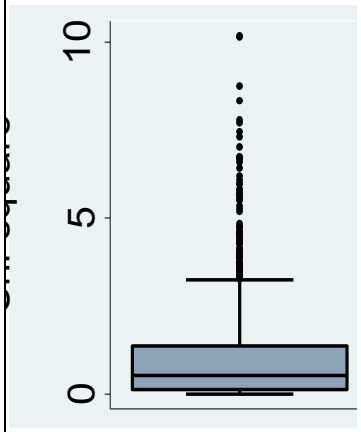
Biostat 202: Opportunities and Challenges of "Big Data" - Causal Inference in 'Big Data'

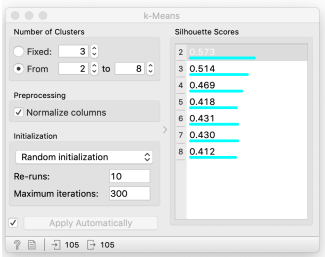
Aaron Wolfe Scheffler,
Division of Biostatistics,
Department of Epidemiology and Biostatistics

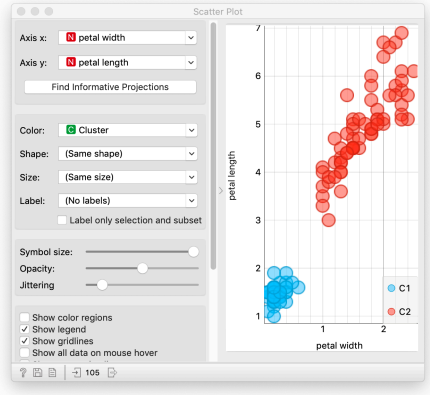


Review

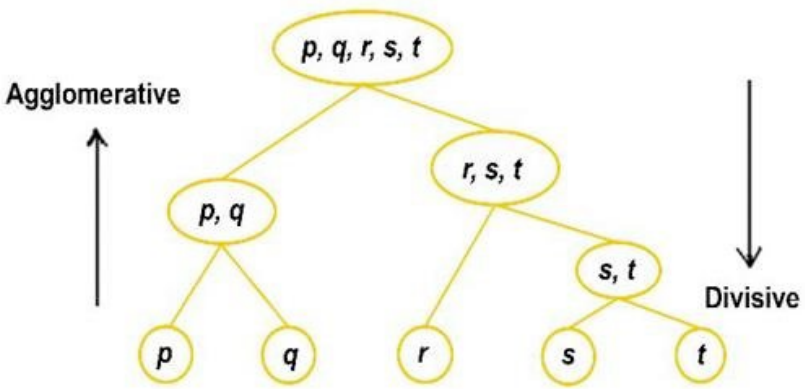
Clustering

1. k-means



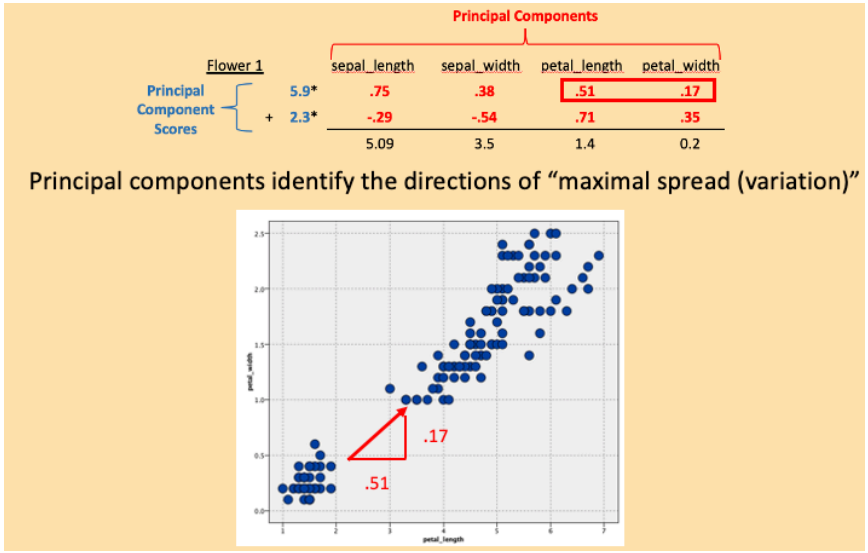


2. heirarchical

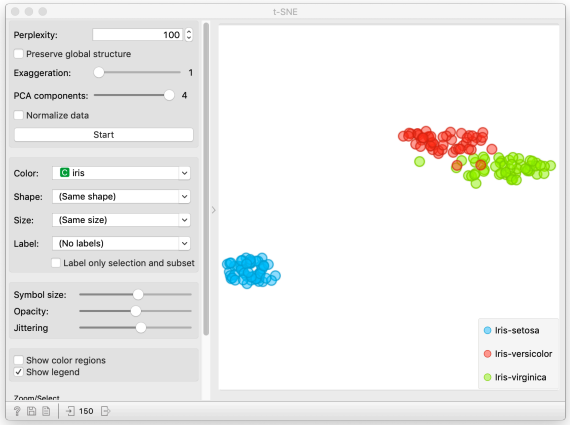


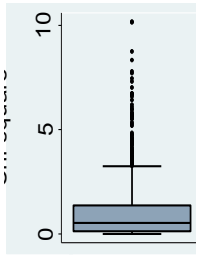
Dimension Reduction

3. PCA



4. t-SNE





Outline

- Introduction: prediction versus causation
- Threats to causal conclusions
- Causal methods in (very) brief
- Interpretability of machine learning methods
- Some uses of big data and machine learning in causal inference
- Summary

Outline

- **Introduction: prediction versus causation**
- Threats to causal conclusions
- Causal methods in (very) brief
- Interpretability of machine learning methods
- Some uses of big data and machine learning in causal inference
- Summary

Prediction versus causality

- Data-mining and machine learning algorithms are only based on associations.
- Often interested in causation, i.e., understand the cause of illness or develop interventions that cause improvement.
- Correlation does not imply causation
  - https://www.tylervigen.com/spurious-correlations
- Often senior researchers have lapses and over-interpret correlations causally
- The (social) media frequently confuses causation and correlation

Which of these are prediction versus causation questions?

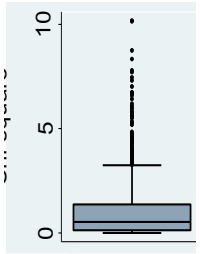
1. Effect of Abaloparatide on fractures in postmenopausal women.
2. Estimating fetal risk of a disease using genetic screening.
3. Effect of cerebral protection device on brain lesions following aortic valve implantation.
4. Estimation of temporal trends in preterm birth rates and their association with obstetric interventions.
5. Survival for children with trisomy 13 and 18.

Prediction vs Causation

Bona fide prediction problems:

- Early detection of cancer.
- Identifying high risk patients for closer monitoring.
- Prediction of disease epidemics.

Key aspects of prediction: Not concerned with causal interpretation of the effect of variables included in the model or interpreting presence or absence of the variable in the model causally. No interest in p-values – focus on prediction accuracy.



Key words

- effect, intervention = causal
- estimate, predict, associate = prediction

Blurring the line?

JAMA Internal Medicine | Original Investigation

Outcomes Associated With Resuming Warfarin Treatment After Hemorrhagic Stroke or Traumatic Intracranial Hemorrhage in Patients With Atrial Fibrillation

IMPORTANCE The increase in the risk for bleeding associated with antithrombotic therapy causes a dilemma in patients with atrial fibrillation (AF) who sustain an intracranial hemorrhage (ICH). A thrombotic risk is present; however, a risk for serious harm associated with resumption of anticoagulation therapy also exists.

Association!

OBJECTIVE To investigate the prognosis associated with resuming warfarin treatment stratified by the type of ICH (hemorrhagic stroke or traumatic ICH).

Conclusions Spontaneous hemorrhagic stroke and trauma-induced ICH confer different prognoses in patients with AF, and recommendations on resumption of warfarin treatment should consider this difference.

Suggested causal link?

JAMA Intern Med. 2017;177(4):563-570. doi:10.1001/jamainternmed.2016.9369
Published online February 20, 2017.

Outline

- Introduction: prediction versus causation
- **Threats to causal conclusions**
- Causal methods in (very) brief
- Interpretability of machine learning methods
- Some uses of big data and machine learning in causal inference
- Summary

Levels of Evidence

Why are these types of studies rated so lowly?

From the Centre for Evidence-Based Medicine, Oxford

For the most up-to-date levels

Therapy/Prevention/Etiology

1a:	Systematic reviews (with trials
1b:	Individual randomized c erval)
1c:	All or none randomized
2a:	Systematic reviews (with ho eity) of t studi
2b:	Individual cohort study or ra co ls (e.g. <80% follow-up)
2c:	"Outcomes" Research; ecological studies
3a:	Systematic review (with homogeneity) of case-control studies
3b:	Individual case-control study
4:	Case-series (and poor quality cohort and case-control studies)
5:	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"

Electronic health record based study

Note! The strength of evidence depends on the data/sample, not the type of model utilized!

Example: Ann Landers/Newsday surveys

Many years ago, the help columnist, Ann Landers, asked her readers "If you had it do to over again, would you have children?" She received about 10,000 responses, 70% of which said no.

Shortly thereafter, Newsday commissioned a nationwide random telephone survey.

9% said no.

