Workshop 005: Developing high-capacity predictive models and intersections with causal inference. (Causal 2)

International Conference on Health Policy Statistics

1/9/2023

Evan Carey, PhD

evan.carey@va.gov

evan.carey@cuanschutz.edu

https://github.com/evan-paul-carey/ML_and_causal_inference_workshop

Instructor: Evan Carey

- MS Applied Biostats, PhD Epidemiology
- Data Scientist with VA hospital system
- Assistant professor of informatics @ Colorado School of Public Health
- Research interests:
  - Interest in answering useful questions using national healthcare data.
  - Interest in coding / algorithmic challenges
  - Clinical topics:
    - Chronic Pain
    - Mental Health Care
    - Operations evaluations

Audience questions…have you…?

- Developed a predictive model before?
- Contrasted multiple different predictive models then picked the best one?
- Implemented propensity score analysis?
- Constructed a causal diagram to select adjustment variables?

Artificial intelligence, machine learning, deep learning, block chain…

- What are some exciting things you have heard about associated with machine learning/AI?

Artificial intelligence, machine learning, deep learning, block chain…

- What are some exciting things you have heard about associated with machine learning/AI?

- Machine learning (ML) can potentially cover a wide variety of topics.

  - Self-driving cars
  - AI generated artwork, music, essays…
  - Chatbots!
  - Clinical decision support applications (image analysis)

Defining Machine Learning

‘Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around those functions’
- Deep Learning Book (https://www.deeplearningbook.org/contents/ml.html)

- Using data, develop a ‘learning algorithm’ (our model).
- Often the focus is prediction of an outcome, given inputs.
- Finding patterns in the data versus finding generalizable trends in the data.

Example – probability of rehospitalization.

- Let's start with a simple example to understand the fundamentals of ML.

You are tasked with creating an algorithm that predicts the probability a 68-year old patient with diabetes will be re-hospitalized within 30 days.

You don't have any data on your patients. But you do know the national average rehospitalization rate generally is 15.9%*.

Task: What is your best guess of the probability of rehospitalization for this patient?

*The Revolving Door: A Report on U.S. Hospital Readmissions. RWJF, 2013, http://www.rwjf.org/en/library/research/2013/02/the-revolving-door--a-report-on-u-s--hospital-readmissions.html

Example – probability of rehospitalization

- You have found more information about your patients from last year – the average rehospitalization rate among your patient population was 19%.

- Given this new information, what is your updated estimate of the probability this 68-year-old patient with diabetes will be rehospitalized?

Example – probability of rehospitalization.

- We have more information from last year's patients in our hospital…

Age	Rehosp
18-40	5%
40-60	9%
60-70	24%
70+	35%

Diabetes	Rehosp
No	17%
Yes	21%

- Q1: Recall our 68-year-old patient with diabetes. **What is your updated prediction for the probability of rehospitalization within 30 days?**
- Q2: Which of these two patients has a higher probability of being rehospitalized?
  - 55-year-old patient with diabetes
  - 70-year-old patient without diabetes

Classification / Regression / Clustering

- Classification
  - Predicting class membership (or probabilities) among distinct classes.
    - Death (Yes / No)
    - Risk Strata (Low / Medium / High)
- Regression
  - Predicting a continuous summary statistic (like the mean)
    - Hospital costs (Mean, median, 90^{th} percentile)
- Clustering
  - Identifying clusters in our data.
  - Project data into smaller dimensionality.
  - *Clustering can be discrete or continuous.

Defining a learning algorithm.

