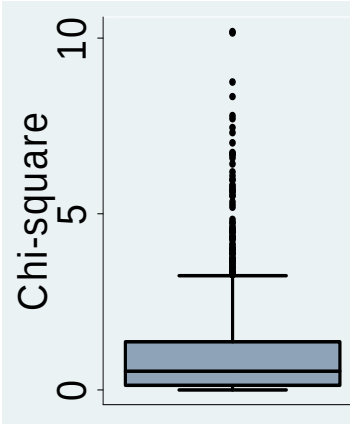
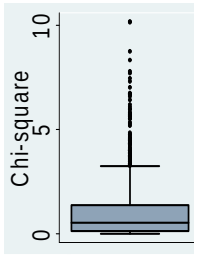
Biostat 202: Opportunities and Challenges of "Big Data" - Causal Inference

Charles E. McCulloch,
Division of Biostatistics,
Department of Epidemiology and Biostatistics



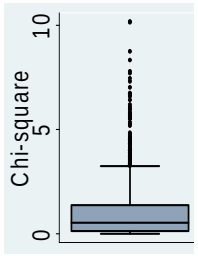
Biostat 202 - 2016

Outline



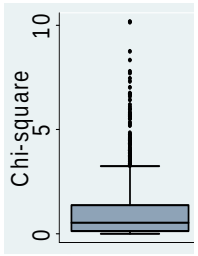
- Introduction: prediction versus causation
- Threats to causal conclusions
- Causal methods in (very) brief
- Interpretability of machine learning methods
- Some uses of big data and machine learning in causal inference
- Summary

Outline



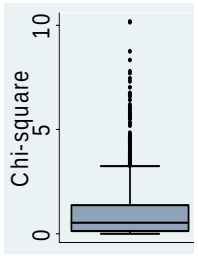
- **Introduction: prediction versus causation**
- Threats to causal conclusions
- Causal methods in (very) brief
- Interpretability of machine learning methods
- Some uses of big data and machine learning in causal inference
- Summary

Prediction versus causality



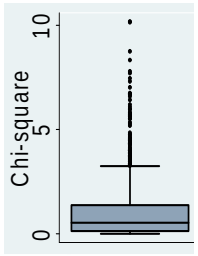
- Everyone in the science community understands the difference between causation and correlation.
- Data-mining and machine learning algorithms are only based on associations.
- Often interested in causation, i.e., understand the cause of illness or develop interventions that cause improvement.
- And I have often seen even senior researchers have lapses and over-interpret correlations causally.

Which of these are prediction versus causation questions?



- Effect of Abaloparatide on fractures in postmenopausal women.
- Estimating fetal risk of screening or not screening for genetic disease.
- Effect of cerebral protection device on brain lesions following aortic valve implantation.
- Estimation of temporal trends in preterm birth rates and their association with obstetric interventions.
- Survival for children with trisomy 13 and 18.

Which of these are prediction versus causation questions?



- Effect of Abaloparatide on fractures in postmenopausal women.
- Estimating fetal risk of screening or not screening for genetic disease.
- Effect of cerebral protection device on brain lesions following aortic valve implantation.
- Estimation of temporal trends in preterm birth rates and their association with obstetric interventions.
- Estimation of survival for children with trisomy 13 and 18.

Original Investigation

Temporal Trends in Late Preterm and Early Term Birth Rates in 6 High-Income Countries in North America and Europe and Association With Clinician-Initiated Obstetric Interventions

Jennifer L. Richards, MPH; Michael S. Kramer, MD; Paromita Deb-Rinker, PhD; Jocelyn Rouleau; Laust Mortensen, PhD; Mika Gissler, DPhil; Nils-Halvdan Morken, MD, PhD; Rolv Skjærven, PhD; Sven Cnattingius, MD, PhD; Stefan Johansson, MD, PhD; Marie Delnord, MSc, MA; Siobhan M. Dolan, MD, MPH; Naho Morisaki, MD, MPH, PhD; Suzanne Tough, PhD; Jennifer Zeitlin, DSc; Michael R. Kramer, PhD

Late preterm and early term births are of emerging clinical and public health importance and concern due to the associated risks of adverse neonatal and childhood outcomes.[1,2] Preterm and early term births may occur spontaneously or be initiated by clinicians through the use of obstetric interventions such as labor induction or cesarean delivery. In the case of maternal or fetal clinical indications (eg, preeclampsia or fetal growth restriction), obstetric interventions may be medically indicated or nonelective.[2] In the United States, clinicians have been urged to reduce nonmedically indicated or elective deliveries before 39 weeks.[3,4]

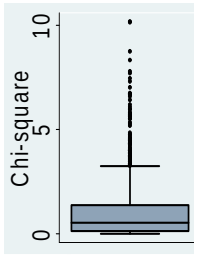
Increasing US preterm birth rates from the 1990s to the

Key Points

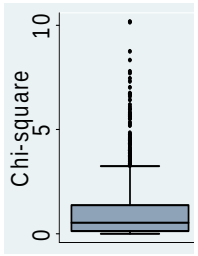
Question Are temporal trends in late preterm and early term birth rates in 6 high-income countries associated with use of clinician-initiated obstetric interventions?

