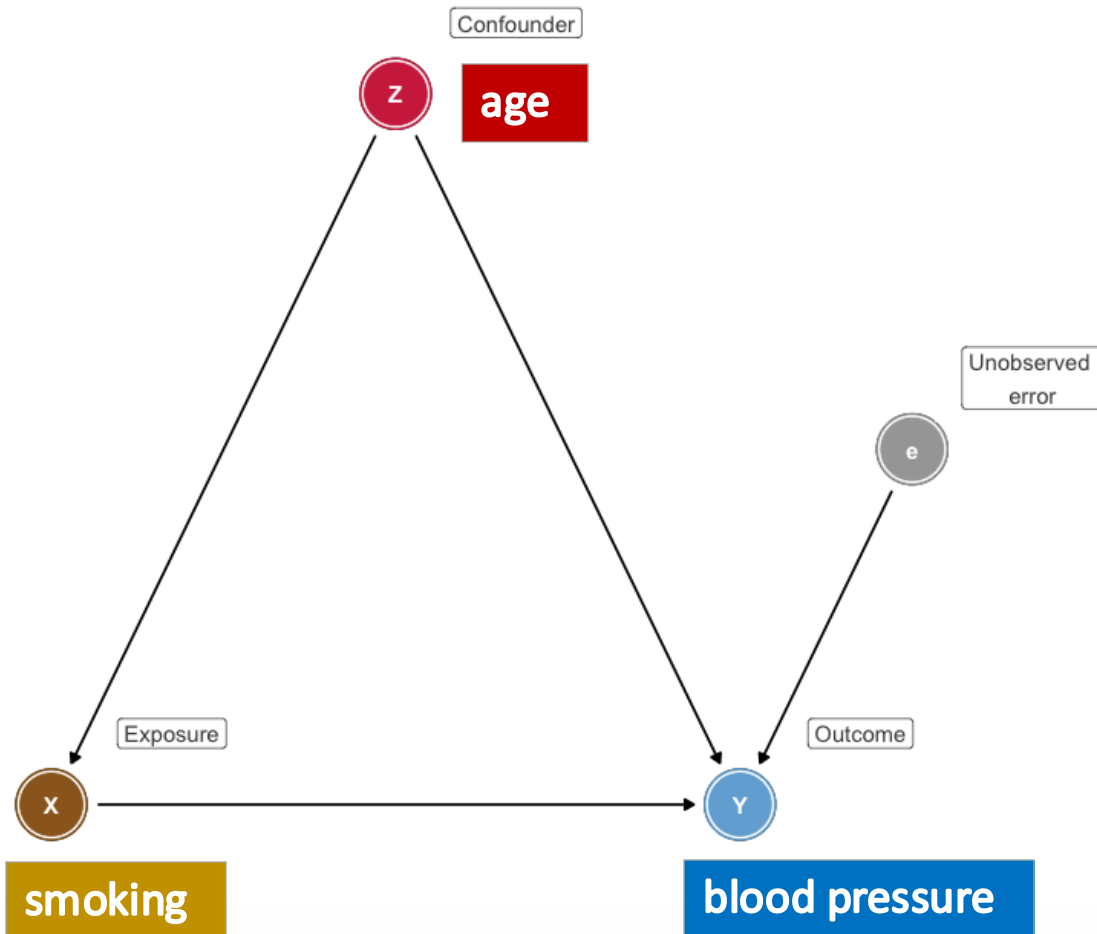


- We must **NOT** control for collider.
- **Colliders CAN HIDE REAL CAUSE EFFECTS**
 - i.e., it would distort the true relationship between the exposure and the outcome

(more in Lab. 4)

How to deal with **confounder** (*common cause*) when modeling?