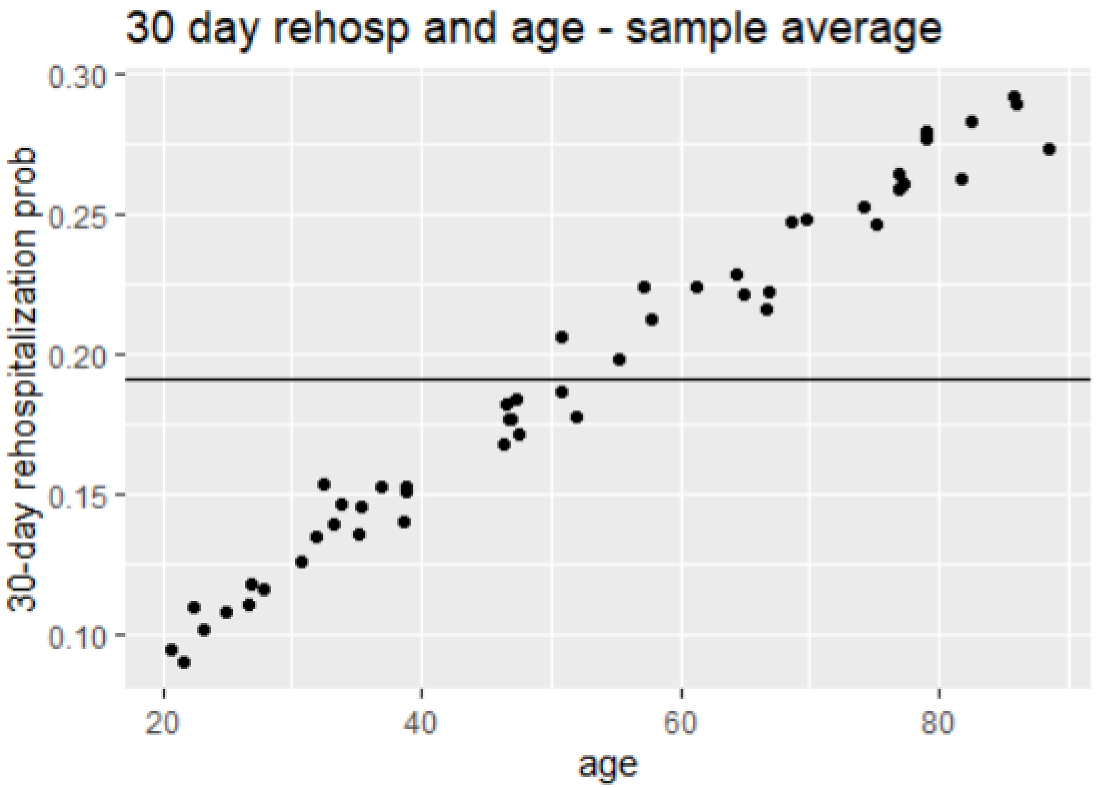
- Using data, develop a ‘learning algorithm’ (our model).

- What do we need to develop a learning algorithm?
  - Data
  - Model
  - Cost function