Findings These population-based retrospective analyses of singleton live births found declining late preterm birth rates in Norway and the United States and declining early term birth rates in Norway, Sweden, and the United States. Obstetric interventions increased in some countries but decreased in the United States, where an association was observed with decreasing early term birth rates.

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

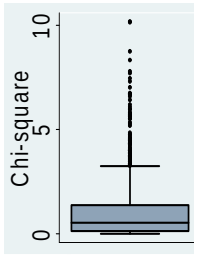
For the most up-to-date levels of evidence, see www.cebm.net/?o=1025

Therapy/Prevention/Etiology/Harm:

1a:	Systematic reviews (with homogeneity) of randomized controlled trials
1b:	Individual randomized controlled trials (with narrow confidence interval)
1c:	All or none randomized controlled trials
2a:	Systematic reviews (with homogeneity) of cohort studies
2b:	Individual cohort study or low quality randomized controlled trials (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

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.

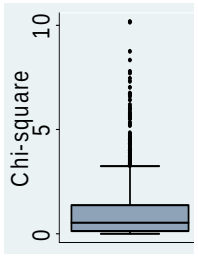
said no.

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



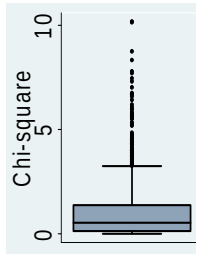
Suppose you are going to conduct a research study using electronic health records and want the results to generalize to all similar patients in the Bay Area?

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

UCSF ~Ann Landers

Kaiser ~Random telephone survey

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?

We studied about 500,000 infants born in Kaiser Northern CA between 1995 and 2011.

Newman, et al Pediatrics, 2016

Example: Excerpt from Table 1

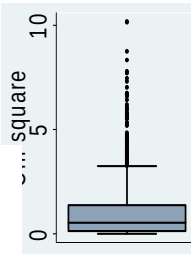
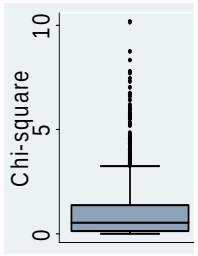


TABLE 1 Demographic and Clinical Characteristics of Infants Exposed to Phototherapy and 2 Groups of Unexposed Infants

	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

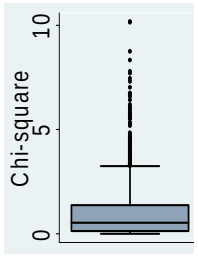


Possibly different on:

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 electronic health record system at Kaiser available.

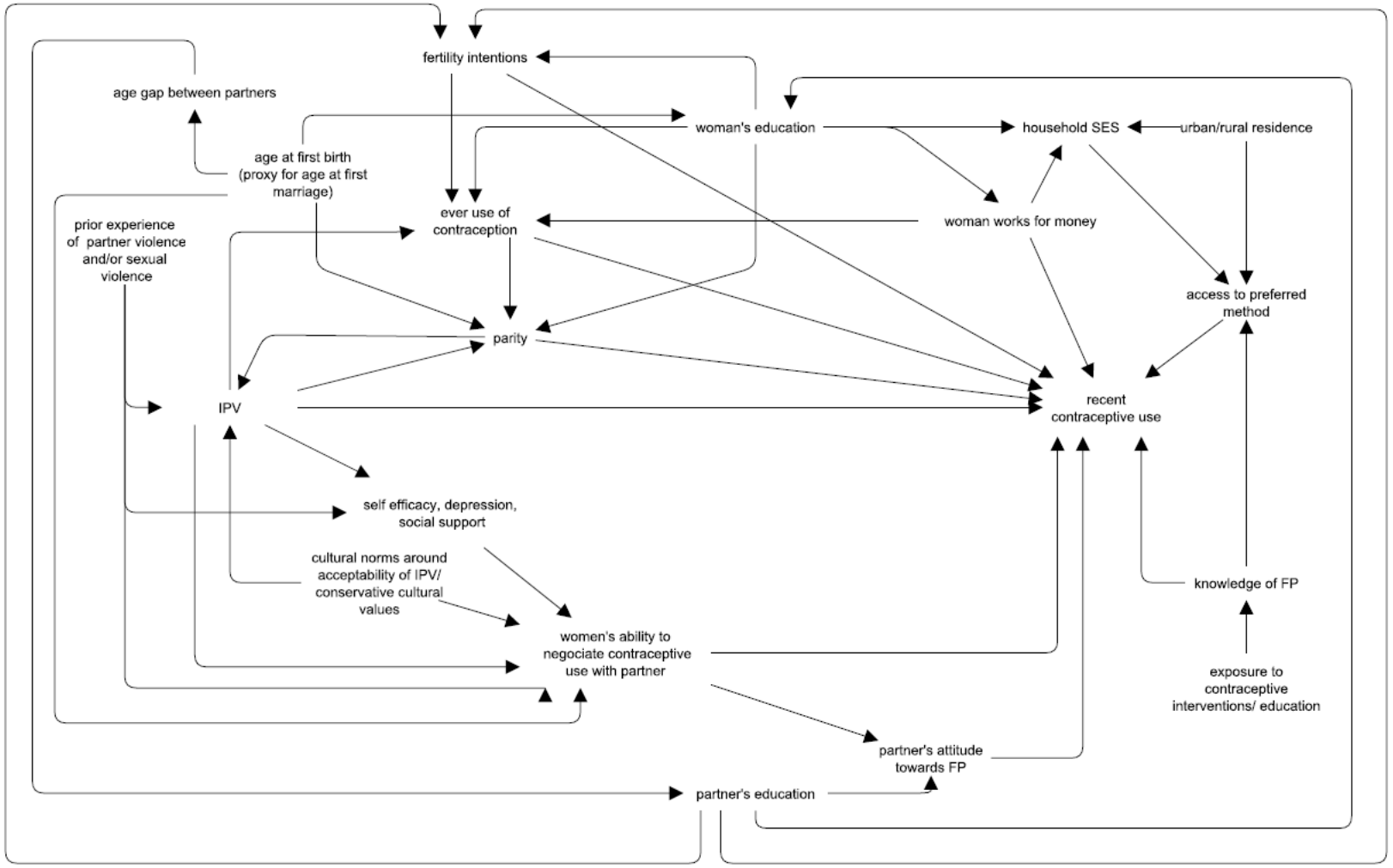
As well as two-way interactions of all variables (e.g., cross-combinations of categorical variables) and nonlinear components of numeric variables. When totaled up, over 2,000 predictors.

