STATISTICS & ML WITH R

Mapping causal & predictive approaches

2024

M. Chiara Mimmi & Luisa M. Mimmi

DAY 4 – LECTURE OUTLINE

Mapping causal & predictive approaches

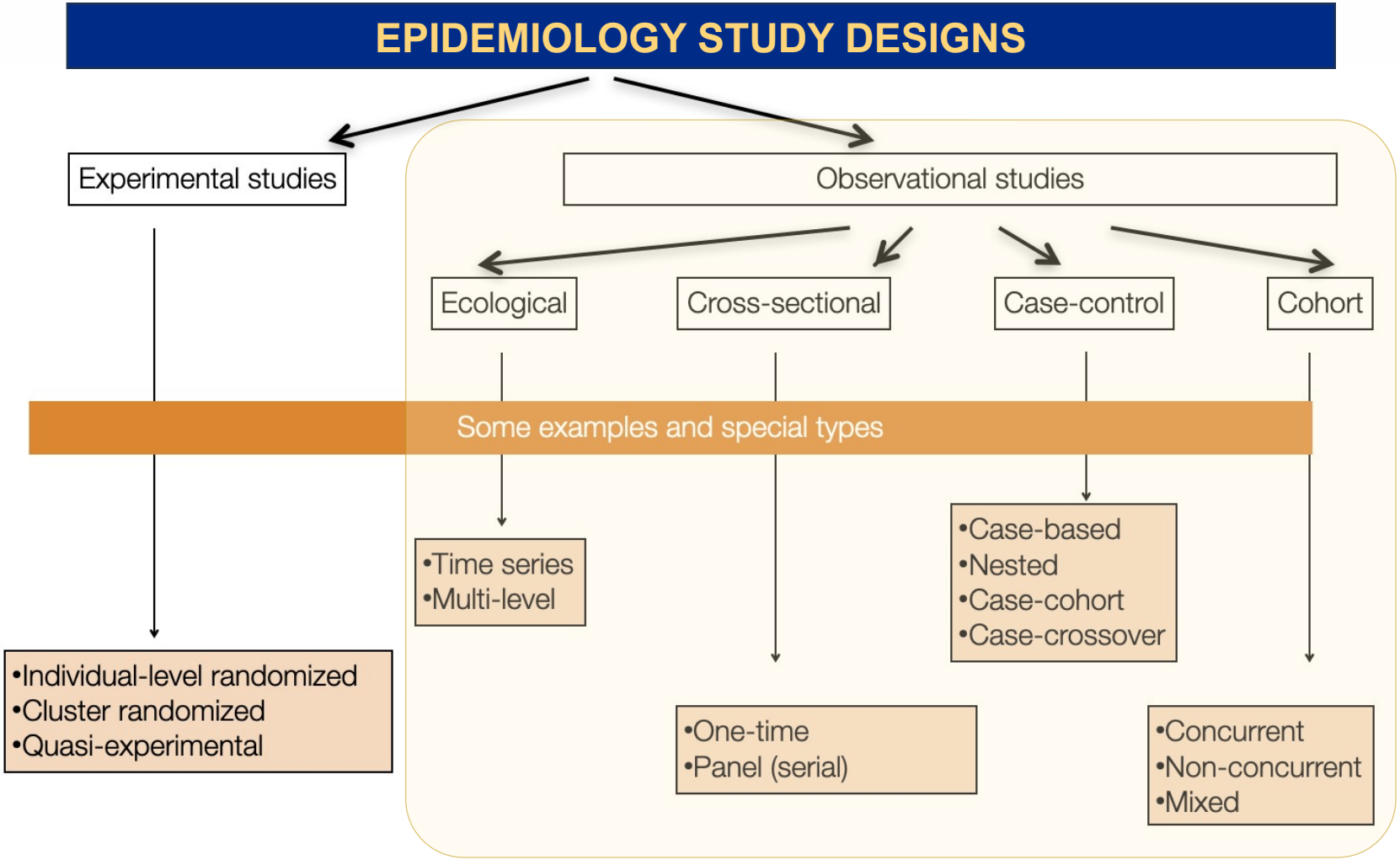
- Illustrating different study designs
- Learning the vocabulary of causal analysis
- Visual understanding of causal pathways, including:
  - Collider variables
  - Confounder variables
  - Mediator variables
- Learn how to address causal pathways in modeling, based on the research question
- Defining causal outcomes and commonly used “estimands” (ATE, ATT, ATU)
- Understand proper statistical methods to estimate ATE, ATT, ATU based on research question and target group
- Introducing Machine Learning (ML)
  - purpose
  - key algorithms’ categories