Samples: **10,000 write-ins; 1,373 random telephone contacts**

Population: **U.S. households with children**

Moral: **Big data doesn't overcome selection bias.**

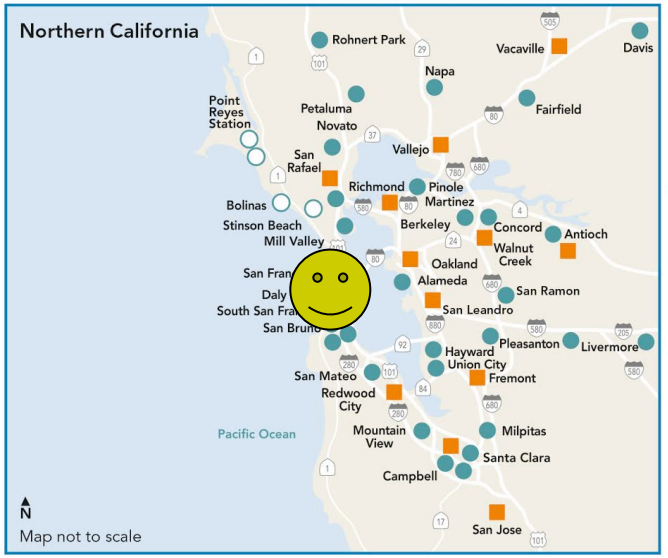
Example: More realistic

Suppose you are going to conduct a research study using electronic health records and want the results to generalize to all similar patients in the Bay Area?

Would you prefer to use data from UCSF or Kaiser? Why?

UCSF ~ Ann Landers

Kaiser ~ Random telephone survey



Kaiser Permanente facilities near you

- Kaiser Permanente medical centers (hospital and medical offices)
- Kaiser Permanente medical offices

UCSF

Example: Phototherapy and Cancer

- Phototherapy (exposing the infant to bright fluorescent light) is widely used to treat jaundice in newborns. Jaundice is caused by a build up of bilirubin levels before the newborn's liver starts working. High levels can cause brain damage. The light breaks down the bilirubin and allows it to be excreted.

- Does phototherapy cause cancer?

- Studied about 500,000 infants born in Kaiser Northern CA between 1995 and 2011.

Newman, et al Pediatrics, 2016

Example: Excerpt from Table 1

TABLE 1 Demographic and Clinical Characteristics of Infants Exposed to Phototherapy and 2 Groups of Unexposed Infants

	Phototherapy	No Phototherapy	
		≥1 TSB −3 to 4.9 mg/dL From Phototherapy Threshold	No TSB −3 to 4.9 mg/dL From Phototherapy Threshold
Number of infants	39 403	62 592	397 626
Year of birth ≥2003, %	79.2	66.1	50.8
Male, %	56.0	53.8	50.2
Down syndrome, %	0.53	0.21	0.05
Other chromosomal anomaly, %	0.3	0.2	0.1
Congenital anomaly, %	3.8	2.4	1.8
Birth wt, mean ± SD, g	3260 ± 589	3354 ± 526	3467 ± 493
Birth wt ≥4500 g, %	1.9	1.9	2.2
Gestational age, mean ± SD, wk	38.2 ± 1.7	38.6 ± 1.6	39.3 ± 1.3
5-minute Apgar <7, %	1.6	0.8	0.8

Concerns comparing the phototherapy and non-phototherapy groups?

Example: Phototherapy and Cancer

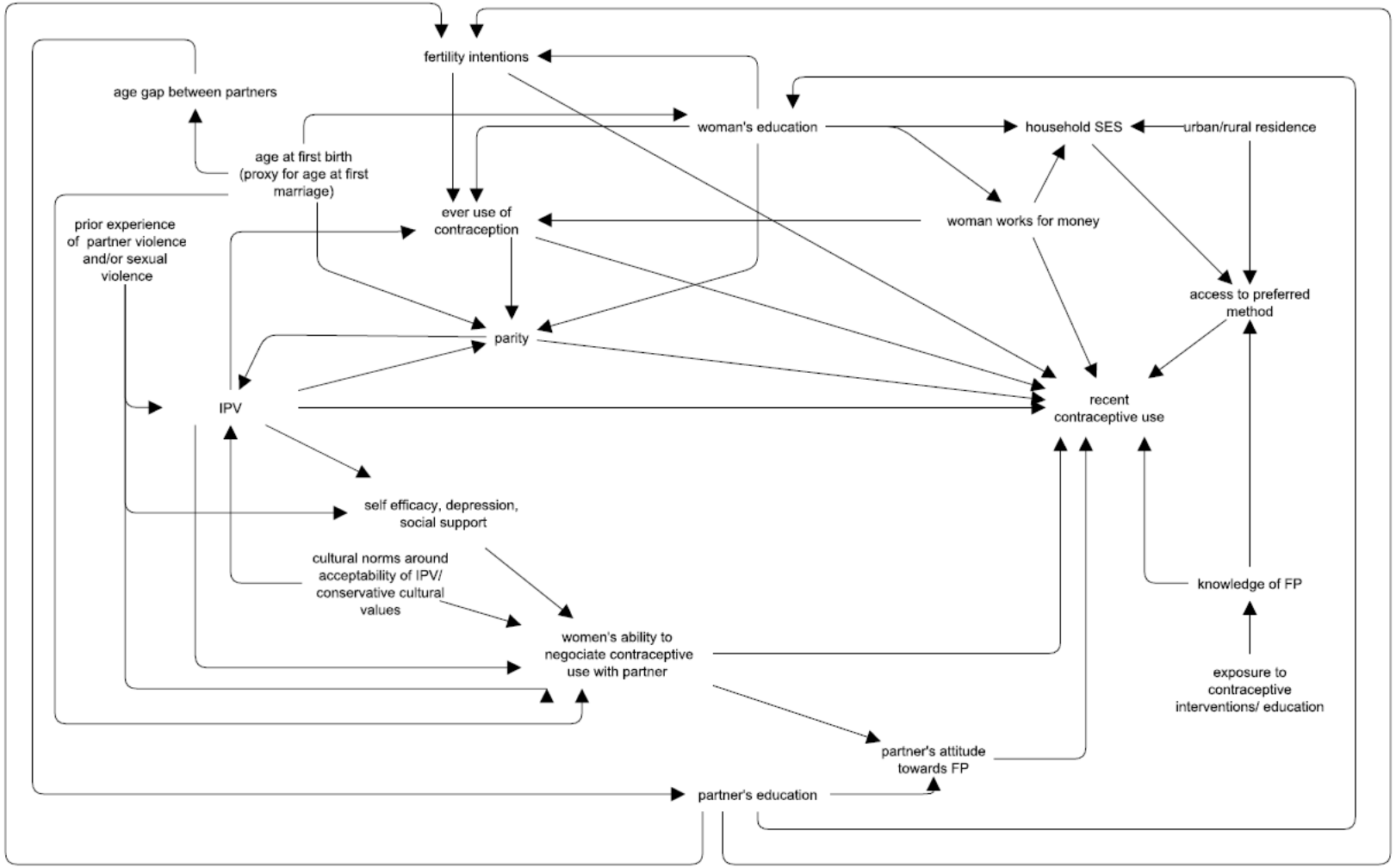
Possibly different on:

- Race, birth year, sex, Down's syndrome, chromosomal or other congenital anomaly, birth weight, gestational age, Apgar score, direct antiglobulin test positive, maternal age, delivery mode, multiple birth, jaundice and more. Full richness of the EHR at Kaiser available.
- What if we want to try and control for these differences using a statistical model (remove bias)?
- As well as two-way interactions of all variables (e.g., cross-combinations of categorical variables) and nonlinear components of numeric variables. When totaled up, over 2,000 predictors.

Outline

- Introduction: prediction versus causation
- Threats to causal conclusions
- **Causal methods in (very) brief**
- Interpretability of machine learning methods
- Some uses of big data and machine learning in causal inference
- Summary