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

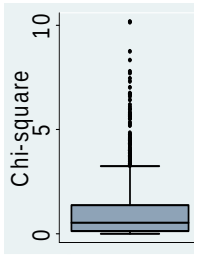
Causal inference

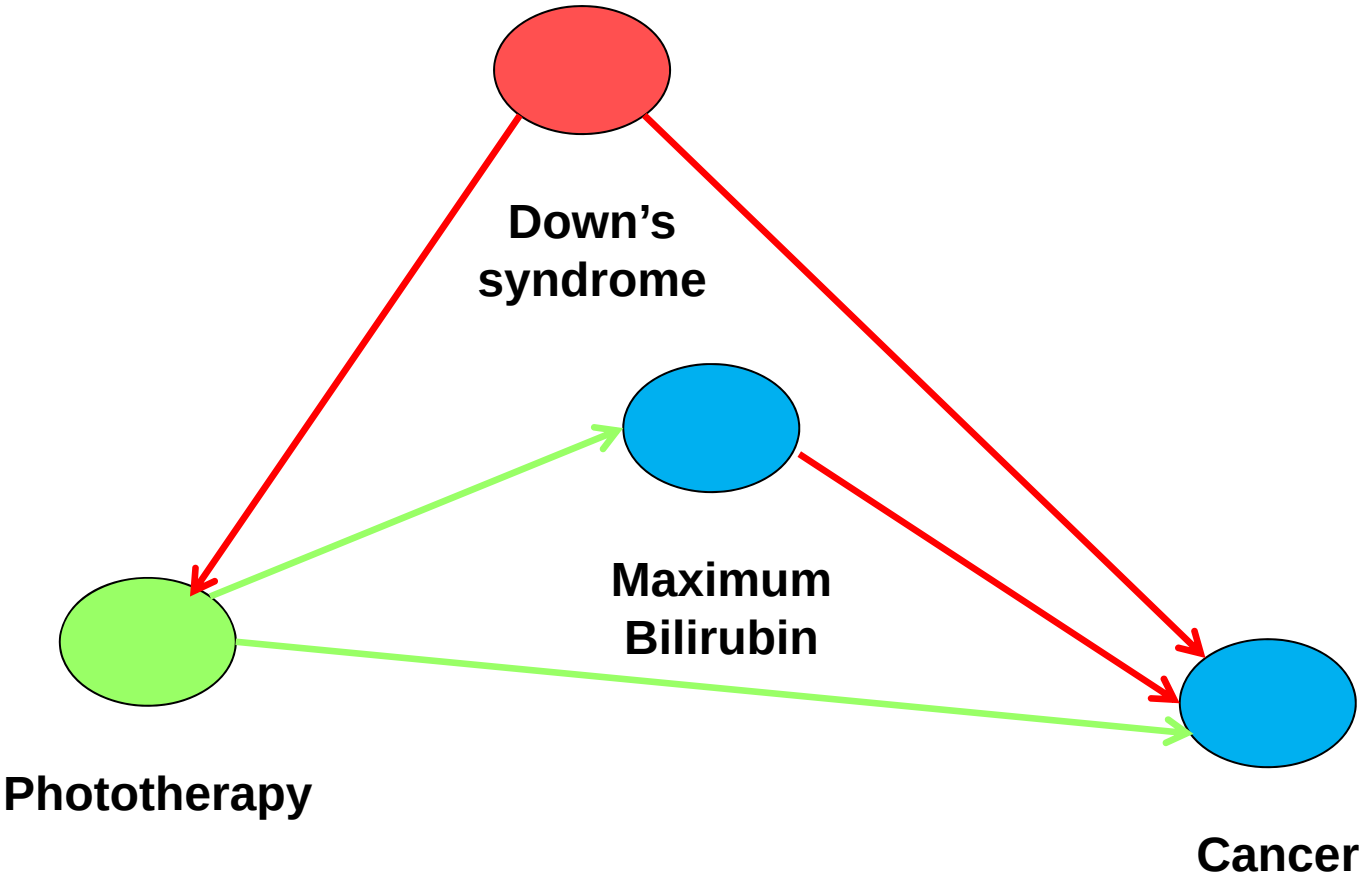


Directed acyclic graph of hypothesized causal relationship between IPV and women's contraceptive use

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

Causal inference – the 5¢ tour

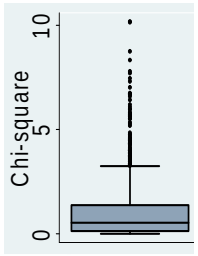




Adapted from Newman, et al Pediatrics, 2016

Biostat 215 covers these topics in detail

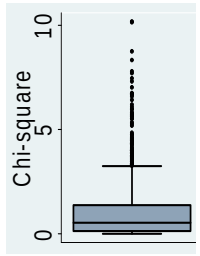
Epidemiologists have long studied the pitfalls of using observational studies to infer causation.