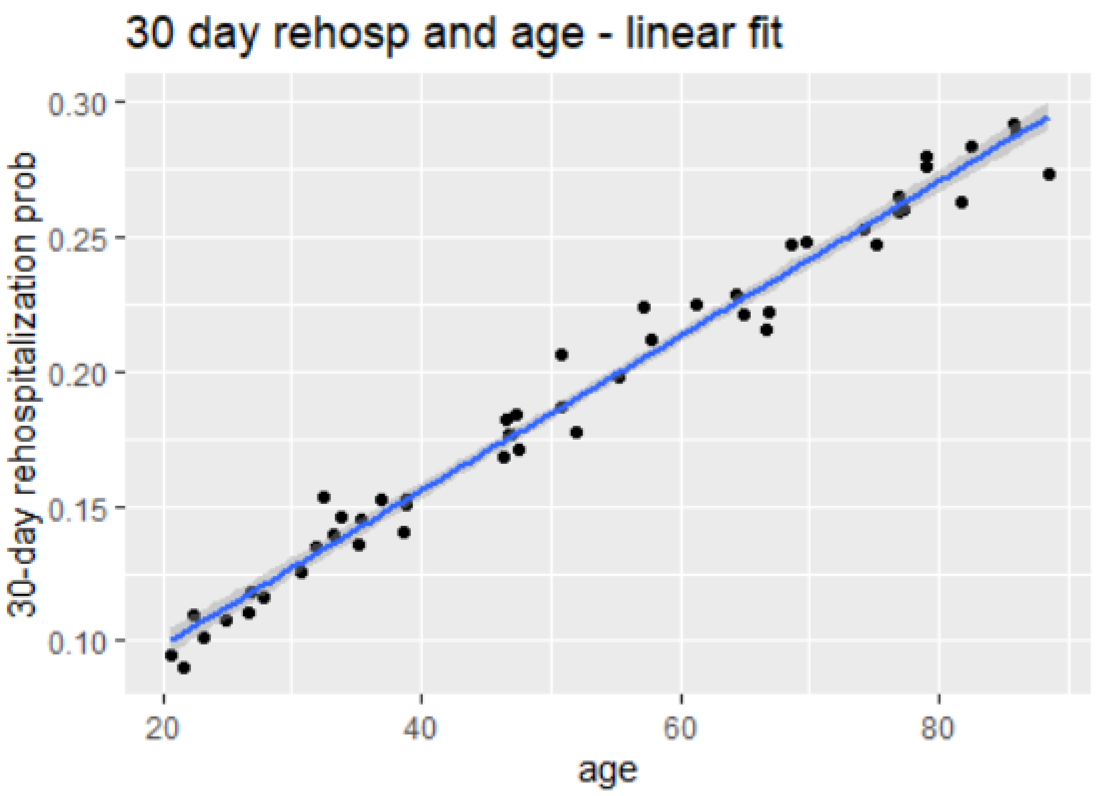
Central Challenge to ML

- The algorithm must perform well on *new* data it has never seen before
  - Next years data
  - New hospitals
  - New patients
- This concept is called generalization.

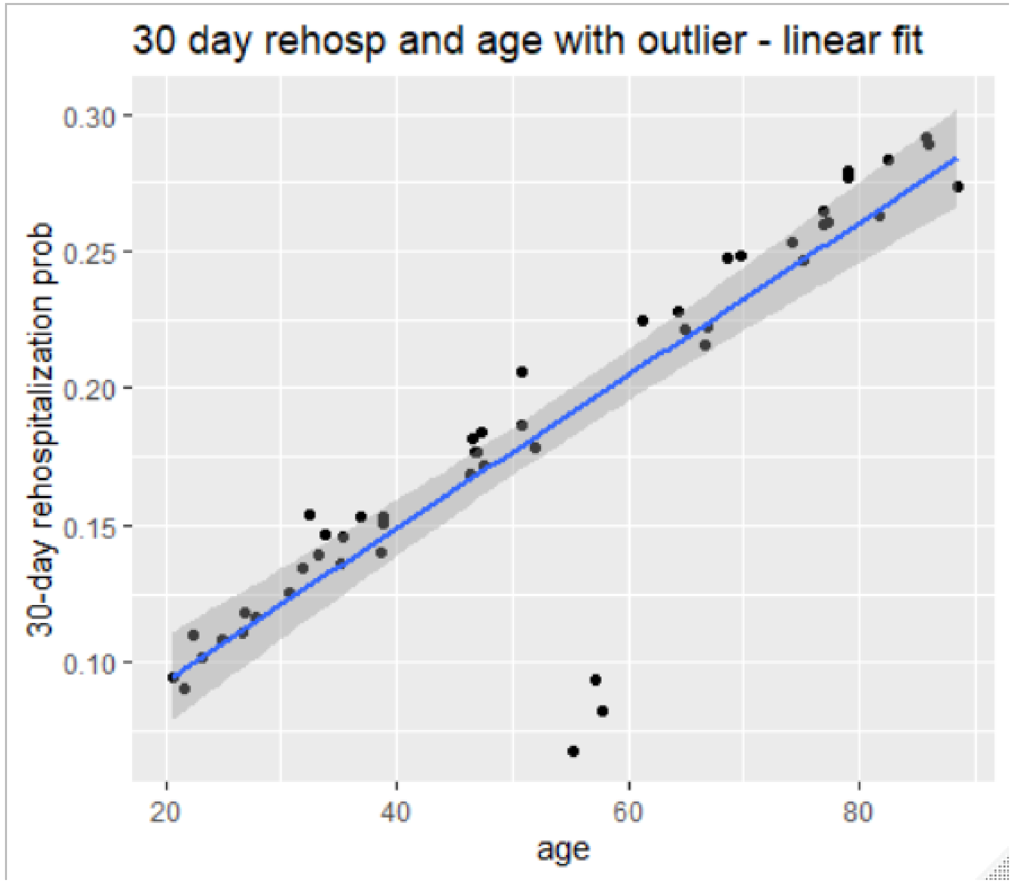
Probability of rehosp cont.