Causal map with CONFOUNDER (Z)



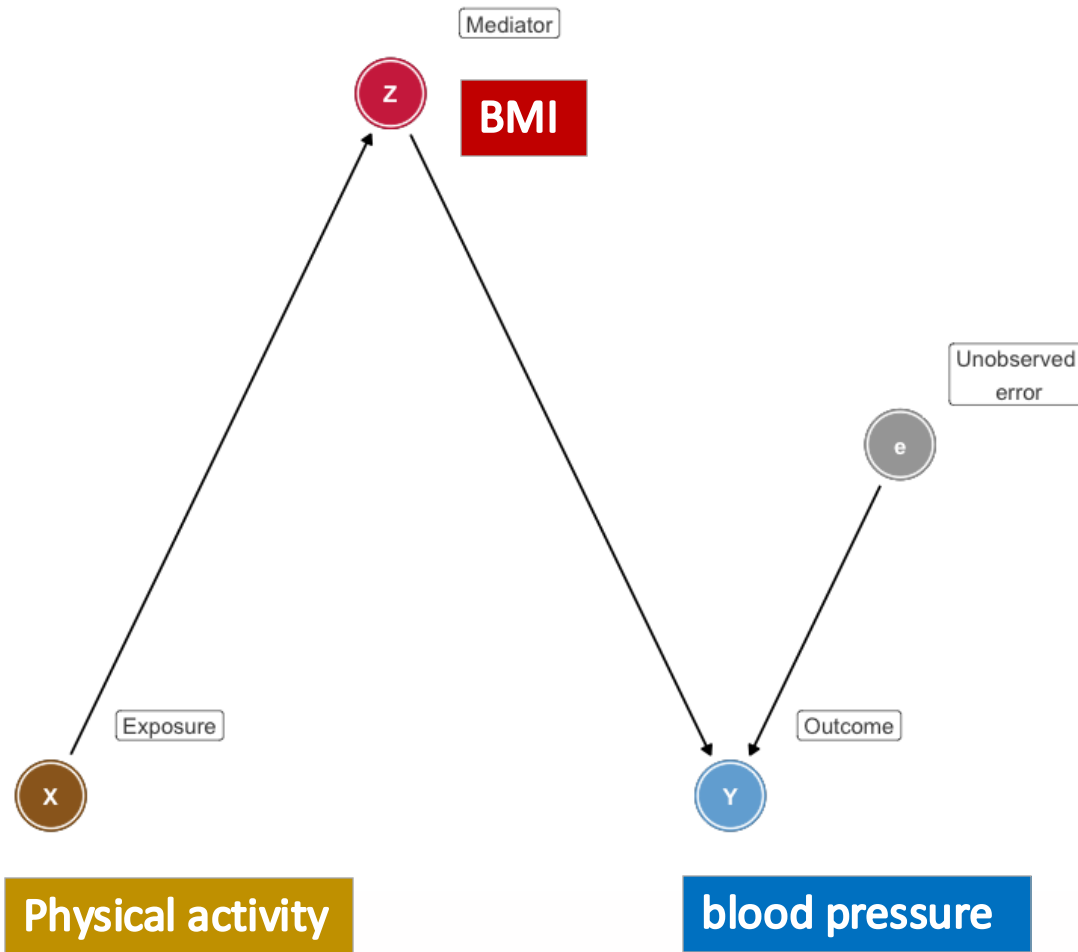
Source: Excellent reading <https://quantifyinghealth.com/control-confounding/>

- We must **control for** a confounder, so we reduce bias. HOW?
- 1) At **design** stage:
 - Random assignment
 - Restriction (*only participants of a certain confounder category*)
 - Matching observations (*confounder distributed evenly by exposure*)
- 2) In **analysis** stage:
 - Stratifying sample in subgroups (*by confounding*)
 - Including term in regression
 - Inverse probability weighting (*equalizing frequency of confounder by exposure*)
 - Instrumental variable estimation

(more in Lab. 4)

How to deal with **mediator** (*mechanism*) when modeling?