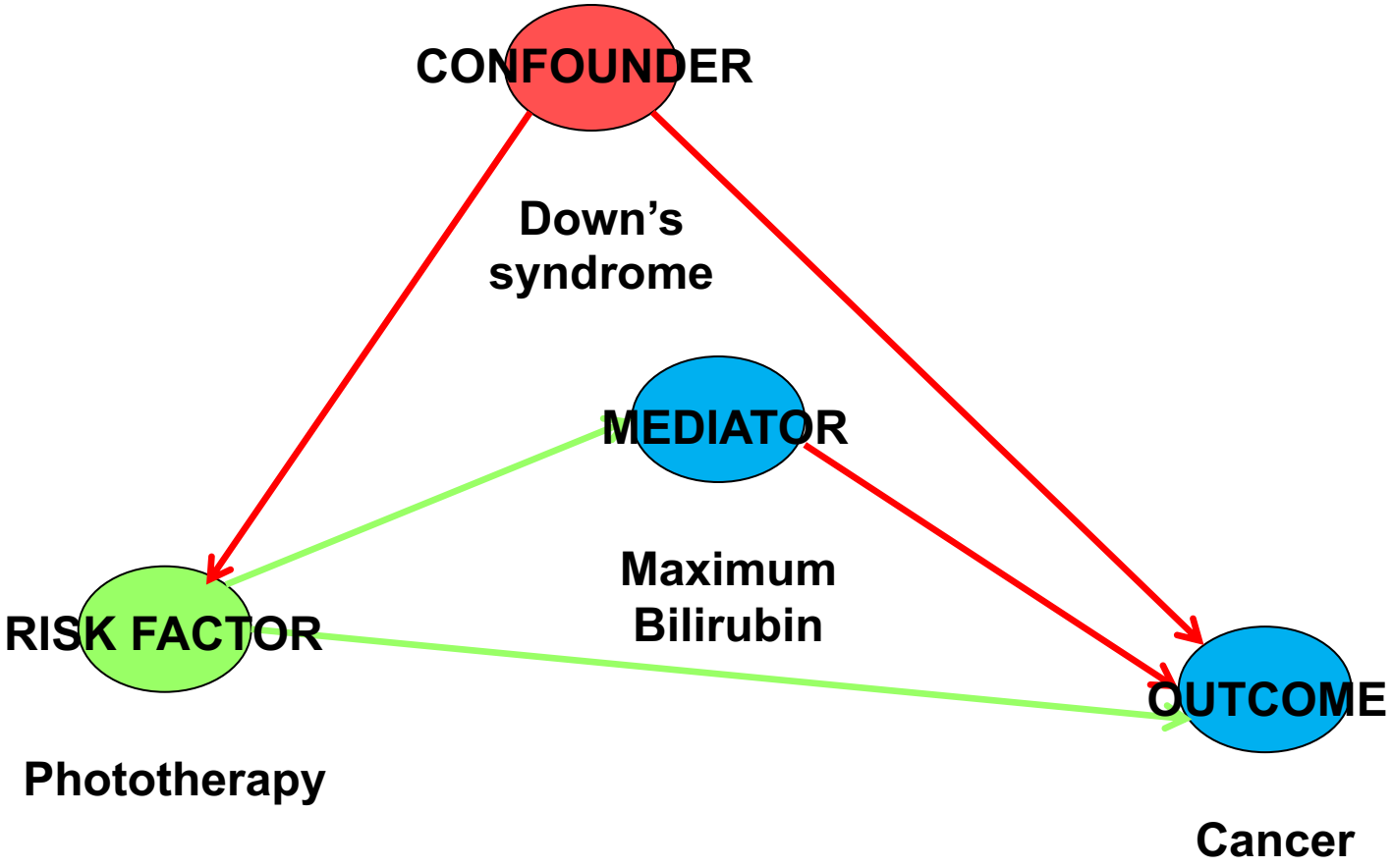
Causal inference



Directed acyclic graph of hypothesized causal relationship between IPV and women's contraceptive use

Maxwell, et al ,PLoS One. 2015; 10(2): e0118234.

Causal inference – the 5¢ tour



Adapted from Newman, et al Pediatrics, 2016

Biostat 215 covers these topics in detail

Epidemiologists have long studied the pitfalls of using observational studies to infer causation.

A fundamental idea in causal inference is to compare the same person with and without an intervention or condition. Can't usually measure the person under both conditions.

- For example, observe the same person with and without a history of knee injury to understand its impact on later osteoarthritis.
- Or recording the outcomes in a catchment area of a trauma center with and without its closure.

Potential outcomes

- Imagine a hypothetical experiment in which you get to observe each participant under both the treatment and control conditions holding all else the same, like a perfect cross-over experiment.

$$Y_{trt}, Y_{ctl}$$

- Often, we only get to observe one of Y_{trt} or Y_{ctl}, depending on whether the participant is in the treatment or control condition.
- Sometimes described as the **"fundamental problem in causal inference."**

Hypothetical example: treatment of depression

- Does addition of an internet based cognitive behavioral component aid in treatment of depression?
- Outcome = improvement in Beck Depression Inventory (BDI).
- Control group treatment is team care approach, which has proven especially effective in the elderly.
- Observational study based on clinical data (non-random).

Potential outcomes framework

		ΔBDI Outcome under		
Patient	Group	Ctl	Trt	Difference
1	Trt	4	8	4
2	Ctl	0	3	3
3	Trt	4	3	-1
4	Trt	3	4	1
…				
Ave in popl'n				1.02 = Average Causal Effect

Thought #5 (from lecture 1)

- If you are interested in causation you can't measure it all.

You can, at most, measure half of it all and that half is subject to selection bias. (Unless you have run a well-conducted randomized experiment).

Potential outcomes are sometimes called “counterfactuals”

Counter – factual

Against – the truth

= Lying

I like “Potential outcomes framework”

or “Hypothetical outcomes framework” better

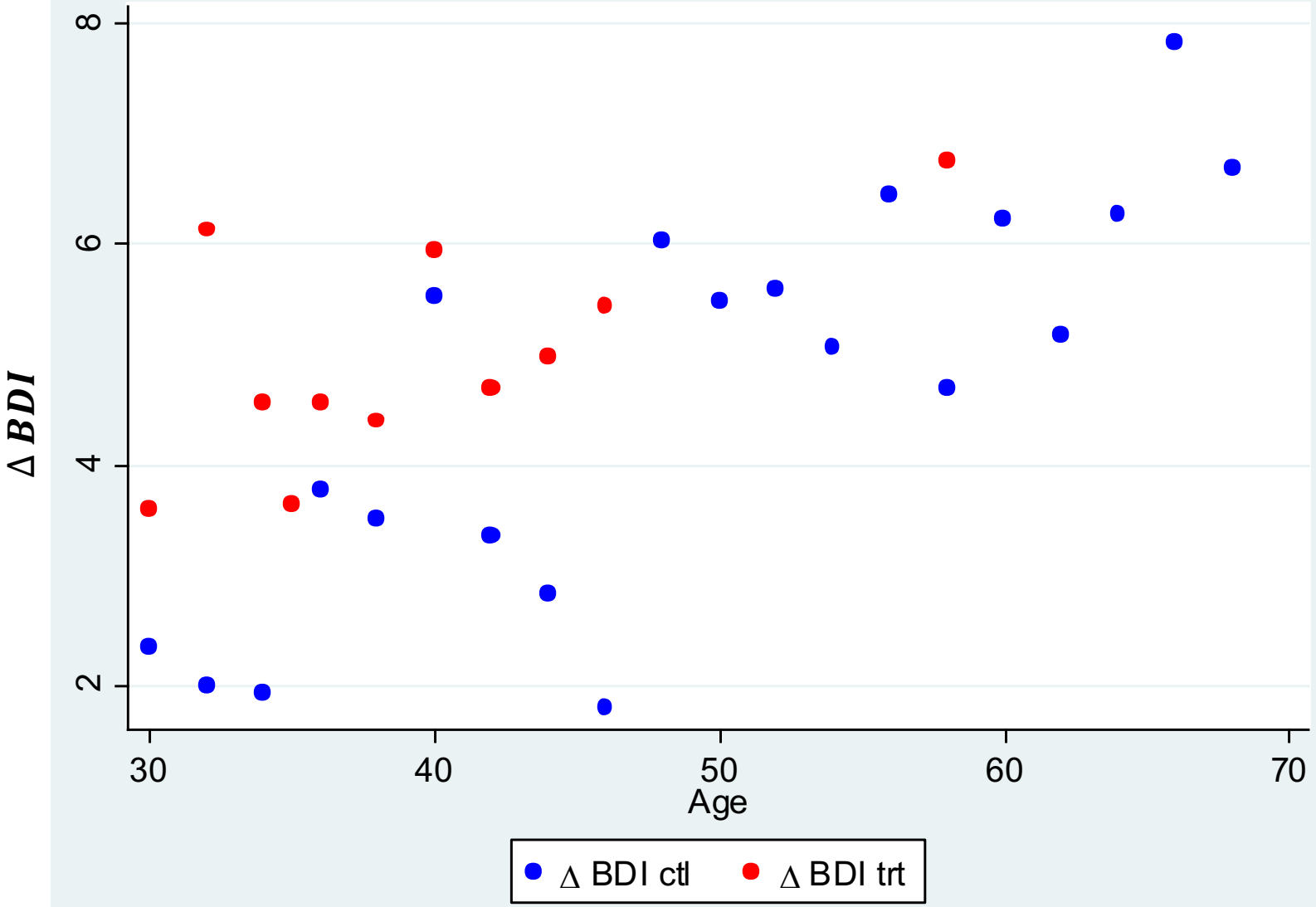
Potential Outcomes – Average Causal Effect

- Often, a reasonable quantity to estimate is the average causal effect (ACE): the average of the individual causal effects across the entire population.

$$\text{ACE} = \frac{1}{n}\sum_{i=1}^{n} Y_{trt\ i} \_ Y_{ctrl\ i}$$

- Nice in theory but you cannot directly observe the individual causal effects! E.g. you cannot observe someone smoke and not smoke over their lifetime.
- So to estimate ACE, we'll need to make some assumptions (more later)

Example: Depression Data



Simple analysis: t-test

Average change in Treatment group = 5.0,

Average change in Control group = 4.6

Estimated difference =0.4,

p=0.57, via t-test,

So not statistically significant.