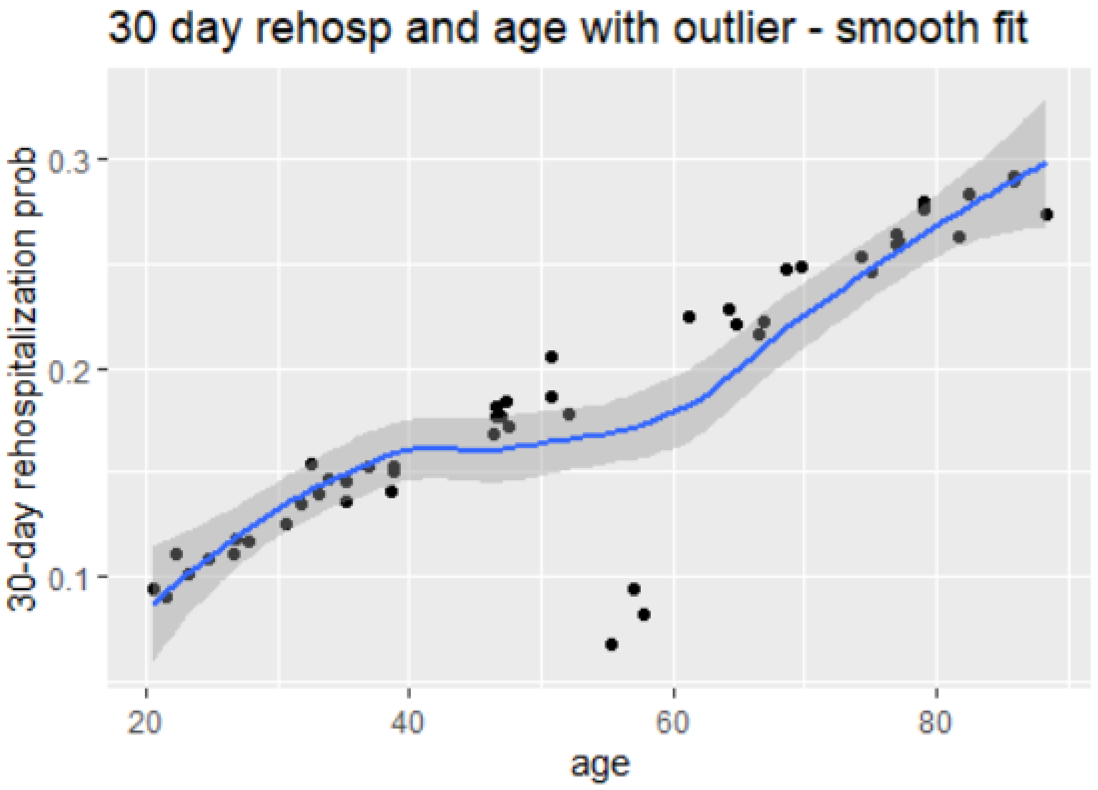
Probability of rehosp cont.



Probability of rehosp cont.



Probability of rehosp cont.