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

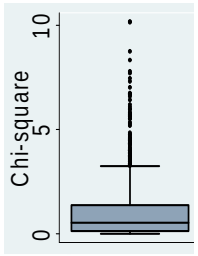


Imagine a hypothetical experiment in which you get to observe each participant under both the treatment and control conditions holding all else the same: Y_{trt}, Y_{ctl}. Like a perfect cross-over experiment.

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.

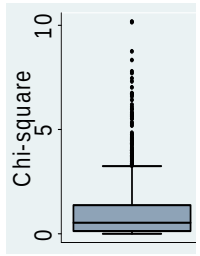
Sometimes described as the “fundamental problem in causal inference.”

Hypothetical example: treatment of depression



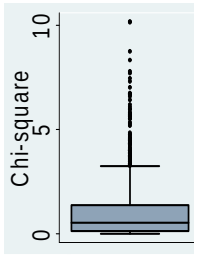
- Does addition of an internet based cognitive behavioral component aid in treatment of depression?
- Outcome = improvement in Beck Depression Inventory.
- Control group treatment is team care approach, which has proven especially effective in the elderly.
- Observational study based on clinical data.

Potential outcomes



		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

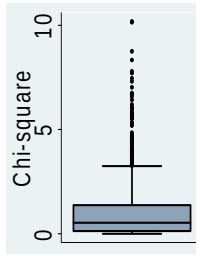
Thought #5 (from lecture 1)



- 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

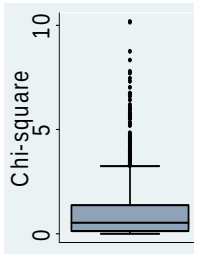
Against – the truth

=Lying

I like “Potential outcomes framework”

or “Hypothetical outcomes framework” better

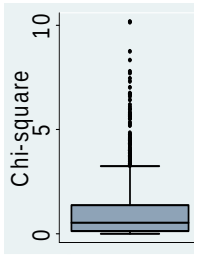
Potential Outcomes – Average Causal Effect