Causal map with MEDIATOR (Z)



- We *could* control for the mediator, depending on which effect we focus on:
 - with **UNADJUSTED MODEL** we get only the total effect (direct + indirect) of X on Y
 - with **ADJUSTED MODEL** we separate the direct effect of X on Y (not mediated), and the indirect effect of M on Y (mediated)
- *Normally both models are shown*
- *The Adjusted model enables to see the **PROPORTION** of the **MEDIATOR** mechanism in the causal path*

(more in Lab. 4)

Measuring causal outcomes of interest

Commonly used “**estimands**” (ATE, ATT, ATU) and how to select and interpret them correctly for making valid inferences

The explanation and examples below follow very closely these two sources:

1. **Noah Greifer, Elizabeth A. Stuart**, [*Choosing the Causal Estimand for Propensity Score Analysis of Observational Studies*](#)
2. **Andrew Heiss**, [*Demystifying causal inference estimands: ATE, ATT, and ATU*](#)

Defining estimands at the *subject level*

- **NOTATION:**

- Y^0 and Y^1 are the potential outcomes in the *absence* and *presence* of treatment for patient i in a study on a new drug on blood pressure,
 - Y_i^0 = patient's blood pressure with receives standard of care
 - Y_i^1 = patient's blood pressure with takes new drug

- **ITE = Individual Treatment Effect (*)** = difference, for subject i , between potential outcome y if treated and if untreated

$$te_i = \delta_i = y_i^1 - y_i^0 \text{ where treatment is } T_i = \{0,1\}$$

- (*) ITE is *never* observable!!
- Hence, we will look at averages...

- **ATE = Average Treatment Effect** = average of ITE differences across subjects

$$E[te_i] = E[Y_i^1 - Y_i^0] = E[Y_i^1 | T_i = 1] - E[Y_i^0 | T_i = 0]$$

- (*) The Avg of the differences = the difference of Averages!
- ATE can hide different distributions of ITEs (e.g., positives and negatives that cancel each other out)
- Important to have a well-defined group or population

Defining estimands at the *subject level*

- **ATT (or ATET) = Average Treatment effect on the Treated** = average treatment effect across all subjects that end up TREATED

$$E[\delta_i \mid T_i = 1] = E[Y_i^1 - Y_i^0 \mid T_i = 1] = E[Y_i^1 \mid T_i = 1] - E[Y_i^0 \mid T_i = 1]$$

- This refers to the avg of the differences conditionally on the fact that both groups “received” the treatment (“ $T_i = 1$ ”)
 - $E[Y_i^0 \mid T_i = 1]$ is essentially the counterfactual for Y_i^1 in a 'parallel universe' where **exactly the same people** who were treated in this universe would not get the treatment
- **ATU = Average Treatment effect on the Untreated** = average treatment effect across all subjects who were NOT TREATED

$$E[\delta_i \mid T_i = 0] = E[Y_i^1 - Y_i^0 \mid T_i = 0] = E[Y_i^1 \mid T_i = 0] - E[Y_i^0 \mid T_i = 0]$$

- This time we seek the Avg of the differences (“ $T_i = 1$ ”) conditionally on the fact that both groups were “assigned” to the treatment
- $E[Y_i^1 \mid T_i = 0]$ is essentially the counterfactual for Y_i^0 in a 'parallel universe' where **exactly the same people** who were NOT treated in this universe would get the treatment

By the way !

- **Treatment** is a binary random variable $T_i = \{0,1\}$
- **Outcome** of interest is $Y_i = \begin{cases} Y_i^0 & (\text{if } T_i = 0) \\ Y_i^1 & (\text{if } T_i = 1) \end{cases}$
- **ATE = Average Treatment Effect** = average of ITE differences across subjects
- **ATT/ATET = Average Treatment effect on the Treated** = average treatment effect across all subjects that end up TREATED

EXAMPLE: Does hospitalization (T) increase health (Y) ?		
ATE =	$E[Y_i^1 T_i = 1] - E[Y_i^0 T_i = 0]$	Avg health of hospitalized group – avg health of NOT hospitalized group
ATT +	$E[Y_i^1 T_i = 1] - E[Y_i^0 T_i = 1]$	Avg health of treated group – [counterfactual] avg health $E[Y_i^0]$ of treated group IF NOT hospitalized
+ Selection bias	$E[Y_i^0 T_i = 1] - E[Y_i^0 T_i = 0]$ (hospitalized have worse Y_i^0 than non hospitalized)	Difference in [counterfactual] avg health $E[Y_i^0]$ of treated group IF NOT hospitalized - those who were NOT hospitalized