Non-comparable groups

The issue:

Estimation of the ***causal*** effect of treatment (the predictor of interest) is ***confounded by*** age (another predictor) since

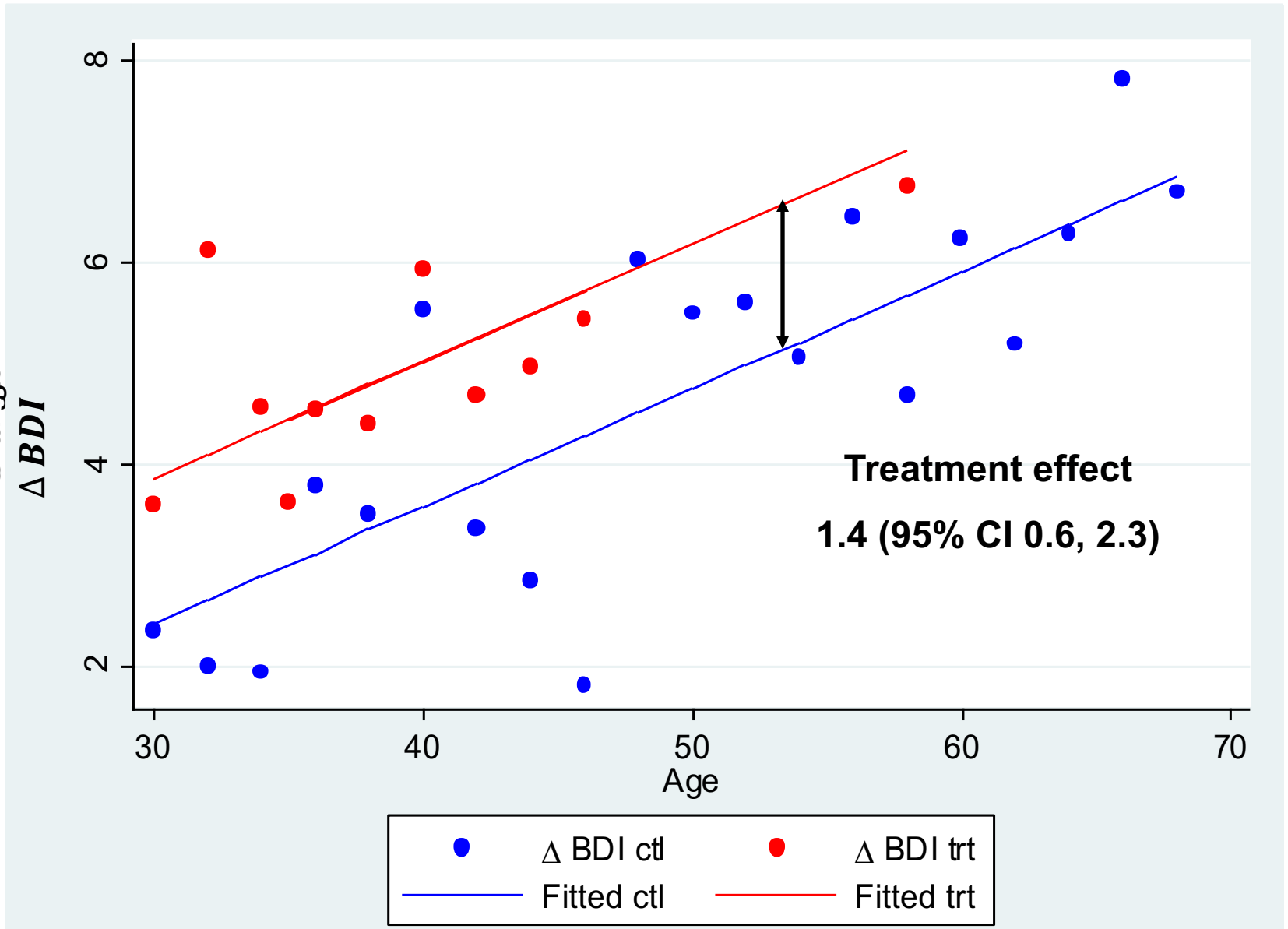
a) age is associated with the outcome (change in BDI) *and*

b) age is associated with treatment

A traditional solution:

Include age in a multipredictor linear regression model along with treatment.

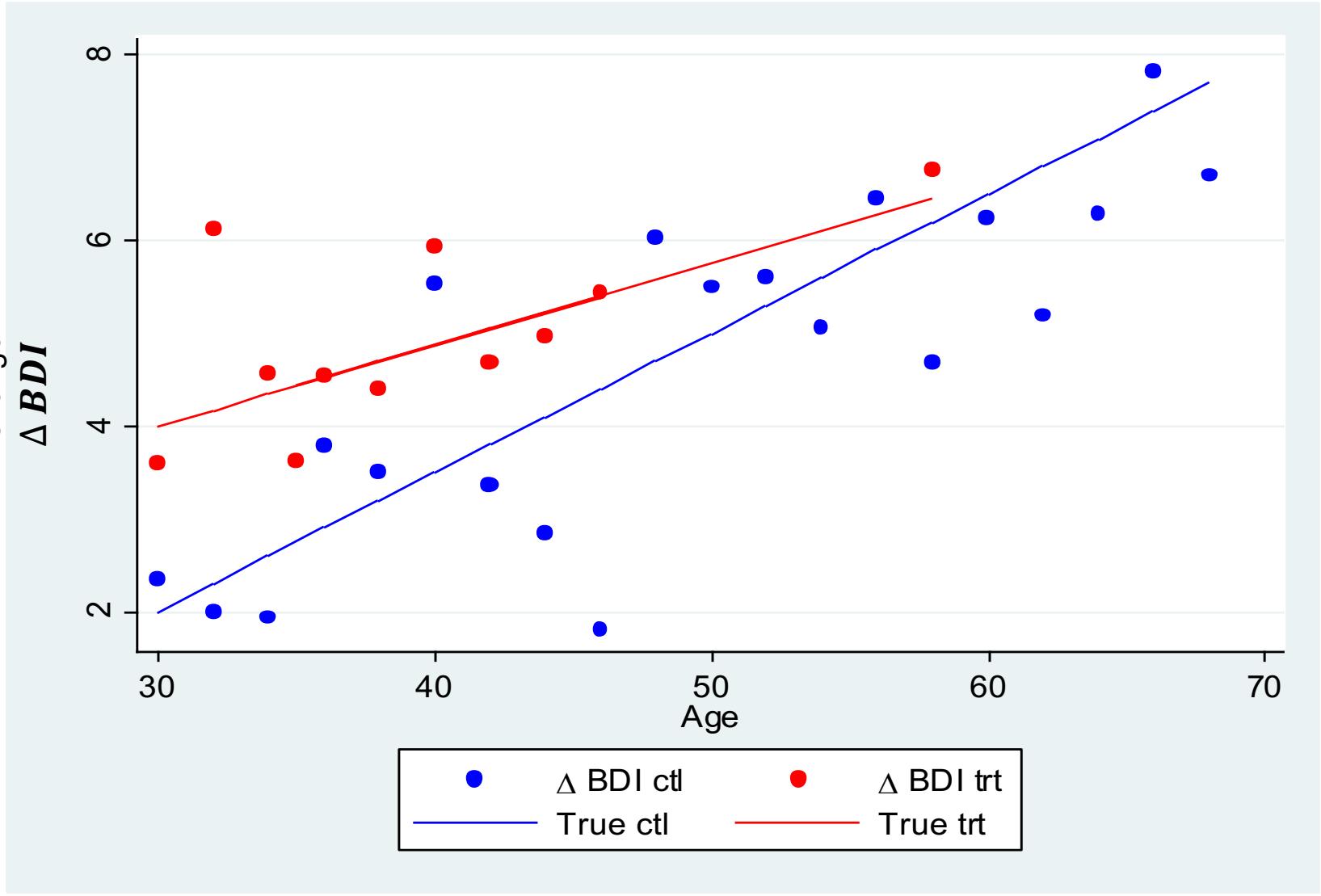
Example


Δ BDI
8
6
4
2
30
40
50
60
70
Age
Treatment effect
1.4 (95% CI 0.6, 2.3)
Δ BDI ctl
Δ BDI trt
Fitted ctl
Fitted trt