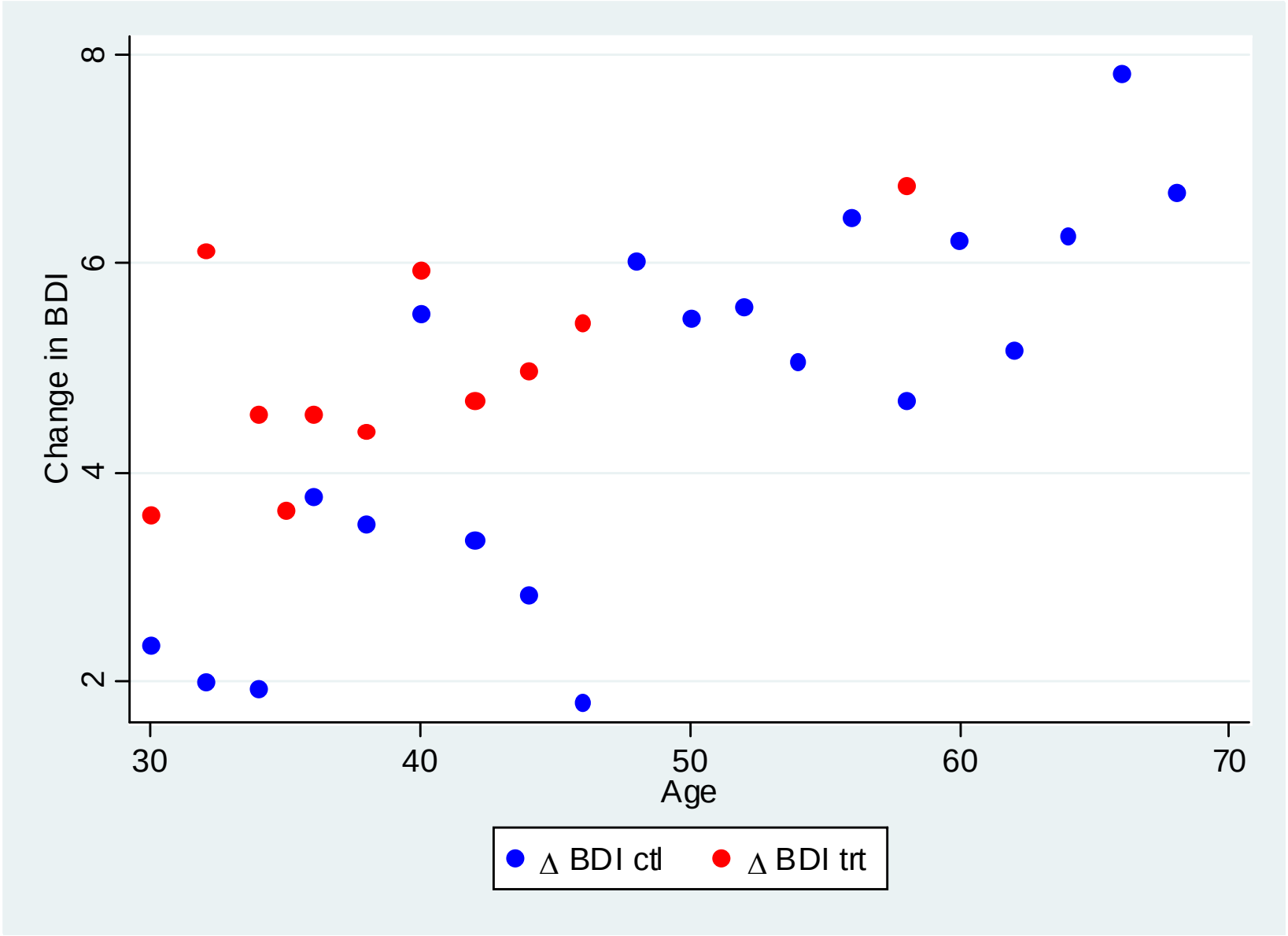


Often, a reasonable quantity to estimate is the average causal effect (ACE): the average of the individual causal effects across the entire population.

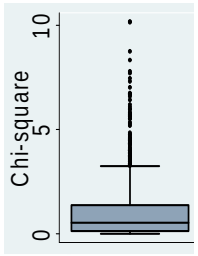
Or perhaps the ACE in a subset of the population. E.g., the causal effect of a smoking cessation program among smokers.

Example: Depression Data





Simple analysis: t-test



Average change in Treatment group = 5.0,

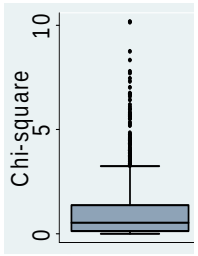
Average change in Control group = 4.6

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

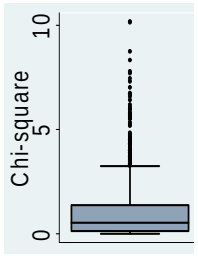
a) age is associated with the outcome (change in BDI) *and*