Under/overfitting and out of sample data

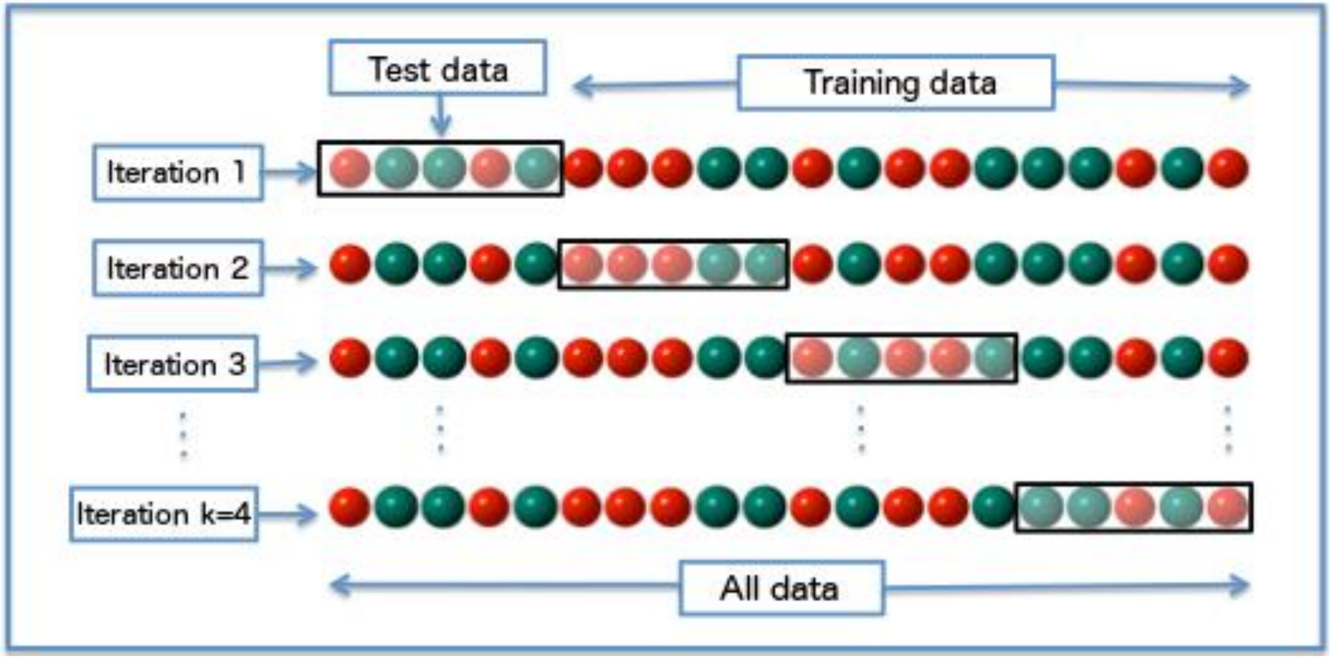
- Given the data we have, how good is our model?
  - This is really just optimization.
  - **Training error** is how well we fit the training data.
  - Increased performance here sometimes decreases performance outside of our sample of data (overfitting)
- In ML, we target **generalization error**
  - Generalization error is how well our algorithm fits data outside our sample.
  - But we don’t have any data outside our sample…
  - Can we pretend we do?

Validation approaches

- If the new data does not come from the same data-generating distribution as the observed data, **<u>full stop</u>**.
- If we assume the new data comes from the same data-generating distribution, then we can implement validation approaches.
  - Create multiple random samples from the data we have
  - Call one ‘training data’ and one ‘Validation’ data.
  - We usually split into training / validation / testing (3 splits).
- Optimization goal:
  - Minimize training error (high error = underfitting)
  - Minimize gap between training and testing error (big gap = overfitting)

Creative Validation Approaches

- Splitting by clusters.
  - Split by year, region, etc…
- Cross validation.