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

Different goals 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

A framework for CAUSAL ANALYSIS

Key terminology and visual causal maps

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 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

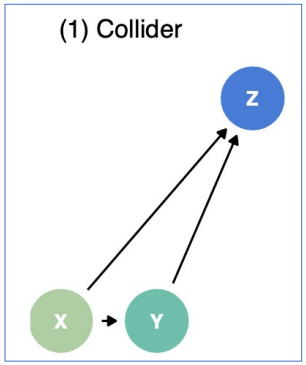
The conceptual framework for causal analysis (2/3)

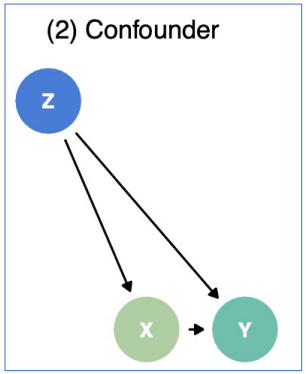
- **Fundamental vocabulary** ("*tricky ones*" 😉):
  - **Bias** = systematic error that can occur at different stages of the study: *data collection*, *analysis* or *interpretation* of the causal relationship exposure-outcome.
    - **Selection bias** = both the exposure and the outcome affect whether 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)
    - …

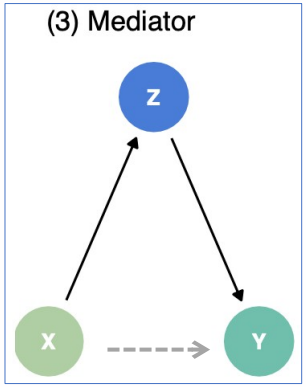
Check out this list of all types of bias:
https://quantifyinghealth.com/list-of-biases/

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 of interest*
  - EXAMPLE: expected value of the difference in potential outcomes across all individuals

- The **estimator** is the method by which we approximate this estimand using data ("recipe")
  - EXAMPLE: in randomized controlled trial, our estimator could just be the average outcome among those who received the exposure A minus the average outcome among those who receive exposure B