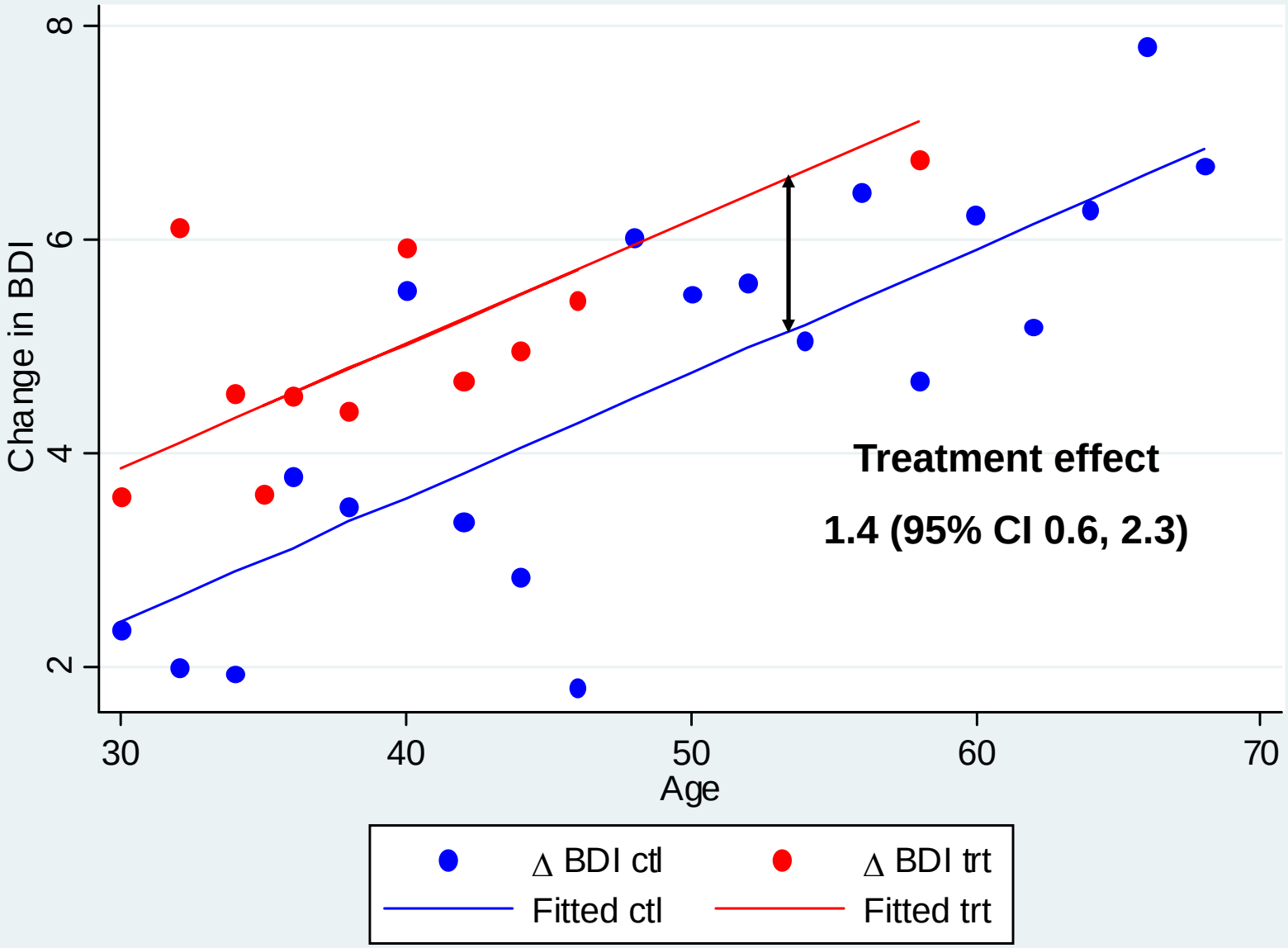
b) age is associated with treatment

A traditional solution:

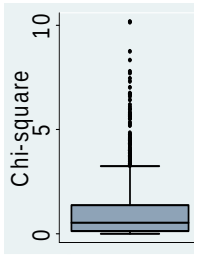
Include age in a multipredictor linear regression model along with treatment.

Example


Chi-square
0
5
10


Change in BDI
2
4
6
8
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).

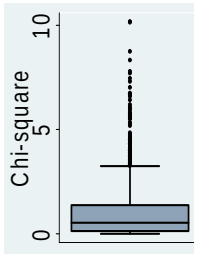
What if model is wrong? As examples: assumption of a linear relationship when it is not. Left out variables? Interactions between variables?

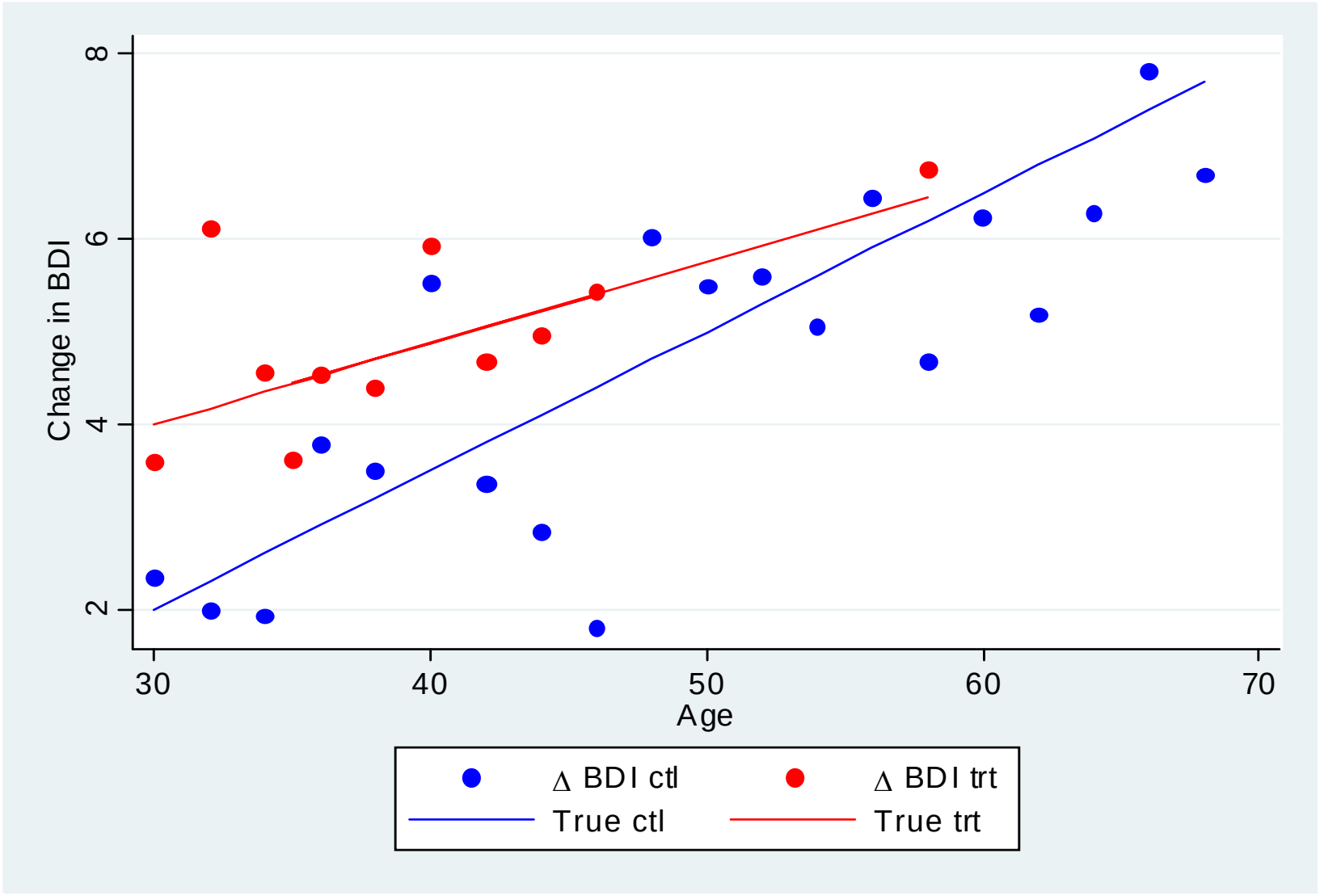
How will we know?