EXE. potential causal outcomes, ITE (δ_i), depends on patients' characteristics)

	Confounder	Treatment	Unobservable		Realized
	Age	Treated	Potential outcomes	ITE	Outcome
ID	Z_i	X_i	Y_i^1	Y_i^0	Y_i
1	Old	1	80	60	80
2	Old	1	75	70	75
3	Old	1	85	80	85
4	Old	0	70	60	60
5	Young	1	75	70	75
6	Young	0	80	80	80
7	Young	0	90	100	100
8	Young	0	85	80	80

$$ATT = \frac{20 + 5 + 5 + 5}{4} = 8.75$$

$$ATU = \frac{10 + 0 - 10 + 5}{4} = 1.25$$

$$ATE = \frac{20 + 5 + 5 + 5 + 10 + 0 + -10 + 5}{8} = 5$$

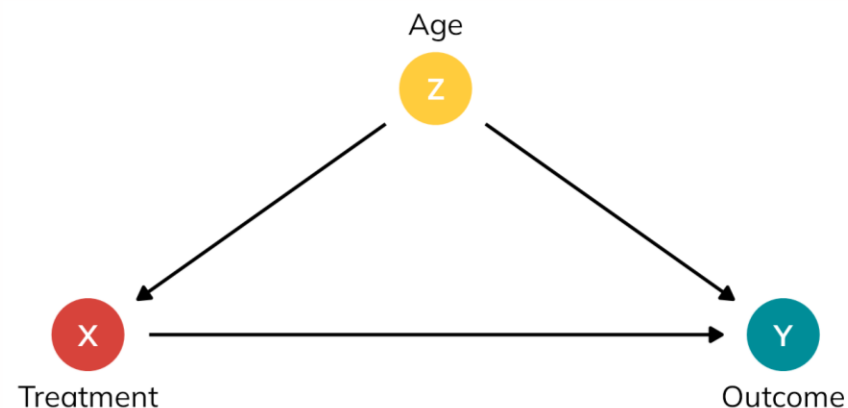
$$ATE = \left(\frac{4}{8} \times 8.75\right) + \left(\frac{4}{8} \times 1.25\right) = 4.375 + 0.625 = 5 \quad (ATE \text{ decomposition})$$

Revisiting the **confounder** seen in DAG visualization

Accounting for **age** for an accurate estimate of
ATE

Acknowledging a confounder variable

ID	Confounder	Treatment	Realized
	Age	Treated	Outcome
	Z_i	X_i	Y_i
1	Old	1	80
2	Old	1	75
3	Old	1	85
4	Old	0	60
5	Young	1	75
6	Young	0	80
7	Young	0	100
8	Young	0	80



- So far, we ignored it, but:
- **AGE** seems to behave as a **confounder**:
 - → it is highly correlated with **treatment** status
 - → it also affects the ultimate value of the **outcome**
- *Hence, we need to account for it statistically*

How to deal with confounder variable?

- Recall that we listed different ways to *control for* a confounder (to reduce bias in estimands)
- Here, we illustrate 2 of them (feasible *at analysis stage*):
- 1) At **design** stage:
 - **Random assignment**
 - **Restriction** (*only participants of a certain confounder category*)
 - **Matching observations** (*confounder distributed evenly by exposure*)
- 2) In **analysis** stage:
 - ✓ **Stratifying sample in subgroups** (*by confounding*)
 - ✓ **Including confounder variable as term in regression**
 - **Inverse probability weighting** (*equalizing frequency of counfounder by exposure*)
 - **Instrumental variable estimation**

1) Stratification to deal with confounder (i.e. combining the weighted averages for old and young people)

	Confounder	Treatment	Realized
	Age	Treated	Outcome
ID	Z_i	X_i	Y_i
1	Old	1	80
2	Old	1	75
3	Old	1	85
4	Old	0	60
5	Young	1	75
6	Young	0	80
7	Young	0	100
8	Young	0	80

$$\begin{aligned} \text{Effect}_{old} &= \bar{Y}_{treated} - \bar{Y}_{untreated} = \\ &= \frac{80 + 75 + 85}{3} - \frac{60}{1} = 20 \end{aligned}$$