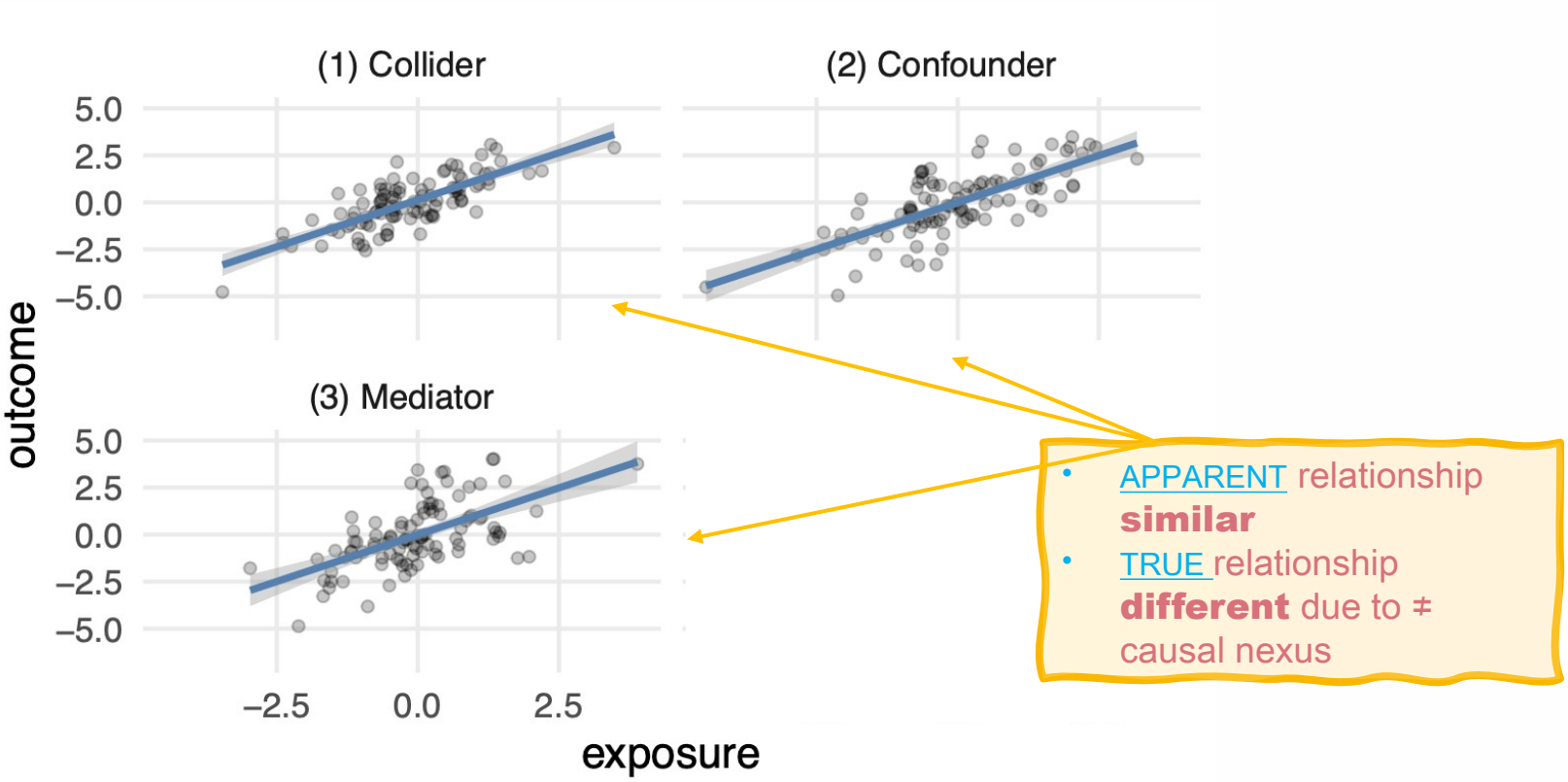
- The **estimate** is the value we get when we plug our data into the estimator
  - EXAMPLE: randomized controlled trial, our estimator could just be the average outcome among those who received the exposure A minus the average outcome among those who receive exposure B

Visualizing causal maps

A helpful tool in guiding statistical modeling

Typical challenges in estimating causal effects: visual intuition

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



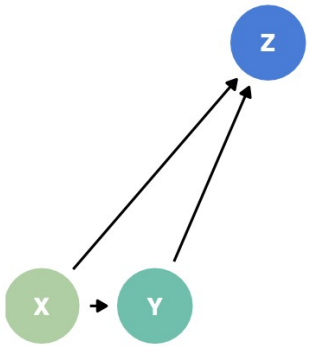
Source: Barrett, M., McGowan, L. D., & Gerke, T. (2024). *Causal Inference in R*. Retrieved from https://www.r-causal.org/

Typical challenges in estimating causal effects: visual intuition

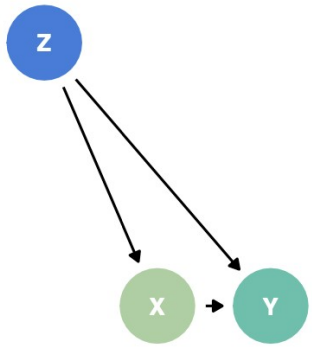
- Directed acyclic graphs (DAGs) can offer visual intuition of the causal nexus at play in the 3 datasets. Failure to adjust models to these situation leads to BIAS
  - X is some continuous exposure of interest, Y a continuous outcome, and Z a known, measured factor