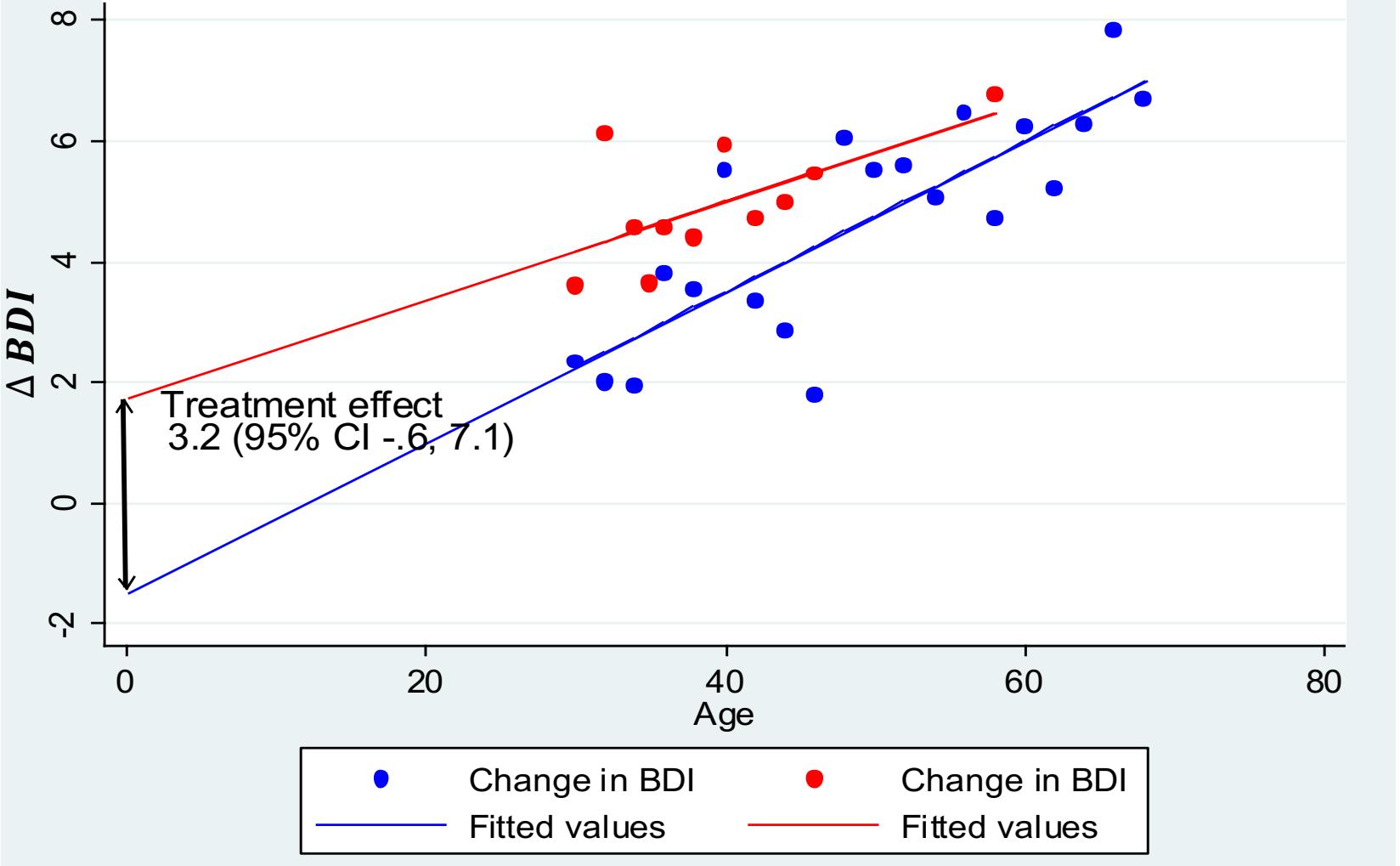
Issues with regression adjustment

- Causal estimate is *defined by* using regression model (contrast the ACE).
- What if model is wrong? As examples: assumption of a linear relationship when it is not. Left out variables? Interactions between variables?
- How will we know? (Lack of overlap/extrapolation.)
- Lack of comparison group for older ages (plenty of controls, not many treated).

Issues with regression adjustment (true model)



Issues with regression adjustment (fit interaction)



Outline

- Introduction: prediction versus causation
- Threats to causal conclusions
- Causal methods in (very) brief
- **Interpretability of machine learning methods**
- Some uses of big data and machine learning in causal inference
- Summary

Interpretability

Example: Prediction of healthy aging. Data source: Health ABC – a long running cohort study that enrolled about 3,000 healthy individuals in their 70s.

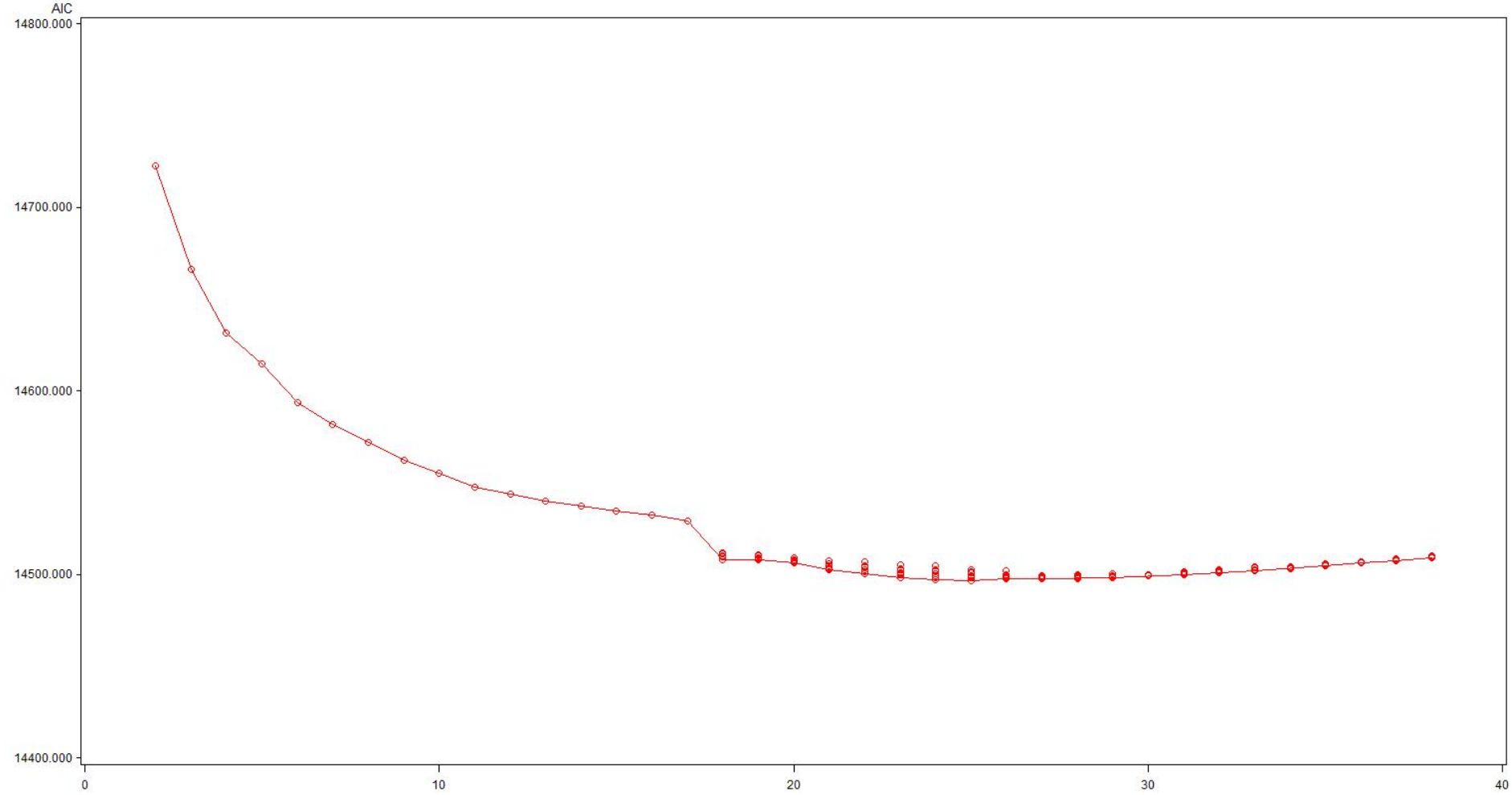
Outcome: Survival free of major diseases and disability.

Predictors: About 100 predictors including demographics, BP, medical history, physical performance (e.g., gait speed), and serum markers (e.g., IL-6).

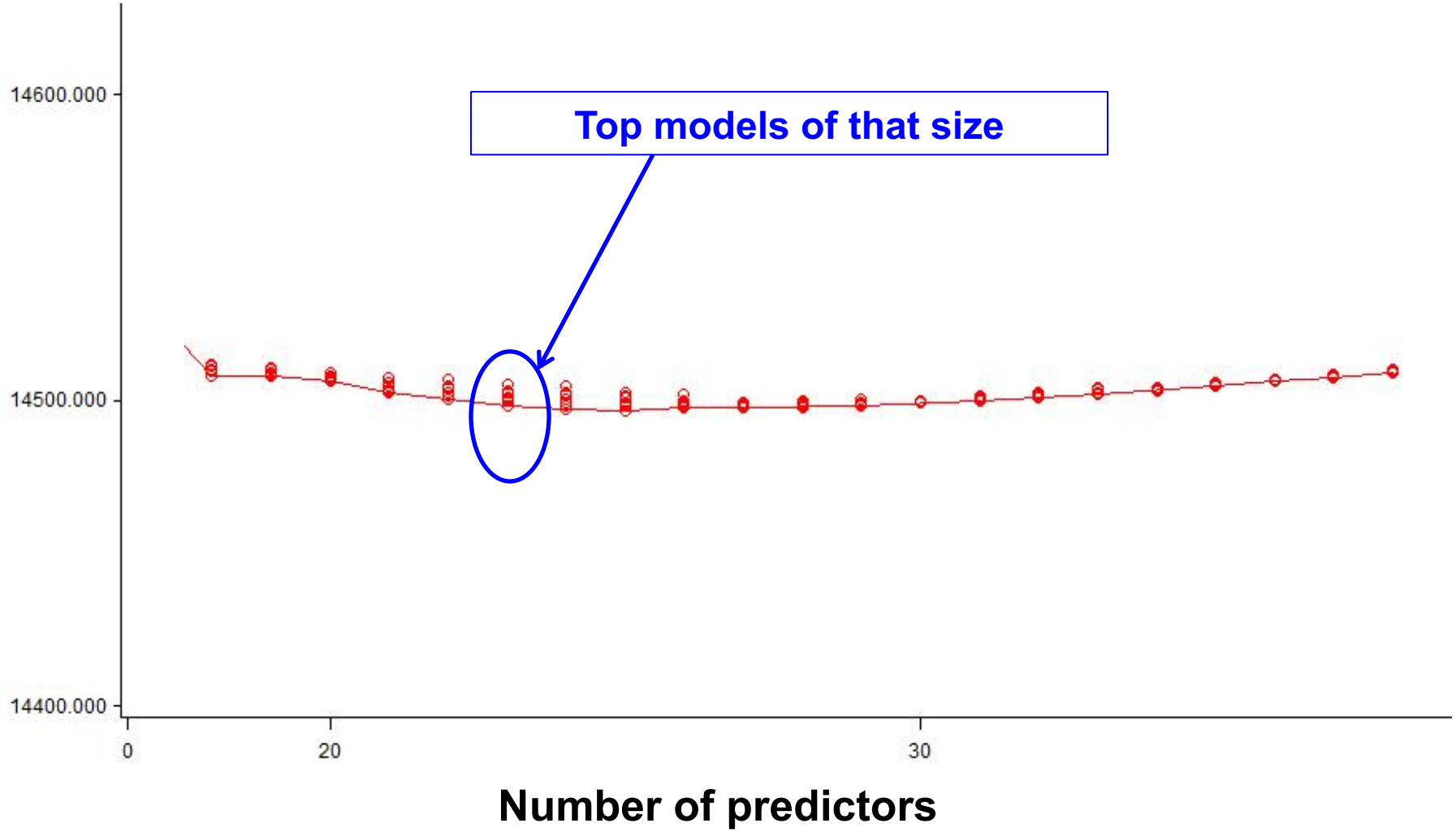
Interpretability

Model selection strategy: Calculated Akaike Information Criterion (AIC) values, which are similar to cross-validated errors, for a large number of Cox survival analysis models. Exhaustively searched for models with about the optimal number of predictors.