(Lack of overlap/extrapolation.)

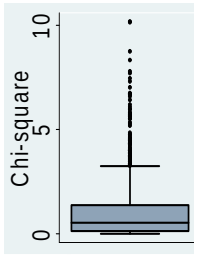
Lack of comparison group for older ages (plenty of controls, not many treated).

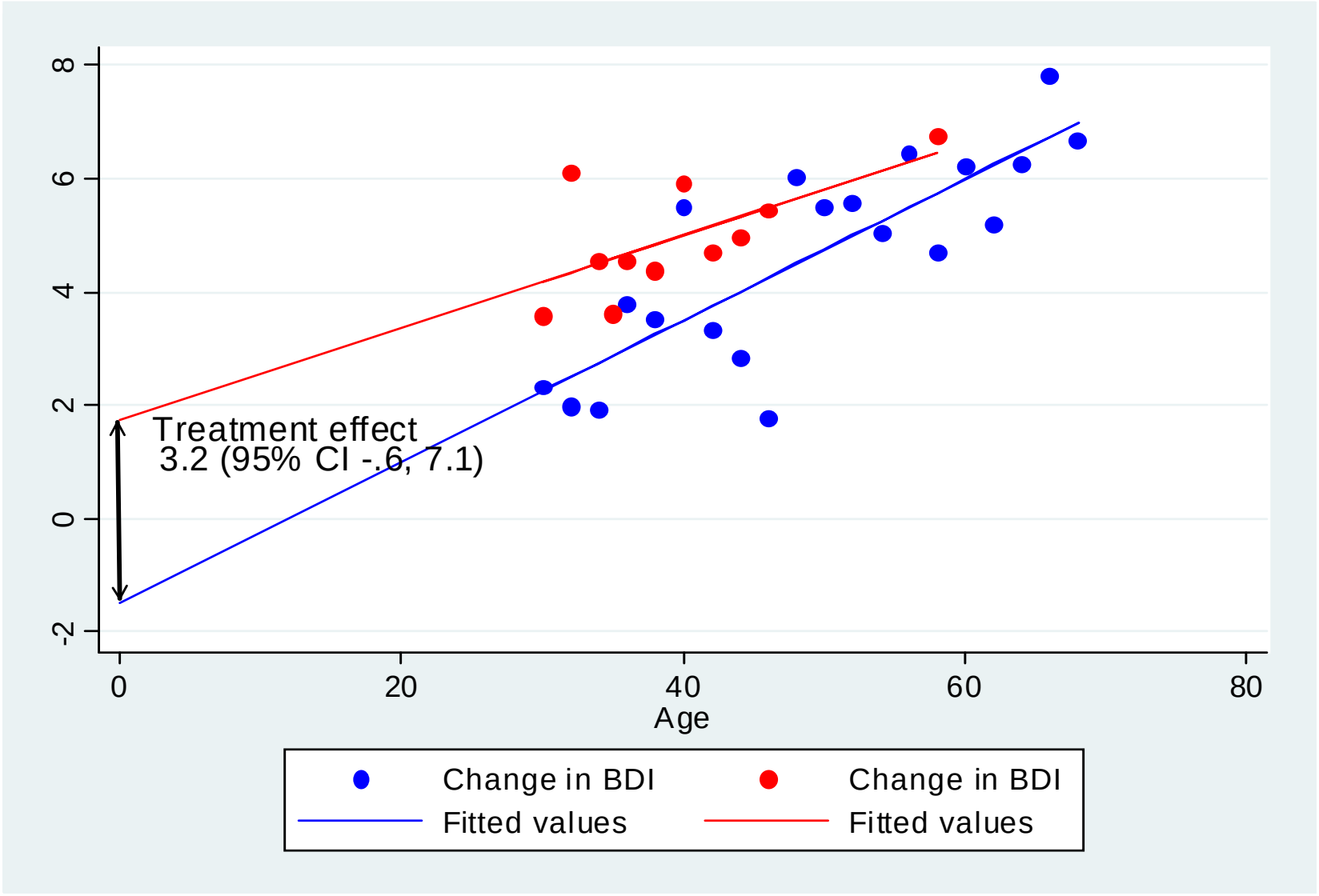
Issues with regression adjustment (true model)



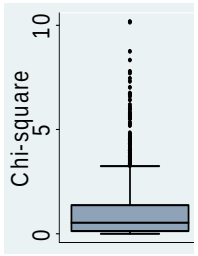


Issues with regression adjustment (fit interaction)





Regression adjustment

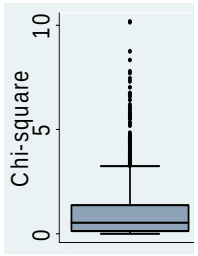


- To fix the issue in this linear regression situation can just center age. Use

  cage = age-Ave(age) = age-42.7

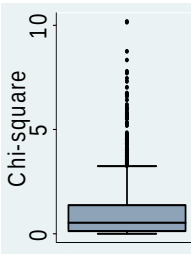
  as a predictor instead of age in the model.
- Then the treatment effect is estimated to be 1.4 (95% CI 0.5, 2.2).
- But this points out the danger in using a statistical model to *define* the causal effect.