(1) a "**COLLIDER**" is caused by both X and Y (it inadvertedly connects the 2). E.g.:
- X = sodium intake
- Y = systolic blood pressure
- Z = urinary protein excretion

(1) Collider



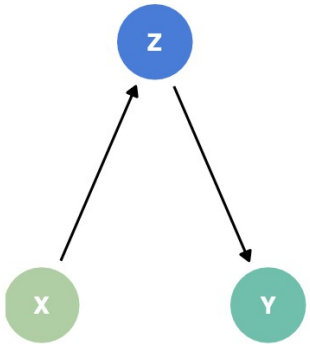
(2) Confounder



(2) a "CONFOUNDER" causes both X and Y. E.g.:
- X = smoking
- Y = lung cancer
- Z = alcohol (consumers also tend to be smokers)

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

(3) Mediator



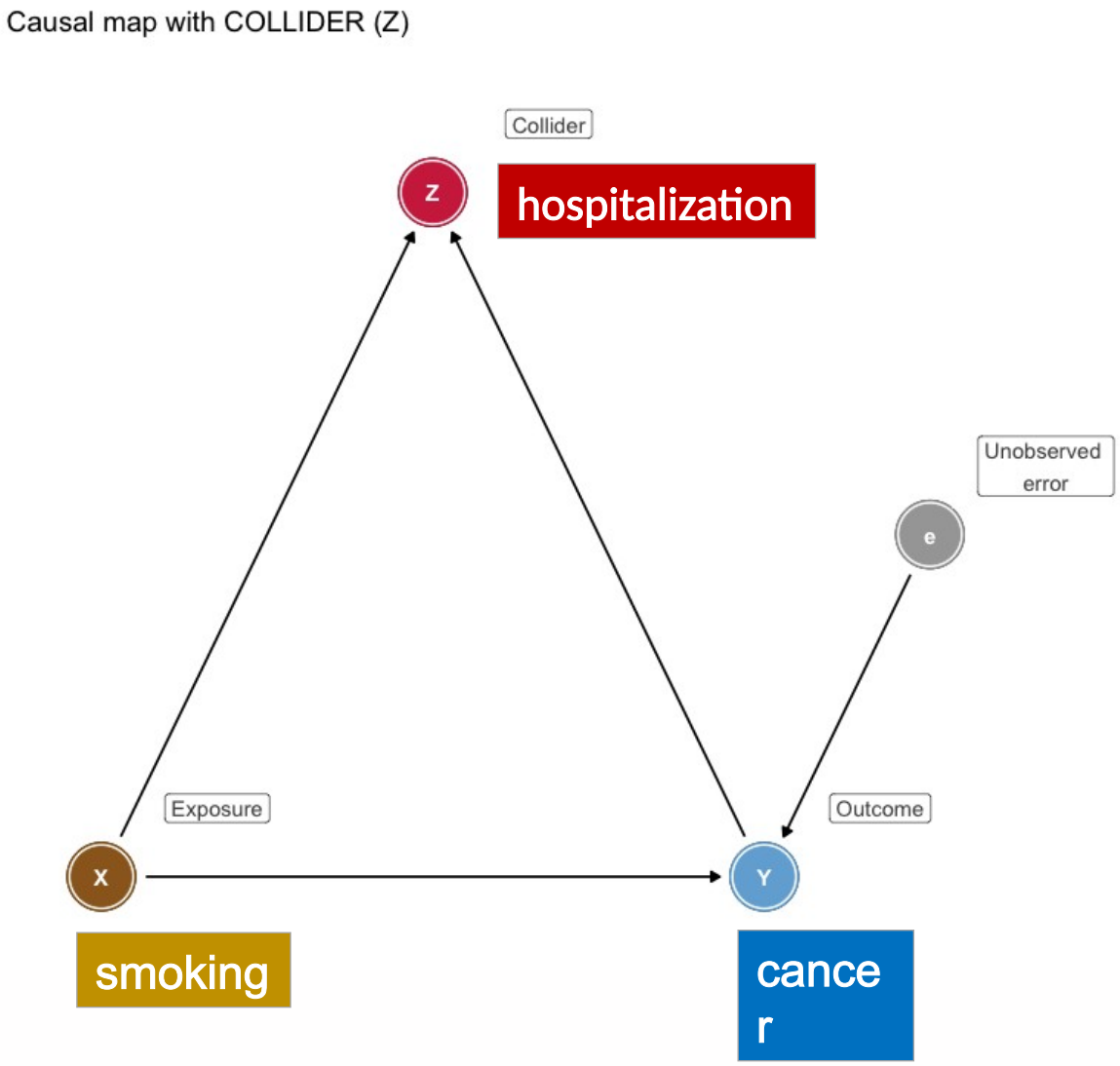
we'll revisit this later in multivariate regression…