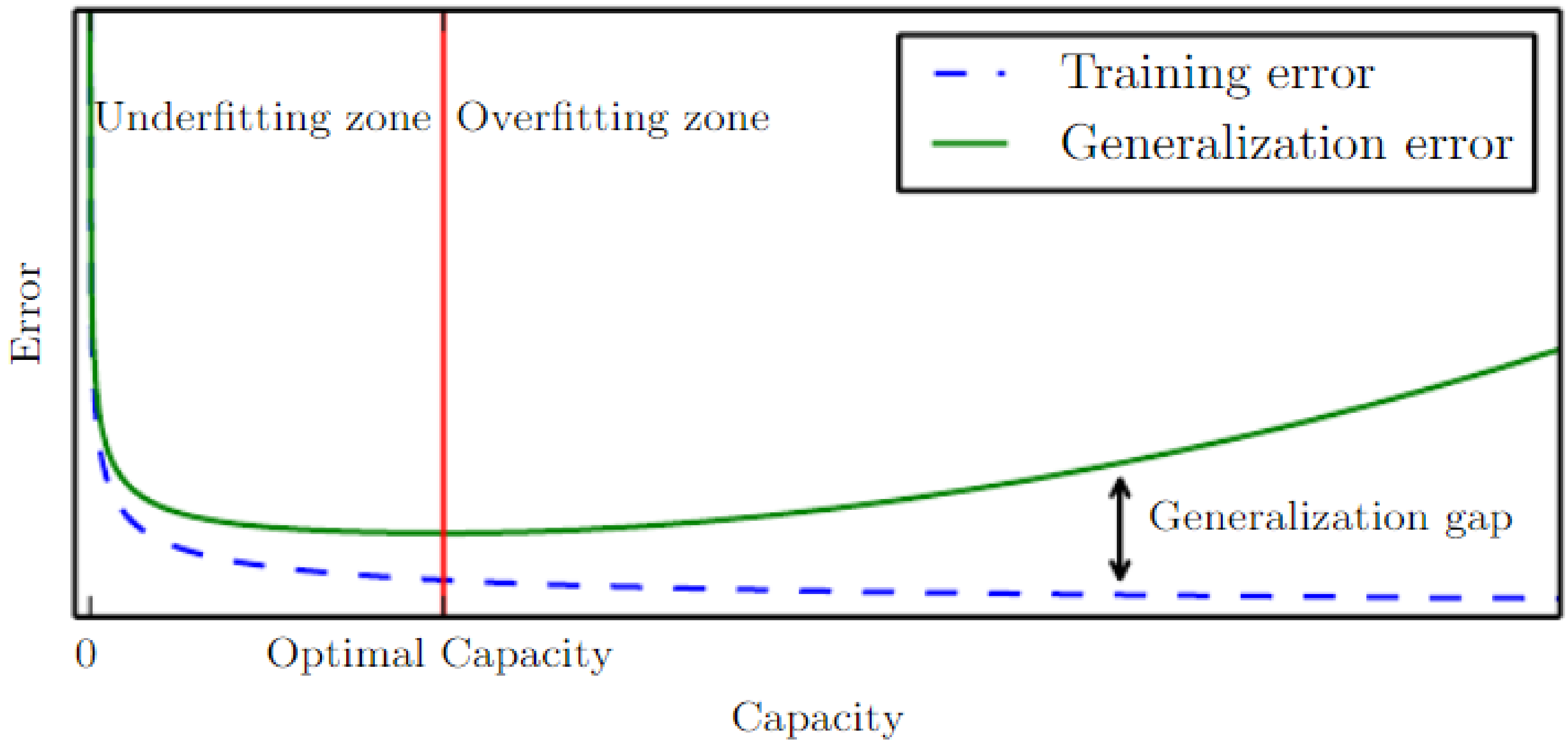


By Fabian Flöck - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=51562781

Balancing Underfitting / Overfitting

- We can balance under and overfitting by making our model more/less complex. This is called ‘model capacity’
- Increasing model capacity generally allows the model to fit more nuanced relationships.
  - <u>In linear modeling</u> – add more inputs, consider non-linear terms (polynomials), consider interactions…
- What is the downside of increased <u>model capacity</u>?

Balancing Underfitting / Overfitting



Deep learning book, figure 5.3
https://www.deeplearningbook.org/contents/ml.html

Model Parsimony

> Among competing hypotheses that explain known observations equally well, choose the simplest one.
> - Occam's razor (c. 1287-1347)

- Model Parsimony
  - Simpler is always better (as long as we still maximize generalization error…)

- How do we implement model parsimony?

Regularization

- Hard code preferences into the model.
  - I prefer Beta’s close to or equal to zero (parsimony)
  - However, if I find enough support for a relationship, it can stay.
  - How to I hard code that into my model?
  - What is an example you have learned of this in ML?

> ‘Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not it’s training error.’
>
> \- Deep learning book, p 117

Training / Validation / Testing splits.

- Training data
  - Optimize the model to minimize **training error**
- Validation data
  - Optimize hyperparameters to minimize **generalization error**
  - If we implement cross validation, this step is internalized to the model optimization.
- Testing data
  - Estimate **generalization error**
  - How good will our model perform on previously unseen data?
  - Usually optimistic.

Focus on predictive model development (supervised ML)

- Given all the data I have observed, what is the best prediction I can make for my outcome (conditional on all data I have observed)?

Transition to RMD notebooks

Back to Causal Inference

Credit to Miguel Hernan’s book

Much of this section is directly inspired by the publications and recent book of Miguel Hernan, which can be accessed here:

https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf

Some Causal Notation – Potential Outcomes

Let's cover some notation we will use from here forward (the same a used in the Hernan book).