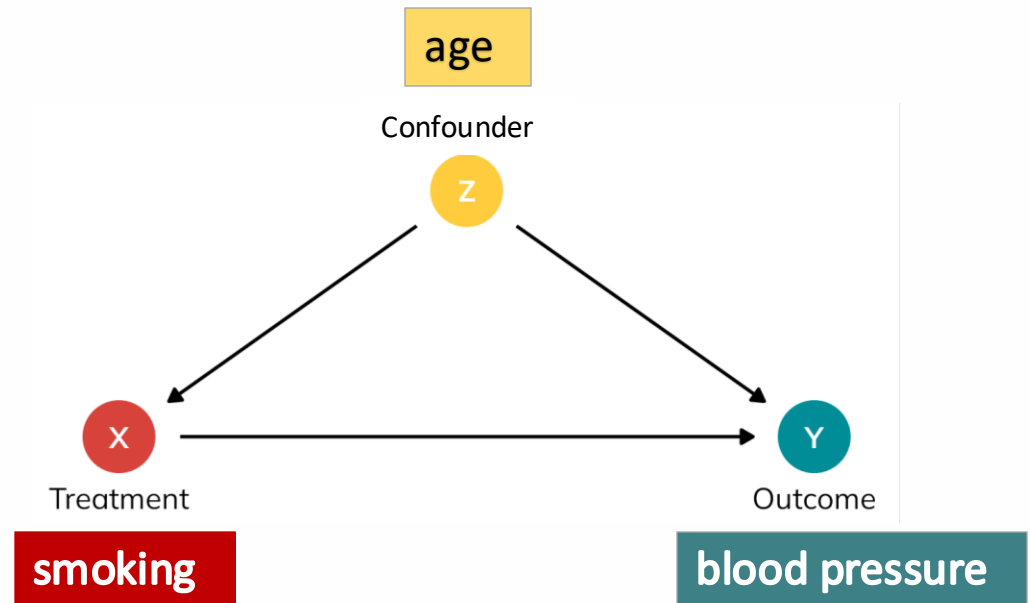
$$\begin{aligned} \text{Effect}_{young} &= \bar{Y}_{treated} - \bar{Y}_{untreated} = \\ &= \frac{75}{1} - \frac{80 + 100 + 80}{3} = -11.667 \end{aligned}$$

$$\text{ATE}_{stratified} = \pi_{old} \text{Effect}_{old} + \pi_{young} \text{Effect}_{young} = \left(\frac{4}{8} \times 20\right) + \left(\frac{4}{8} \times -11.667\right) = 4.1667$$

After stratification based on the confounder we get a very close **approximation of the ATE** (=5)

2) Introducing the **confounder** as term in the regression model

- Let's consider an example (to be discussed in the Lab) based on the NHANES dataset,
- where:
 - $Z = \text{Age} = \text{confounder}$
 - $X = \text{SmokeNow}$
 - $Y = \text{BPSysAve}$ (blood pressure)
- In the Lab, we will fit a regression model for the outcome and compare the results **WITH** and **WITHOUT** the confounder in the model



```
# Unadjusted linear model
model_unadjusted <- lm(BPSysAve ~ SmokeNow, data = nhanes_conf)
```

```
# Adjusted model
model_adjusted <- lm(BPSysAve ~ SmokeNow + Age, data = nhanes_conf)
```

2) Introducing the **confounder** as term in the regression model reduces bias and isolates the true effect of the X on Y

- In the **Adjusted Model** (with **Age**), the estimate of the causal effect of smoking (**SmokeNowYes**) on blood pressure (BPSysAve) is more accurate
 - the regression “adjusts” for the influence of **Z** in the causal path **X** → **Y**
 - this prevents *falsely attributing* to **Smoking (X)** an effect that might actually result from **X**

	NO Confounder	Confounder
(Intercept)	124.384**	100.997**
	[123.537, 125.231]	[98.863, 103.131]
	s.e. = 0.432	s.e. = 1.089
SmokeNowYes	-4.264**	0.684
	[-5.524, -3.004]	[-0.554, 1.922]
	s.e. = 0.643	s.e. = 0.631
Age		0.432**
		[0.395, 0.468]
		s.e. = 0.019
Num.Obs.	3108	3108
R2	0.014	0.158
* p < 0.05, ** p < 0.01		

Ways to deal with **confounders** must be carefully pondered

All these solutions have pros and cons that must be considered...