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

How to address causal pathways in modeling

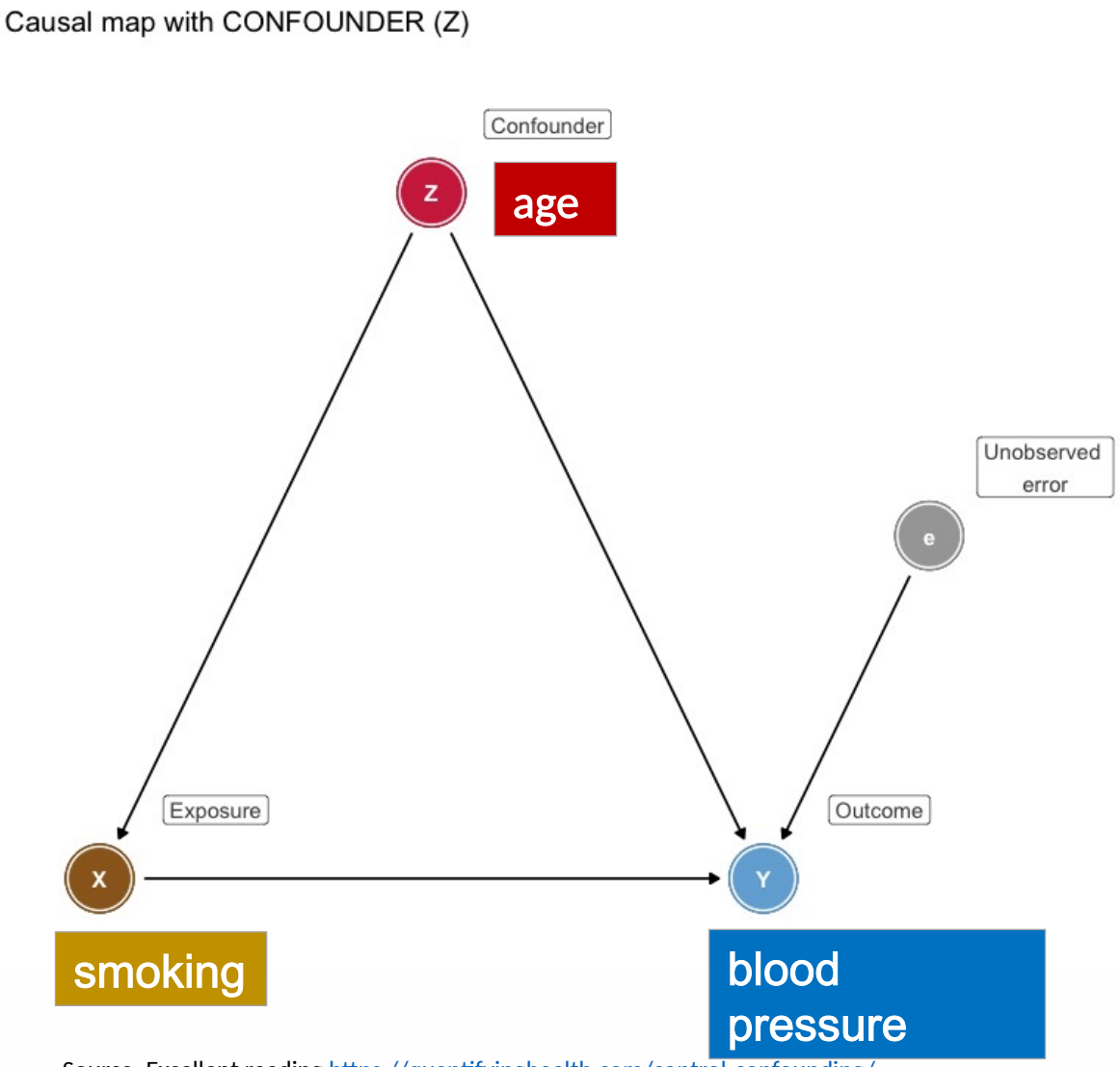
This will be re-visited through practical examples in Lab 4