Example: Disease free-survival
AIC plot



Example: Disease free-survival AIC plot – zoom in near optimal



Machine learning/causal inference

Results: About 120 models were within 10 of the model with the lowest (best) AIC. (Rule of thumb: statistically speaking AIC larger by 10 is a worse fitting model).

Observation: Possible to get nearly the same quality of prediction using very different models.

Predictor selection: Prediction algorithms cannot be depended upon to identify the "best" predictors. Which predictors get included or excluded from the model will vary with small perturbations in the dataset or analysis method.

Example: Two models

For example, these two models were virtually identical in performance:

19 variable model: BLAGE WTK HGT WGTPDB HGTDB PULSE BESFEV BPAAIT CESD PACE400 PACE20 GRIP SEMITND TANDSTD COMP FALL RDW ALBUMIN STPCK

29 variable model: BLAGE WHITE WTK HGT WGTPDB HGTDB PULSE FAT BESFVC FEV1R BPAAIT TENG DSS CESD PACE400 PACE6 PACE20 GRIP SEMITND TANDSTD STANDY FFHIP COMP FALL CYSTC LDL RDW ALBUMIN STPCK

Cognition variables

Red indicates not in the other model

Machine learning/causal inference

And this is with a standard statistical analysis approach (Cox regression). Many other machine learning methods are worse because they don't give interpretable coefficients.

Machine learning/causal inference

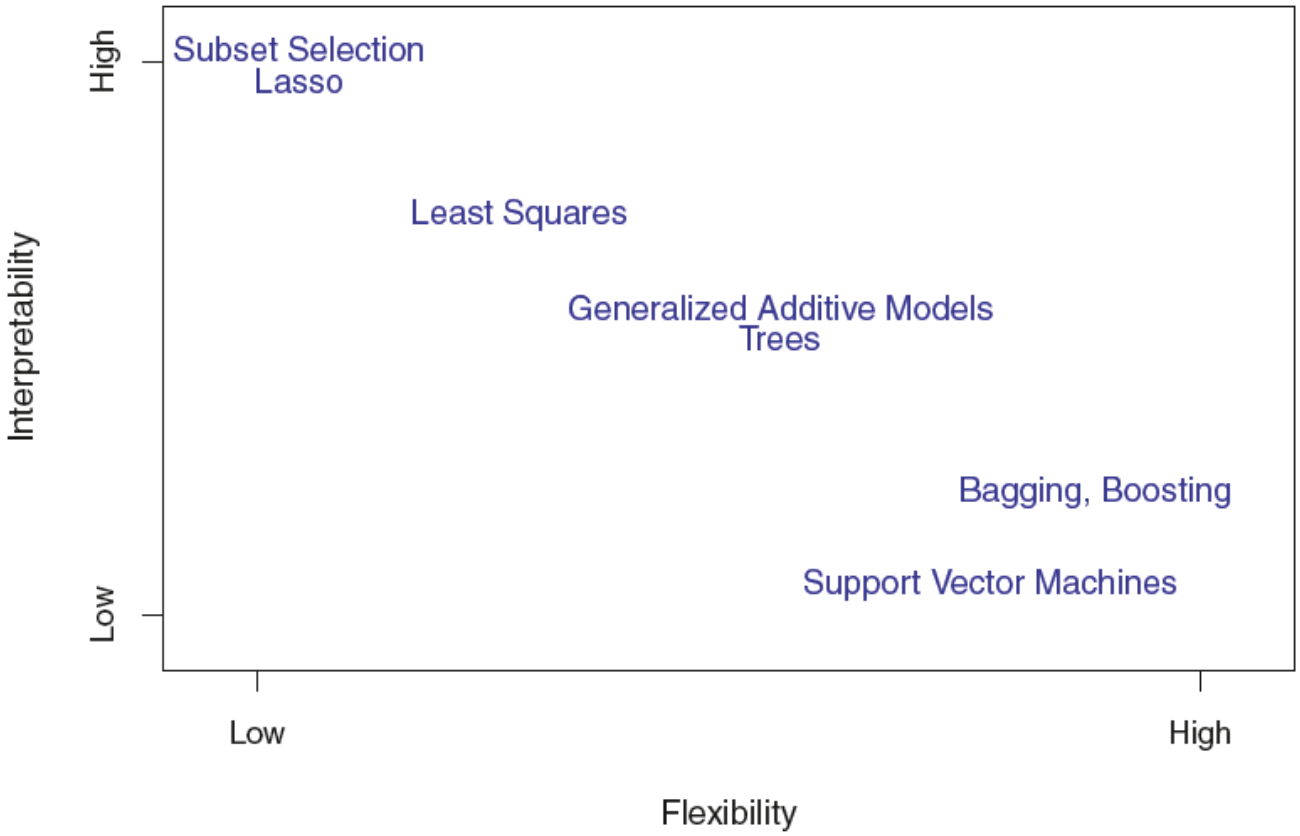


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

From Introduction to Statistical Learning, James, et al., 2013

Outline

- Introduction: prediction versus causation
- Threats to causal conclusions
- Causal methods in (very) brief
- Interpretability of machine learning methods
- Some uses of big data and machine learning in causal inference
- Summary

Outline

- Some uses of big data and machine learning in causal inference
  - **Conduct an RCT using large scale data**
  - Propensity scores
  - Potential outcomes estimation
  - Subgroup analyses

Clinical trials:

- Randomized controlled trials (RCT) often function in “laboratory” like conditions, i.e. artificial
- Pragmatic clinical trial:
  - Focus on measuring effect of an intervention in a more pragmatic setting
  - See lecture by Mark Pletcher on pragmatic trials (uploaded to CLE)
- Intervention applied in more realistic setting e.g. across many sites
- Predictors and outcomes measured through EHR records
- **Goal**: put clinical trials in more realistic settings, leverage EHR data, and save $$$

Pragmatic clinical trials:

TABLE 1 Attributes of Different Clinical Trial Designs

Clinical Study Design	Relative Cost	Design and Data Collection	Patient Population	Potential for Bias	Advantages and Disadvantages
Observational studies (including registry studies)	$	Can be retrospective or prospective in design; data quality is variable	Typically unselected population (e.g., Medicare)	Without randomization, comparative effectiveness studies cannot be performed	Large population; often many unmeasured variables or unexplained factors
Traditional RCTs	$$$$-$$$$$	Prospective design; data collection occurs at specialized study centers	Highly-selected patient population at study centers; may lead to results that are not generalizable	Randomization eliminates confounding bias	Current gold standard for comparative-effectiveness studies
Registry-based RCTs	$$-$$$	Prospective design; data collection often occurs at diverse clinical sites	Typically designed to study a specific patient population (e.g., those undergoing PCI)	Randomization eliminates confounding bias	Large number of outcomes; harnesses power of already-established clinical registry
Large, pragmatic clinical trials	$$-$$$$	Prospective design; data is collected ubiquitously as part of clinical care	Depending on electronic infrastructure, can be broad or selective; can incorporate enrichment criteria	Randomization eliminates confounding bias	Simple design; large number of outcomes; requires infrastructure that can facilitate easy and quick enrollment

Jones, et al (2016). The Changing Landscape of Randomized Clinical Trials in Cardiovascular Disease.

Outline

- Some uses of big data and machine learning in causal inference
  - Conduct an RCT using large scale data collection systems (Mark Pletcher's lecture)
  - **Propensity scores**
  - Potential outcomes estimation
  - Subgroup analyses

Propensity scores:

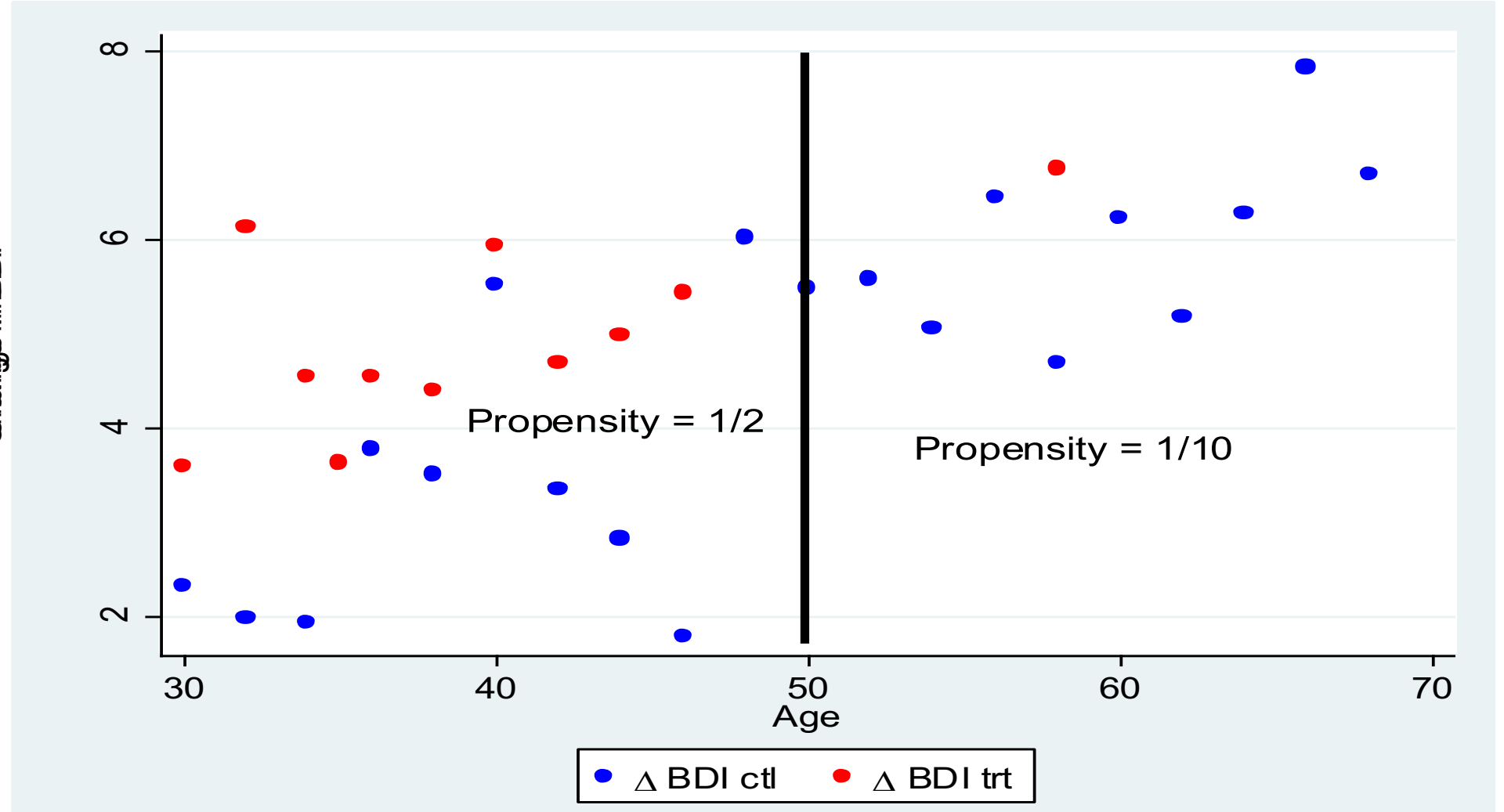
"[Propensity score] matching is the observational study analog of randomization in ideal experiments…– Rubin and Thomas (2000), *Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates*

Propensity scores:

- Define prop(x) be the probability of being on treatment as a function of x, all the variables that determine treatment. For each subject this is called a **propensity score**.

- In our depression example, suppose temporarily that the probability of selecting treatment only depends on age.

Propensity scores: example



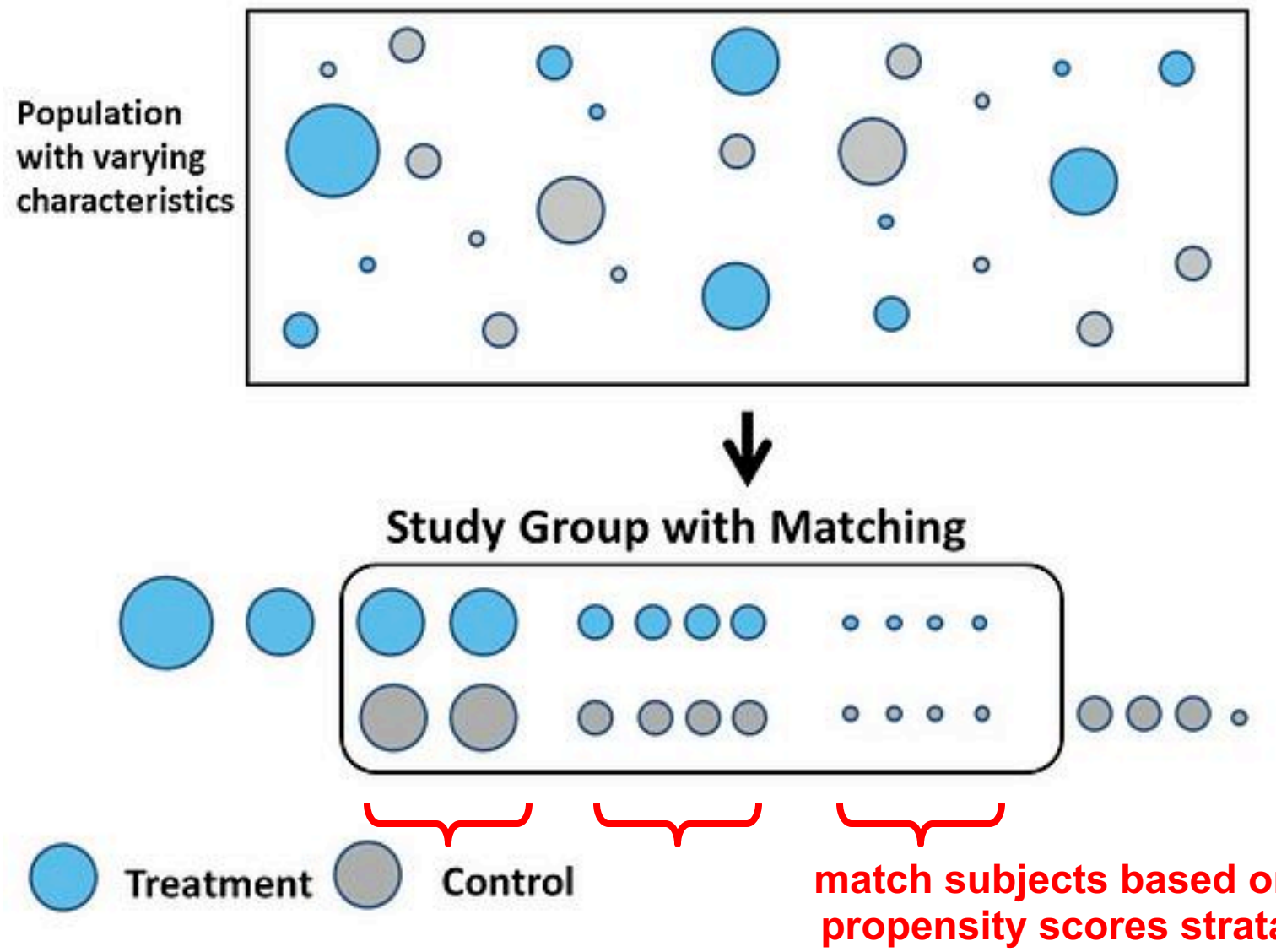
Propensity scores: theory

1. Very important theoretical property: Consider individuals with the same value of prop(x). The ones receiving treatment have the same distribution of x as do those who do not. *So values of the predictors are more likely to be balanced between the two groups which share a propensity for treatment assignment.*
2. Therefore you only need to adjust for prop(x) to balance the groups.

Propensity score stratification

1. Select covariates x that may influence treatment assignment (e.g. age)
2. Estimate propensity of treatment for each subject given x (using supervised ML algorithm)
3. Match individuals in the treatment group with individuals in the control group based on their propensity scores (directly or via stratification)
4. Perform analysis on matched sample

E.g. Propensity score matching



https://www.summitllc.us/propensity-score-matching

Propensity scores: depression example

Adjusting for propensity score categories and using a linear regression model:

Estimated difference =1.4, CI [0.4, 2.3],

Contrast previous result:

No adjustment for propensity score categories:
Estimated difference = 0.4, (p=0.57, via t-test)

Propensity scores: practical issues

- What if not all the variables that determine treatment are measured? Or included correctly in the model?
- Suggests being more inclusive with both predictors and interactions.
- So something that *is* easier with large databases and machine learning.

Propensity score estimation using Orange

- Getting estimates of the propensity score is a prediction problem.
- So use your favorite machine learning algorithm to predict the probability of treatment (use a “Edit Domain” widget to set Treatment as the Target variable temporarily).
- The propensity scores are these predicted probabilities.