DESIGN

WHAT CAN BE DONE AT DATA COLLECTION?

- **RANDOM ASSIGNMENT** to treatment exposure
- **RESTRICTION** of participants of a certain confounder category
- **MATCHING** by distributing the confounder evenly (between exposed and unexposed)

ANALYSIS

WHAT CAN BE DONE DURING ANALYSIS?

- **STRATIFICATION** (as we saw by **Age**) by estimating the relationship outcome within different subsets of the confounder
- **REGRESSION** (as we saw with **Age**) can scale to many confounders
- **INVERSE PROBABILITY WEIGHTING** to balance confounder-weighted distributions and achieve comparability between treated and untreated groups

What about the other **estimands**? (ATT, ATU)

Choosing the **estimands** and the proper statistical method to estimate the effect

- In a **randomized trial**, the treated and untreated groups will, on average, have the same distributions of patient characteristics, so the **ATT, ATU, and ATE** will be the same
- **Without randomization**, however, the treatment groups can have quite different distributions of characteristics, **ATT, ATU, and ATE** will differ when these characteristics also relate to the treatment effect
 - So, when using observational data: *for whom should the treatment effect be estimated?*

Choosing the **estimands** based on the research question

BEFORE analyzing a dataset, let's consider **which question** we are asking, and **about which target** population group,

THEN choose a **statistical method** that corresponds to the chosen estimand.

Estimands	Target Population	Example research question and research/policy addressed
ATT	Treated patients	<i>Examining an intervention that would only reach those currently receiving it:</i> <i>- e.g. decision to replace / withhold a treatment for currently treated patients</i>
ATU	Untreated patients (control)	<i>How would untreated patients respond to a new potential treatment/exposure?</i> <i>- e.g. decision to extend a medical practice (drug prescription/vaccine) to a group that would not otherwise receive it</i>
ATE	Full sample / population	<i>Should a specific policy be applied to all eligible patients? How would the outcome be on average?</i> <i>- e.g. regulating a system-wide policy for a previously unregulated practice</i> <i>- useful when treatment decisions are not well informed (ATE does not depend on current treatment assignment)</i> <i>- NOT OK when patients' benefit depend on clinical judgment</i>

Recapping key points of the lecture

- Capturing the **causal nexus** between a **treatment** and an **outcome** may be tricky due to:
 - constraints in study design, sampling, or data collection process
 - repeated measures over time
 - effects of *confounder*, *collider* or *mediator/mechanism* variables
 - etc.
- **Visual causal maps** may help guiding the analysis, by summarizing how variables affect each other
 - e.g. **DAGs** (Directed Acyclic Graphs)
- Another key step is deciding which **estimand(s)** we are seeking with reference to the **specific target population**:
 - **ATE** – if a treatment targeted to the general population (e.g. a universal policy)
 - **ATT** – if a treatment targeted subjects already exposed (e.g. \neq drug for treated patients)
 - **ATU** – if a new treatment apply to currently untreated patients (e.g. new drug)
- Each estimand implies its own assumptions, interpretation, and **statistical methods**
- *coming up next...*

Shifting emphasis on empirical outcome prediction

Introduction to **Machine Learning (ML)**
models

A conceptual framework to understand different types of statistical **modeling** (part 2/2)

1. **association/correlation** → observational studies
2. **causal explanation** → experimental studies
3. **empirical prediction** → algorithmic machine learning and data-mining modeling
 - aimed at **predicting new or future observations** (without necessarily explaining how)
 - relies on **big data**
 - prevalent in fields like natural language processing, **bioinformatics**, etc.. In **epidemiology**, there is more of a mix causal explanation & empirical prediction

- **NOTES:**

- ✓ “Prediction” does not necessarily refer to future events, but rather to *future* datasets that were previously unseen to the algorithm

...stay tuned for next chapter on ML