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



- 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

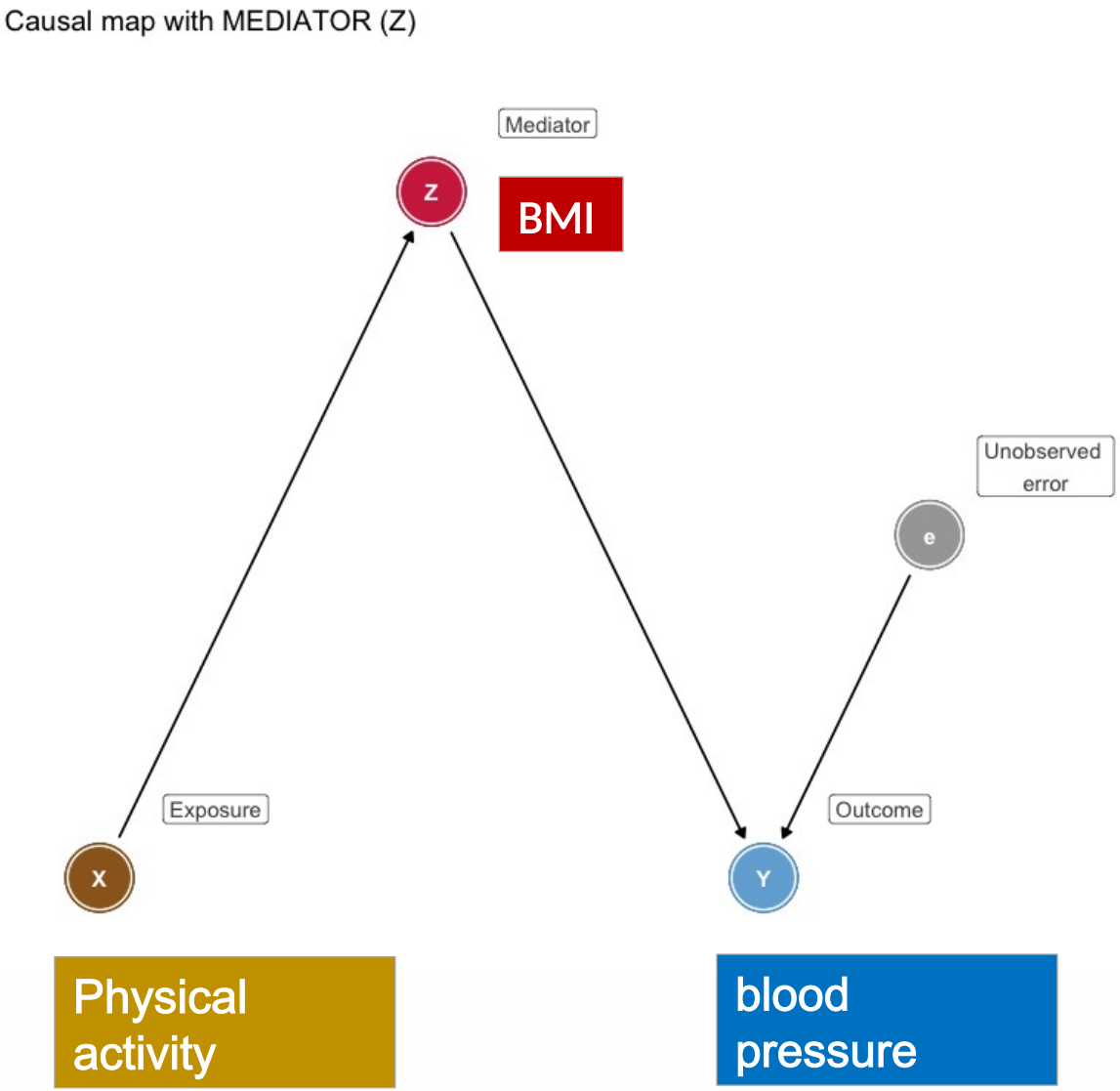
How to deal with confounder *(common cause)* when modeling?