Dichotomous treatment variable (the 'action'): *A* (1: treated, 0: untreated)

Dichotomous outcome variable: *Y* (1: outcome yes, 0: outcome no)

Vector of confounding variables *L*

$Y^{a=1}$ (Y under treatment *a*=1): The outcome we observe when treatment a=1

$Y^{a=0}$ (Y under treatment *a*=0): The outcome we observe when treatment a=0

$Y^{a=1}$ and $Y^{a=0}$ are referred to as 'potential outcomes'.

Causation versus Association

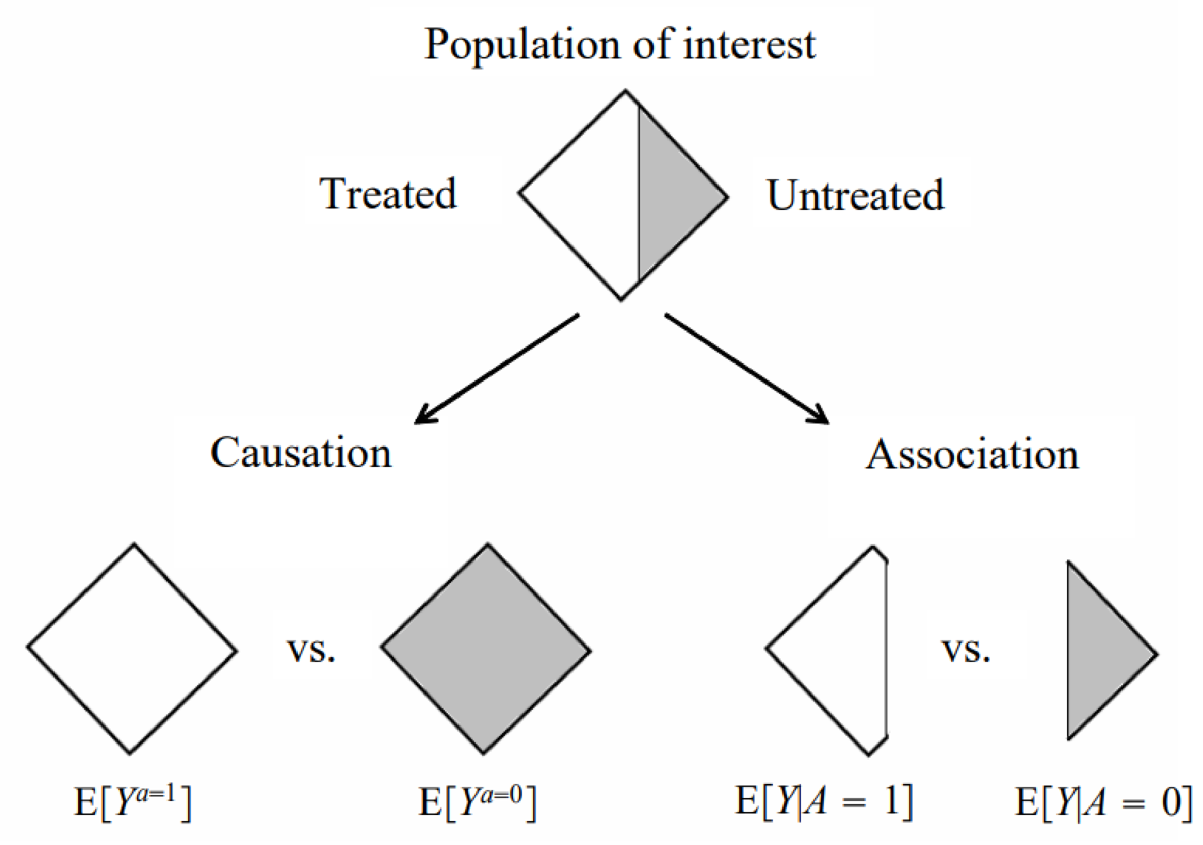
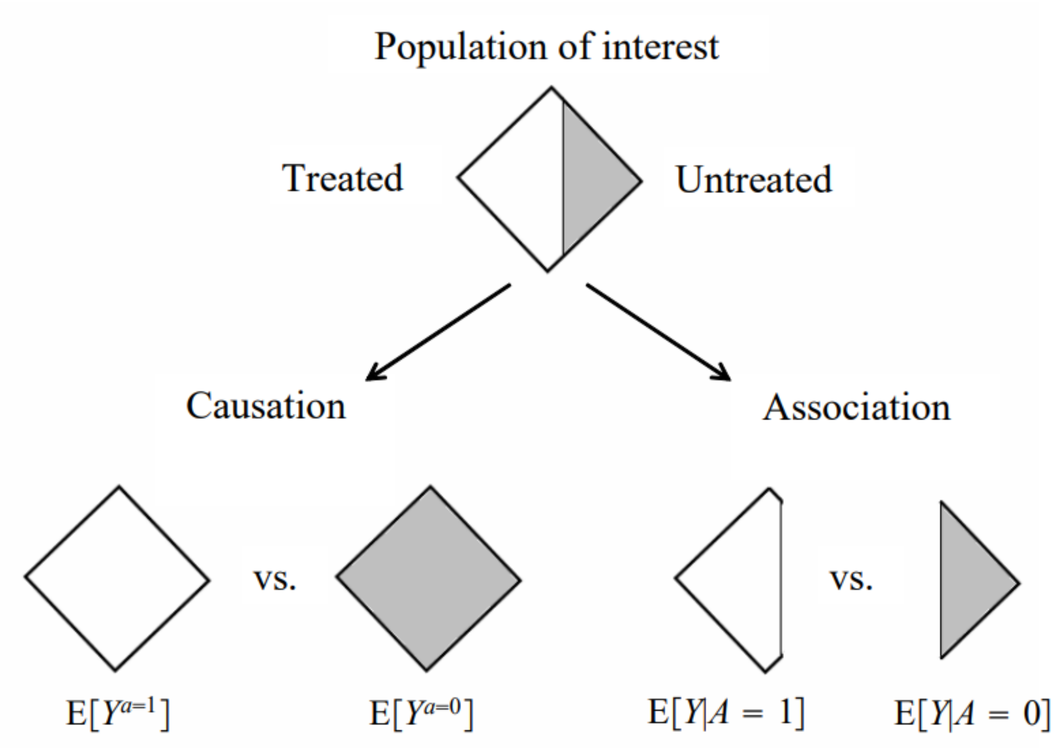