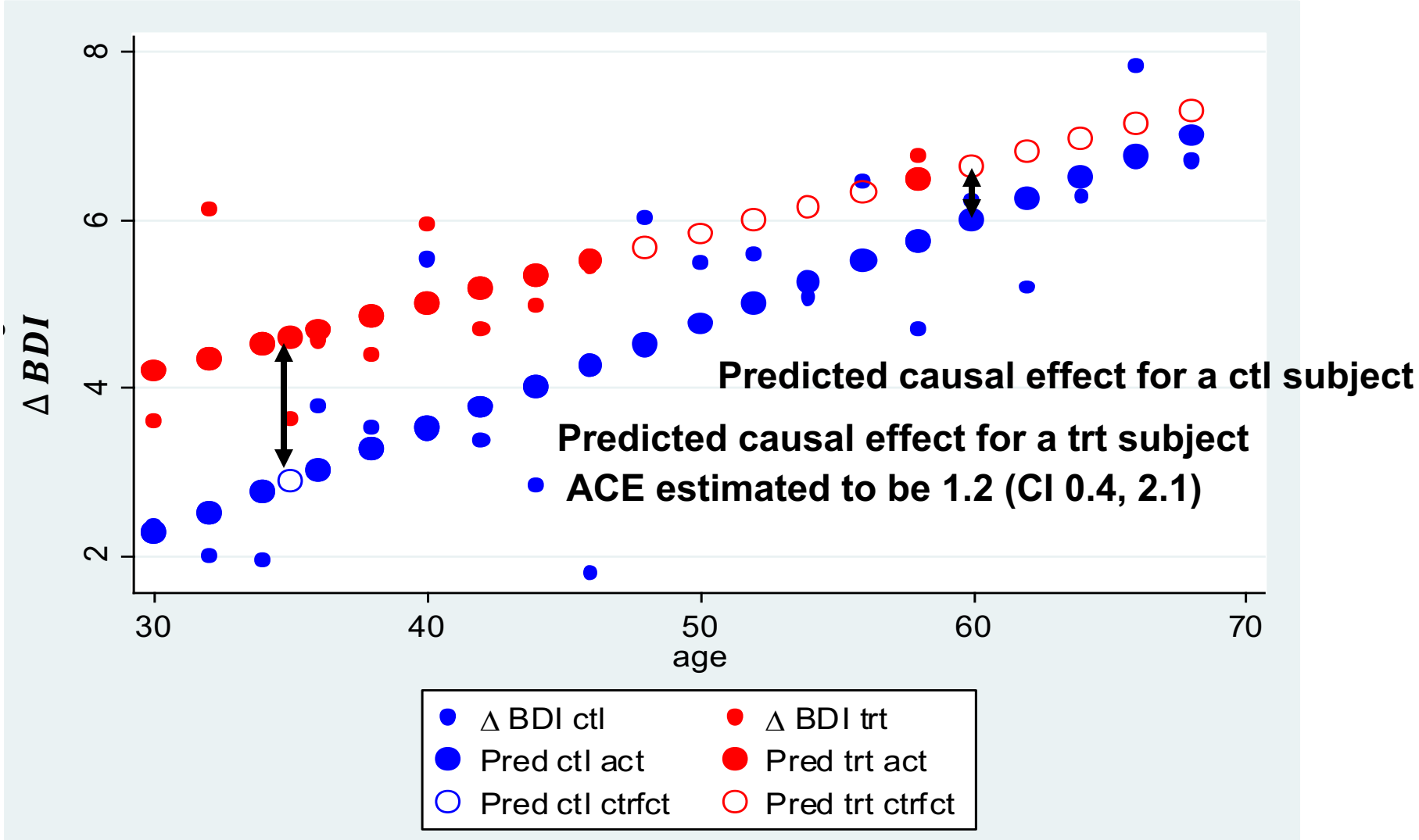
Outline

- Some uses of big data and machine learning in causal inference
  - Conduct an RCT using large scale data collection systems (Mark Pletcher's lecture)
  - Propensity scores
  - **Potential outcomes estimation**
  - Subgroup analyses

Regression estimation

- In the depression example, we used a linear regression model. We can use that to get an estimate of the causal effect for each person. Then we could calculate the average causal effect.
- Also called *G-computation* in the causal literature and *marginal estimates* in economic statistics.
- Can get marginal estimates for subpopulations, e.g., causal effect in younger participants.

Potential outcomes estimation



Potential outcomes estimation using Orange

- Create duplicate rows of data with the treatment assignment switched and missing outcome data:

Patient_ID	Sex	Treatment	Systolic_BP	…other predictors
1	F	1	120	
2	M	1	135	
3	F	0	142	
…				
1	F	0	missing	
2	M	0	missing	
3	F	1	missing	

Potential outcomes vs propensity score adjustment

- Potential outcomes: try to **estimate** the potential outcome using supervised learning
- Propensity score adjustment: try to construct **balanced** groups
- Both rely on models
- Can be used alone or in tandem (“doubly robust”)
- Neither helps us extrapolate to samples that do not represent the training data

Outline

- Some uses of big data and machine learning in causal inference
  - Conduct an RCT using large scale data collection systems (Mark Pletcher's lecture)
  - Propensity scores
  - Potential outcomes estimation
  - **Subgroup analyses**

Subgroup effects and precision medicine

- Since the causal effect is not the same for everyone, are there subgroups for which the treatment is more or less efficacious (or even harmful)?
- If we can identify the subgroups we can tailor treatment (precision medicine).
- How many subgroups might there be?
- Let's consider the phototherapy/cancer example.

Subgroup problem

- Race (5 categories), birth year (say 3), sex (2), Down's syndrome (2), chromosomal (2) or other congenital anomaly (2), birth weight (6), gestational age (say 3), Apgar score (2), direct antiglobulin test positive (2), maternal age (2), delivery mode (2), multiple birth (2), jaundice (2).
- So $5\times3\times2\times2\times2\times2\times6\times3\times2\times2\times2\times2\times2\times2$

  =276,480 subgroups

  (for 500,000 babies)

Subgroup analysis issues

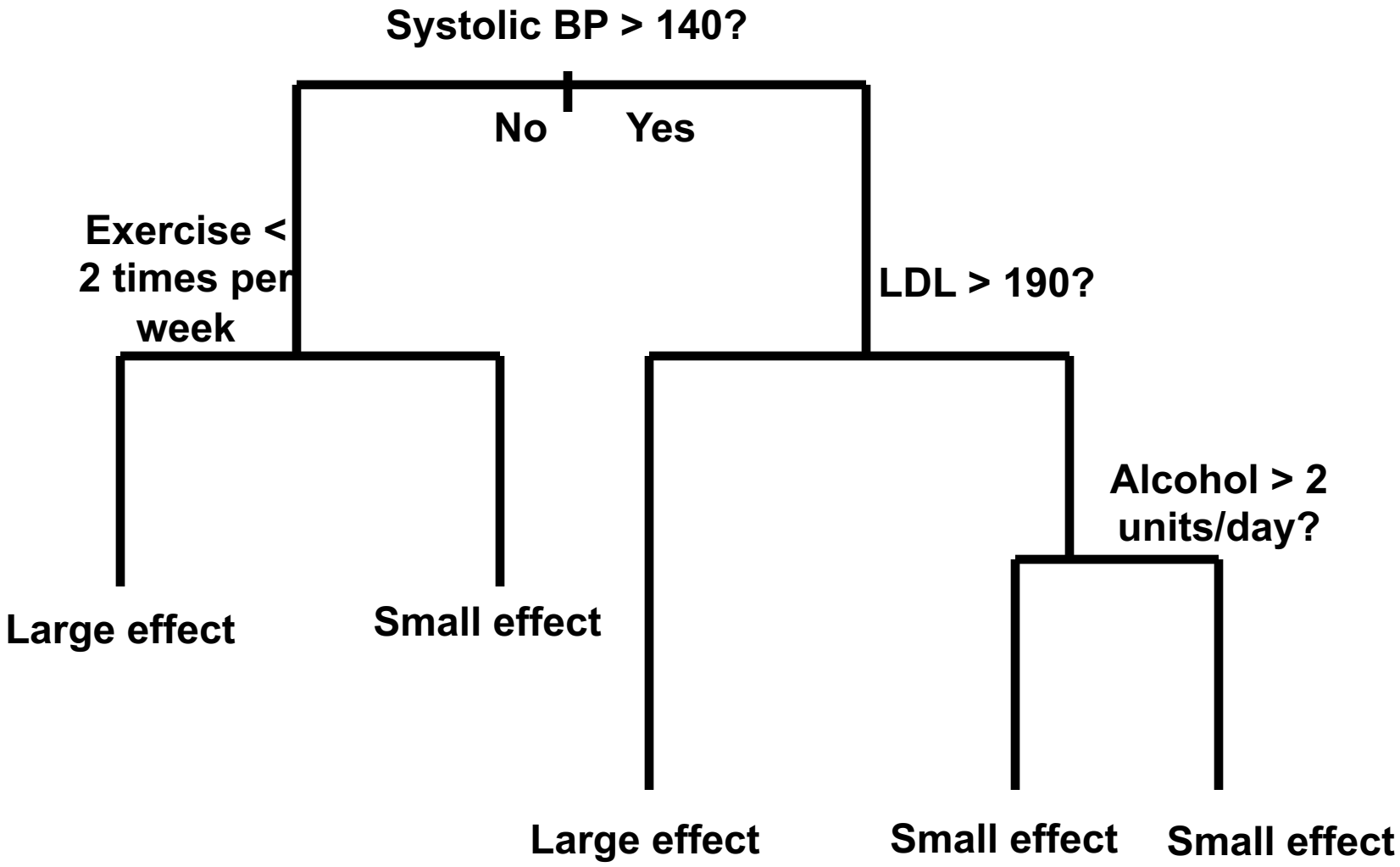
- Insufficient data in each subgroup.
- Huge multiple testing problem (if we test each one at α = 0.05 and there are no effects we expect $0.05 \times 276{,}480 = 13{,}824$ false positives

Subgroup problem – proposed solution

- Use the potential outcomes estimation approach in the previous section (and your favorite machine learning algorithm) to get an estimated treatment effect for each person using all the possible subgroup variables as predictors.
- Code each subject as having large/small treatment effect
- Use a recursive partitioning algorithm to find subgroups with different treatment effects and try to predict people who fall into the large/small groups

Reference for using this idea in RCTs: Foster, Taylor and Ruberg, Statistics in Medicine, 2011.

Subgroup analysis: decision tree



Summary

- Distinguish questions by whether they can be answered by prediction only or require more detailed analysis.
  - effect, intervention = causal,
  - estimate, association, correlation = prediction
- Machine learning methods are not designed for drawing causal conclusions.
- However, *aspects of* causal analyses may be prediction problems for which we can use the methods in the class
  - Estimating propensity scores and predicting potential outcomes (e.g. regression estimation)
  - Identifying subgroups on the basis of different estimated treatment effects