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

- We must *control for* a confounder, so we reduce bias:
- 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 counfounder by exposure*)
  - **Instrumental variable estimation**

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



- 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*

Measuring causal outcomes of interest

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

Defining potential outcomes at the *subject level (experimental unit)*

- **NOTATION:**
  - and are the potential outcomes in the *absence* and *presence* of treatment
  - for patient i in a study on a new drug on blood pressure,
    -
    - = with takes new drug

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

  where treatment is

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

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

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

Defining potential outcomes at the *subject level (experimental unit)*

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

]

  - This refers to the avg of the differences conditionally on the fact that both groups "received" the treatment ("")
  - is essentially the counterfactual for 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

]

  - This time we seek the Avg of the differences ("") conditionally on the fact that both groups were "assigned" to the treatment
  - is essentially the counterfactual for 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 vriable
- outcome of interest is
- **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

Ex: Does hospitalization (T) increase health (Y) ?	
(ATE)	Avg health of hospitalized group – avg health of NOT hospitalized group
(ATT) + …	Avg health of treated group – [counterfactual] avg health of treated group IF NOT hospitalized
(Selection bias) … + *(hospitalized have worse than non hospitalized)*	Difference in [counterfactual] avg health of treated group IF NOT hospitalized - those who were NOT hospitalized

EXE. potential causal outcomes (depends on patients characteristics)

	Confounder	*Treatment*	*Unobservable*			*Realized*
	Age	**Treated**	**Potential outcomes**		**ICE or δ_i***	**Outcome**
ID	Z_i	X_i	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	Y_i
1	Old	1	80	60	20	80
2	Old	1	75	70	5	75
3	Old	1	85	80	5	85
4	Old	0	70	60	10	60
5	Young	1	75	70	5	75
6	Young	0	80	80	0	80
7	Young	0	90	100	–10	100
8	Young	0	85	80	5	80

* ICE = individual causal effect

(ATE decomposition)

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

+ = 4.1667

After stratification based on the confounder we get a very close ***approximation of the ATE***

Other ways to deal with confounders

- *GIVEN THAT IN REAL LIFE WE NEVER HAVE THE ALTERNATIVE POTENTIAL OUTCOMES FOR EACH SUBJECT i , HOW DO WE DEAL?*
  - **STRATIFYING** by Age was easy, but what if there is >1 confounder? What if it is continuous?

- *WE'VE GOT TO DO SOMETHING ELSE TO GET TO COMPARABLE GROUPS:*
  - **MATCHING** methods = dropping units from the sample or partitioning units into pairs or subclasses (e.g., *Propensity Score Matching*)
  - **WEIGHTING** methods = weighting the units so that the weighted distributions are similar between treatment groups (e.g., *Inverse probability weighting*)
- After *adjustment* the treatment effect is estimated in the resulting sample (incorporating the weights resulting from the matching or weighting)

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?*
  - ~~Some methods, such as PSM in its most commonly used form, cannot target the ATE, and so are inappropriate when the ATE is of interest!~~

Choosing the estimands based on the research question

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

EXE [see LAB 4]: how to exploit *"*matched*"* untreated observation to estimate the ATT

- …. ANDREW HEISS ESEMPIO DI PSM (uso il suo che e' troppo bello!!!!

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