Figure 1.1 Hernan 2021, https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf)

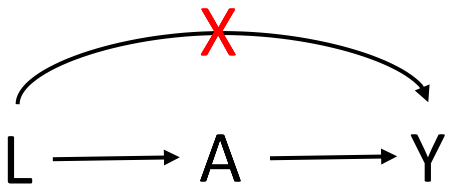
Visual for this idea



- We are basically saying the shaded triangle on the right is a good representation of the shaded triangle on the left (and same for unshaded).
- By good representation, we specifically mean the expected value of the outcome conditional on A = 1 or A = 0.

Figure 1.1 Hernan 2021, https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf)

Outcome regression – interrupting the L-Y connection.



- We can remove potential bias by interrupting the L → Y connection.
- Traditional 'adjusted regression' does this!
- However, we can implement higher capacity models than simple linear regression to estimate these outcome probabilities….

$$\mathrm{E}\left[Y^{A} \mid A = 1\right] \text{ for all values of A}$$

$$\mathrm{E}\left[Y^{A} \mid A = 0\right] \text{ for all values of A}$$

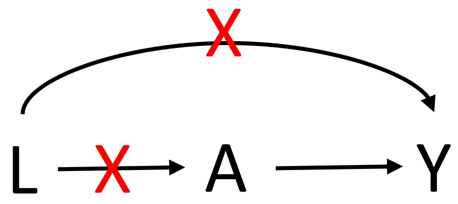
Probability of treatment – interrupting the L-A connection.

$$\begin{tikzcd} L \arrow[r, "X"] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

- We can remove potential bias by interrupting the L → A connection.
- This is propensity score analysis (inverse probability of treatment weighting).
- We can implement higher capacity models than simple logistic regression to estimate these treatment probabilities….

$$\mathrm{E}[A \mid L]$$

What if we do both? Interrupting the L→ A and the L → Y connection.



- We can implement both methods simultaneously.
- This is a so-called doubly robust estimator.
- If either model is correctly specified, the resulting estimate is unbiased. (one model can be wrong).