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

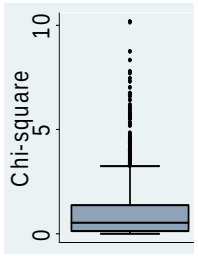


Example: Prediction of healthy aging. Data source: Health ABC – a long running cohort study that enrolled about 3,000 healthy individuals in their 70s.

Outcome: Survival free of major diseases and disability.

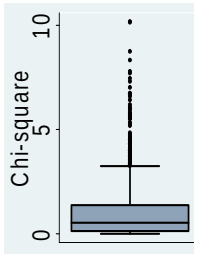
Predictors: About 100 predictors including demographics, BP, medical history, physical performance (e.g., gait speed), and serum markers (e.g., IL-6).

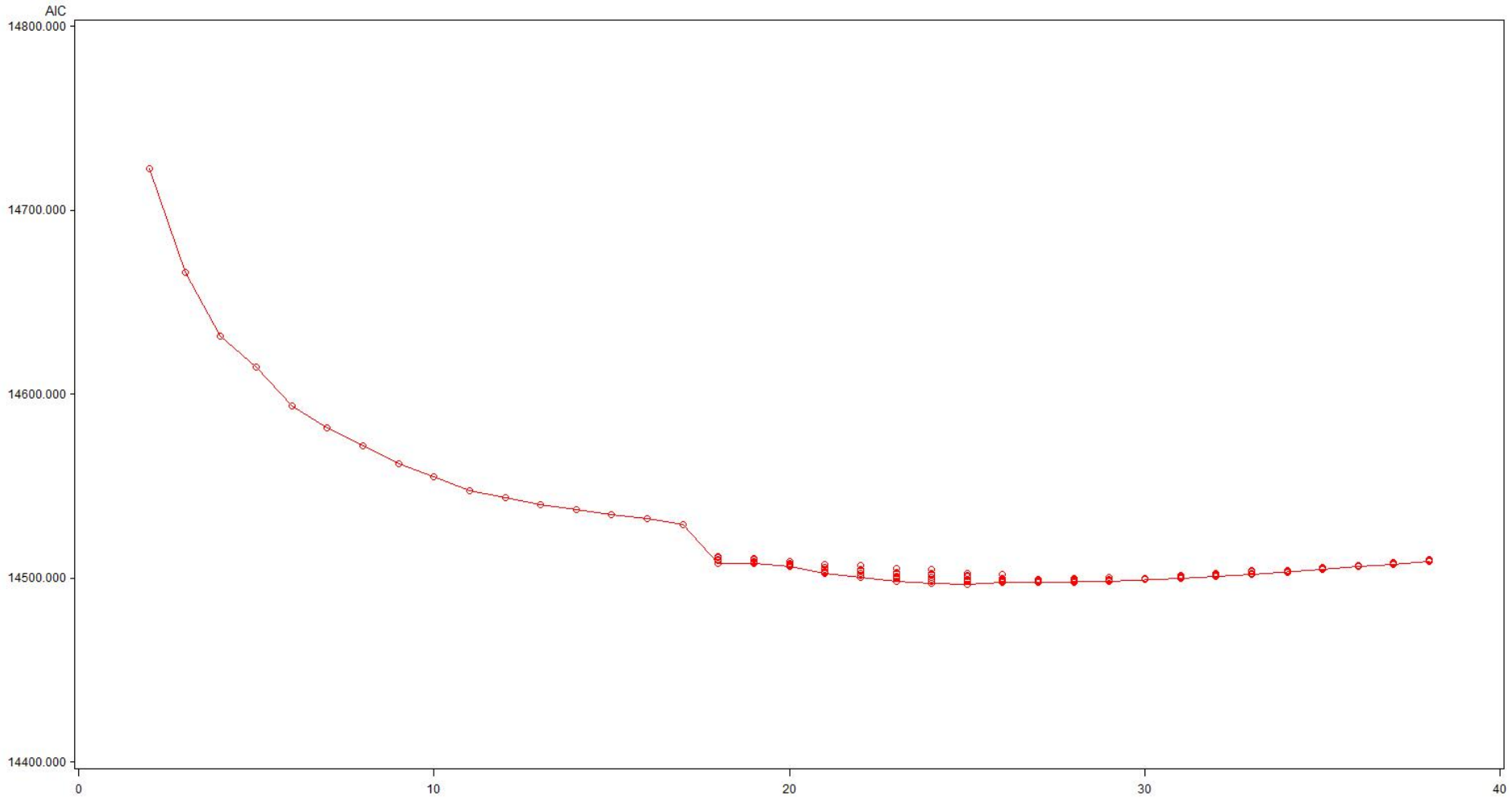
Intepretability



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.

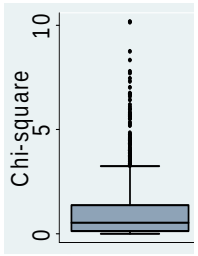
Example: Disease free-survival
AIC plot

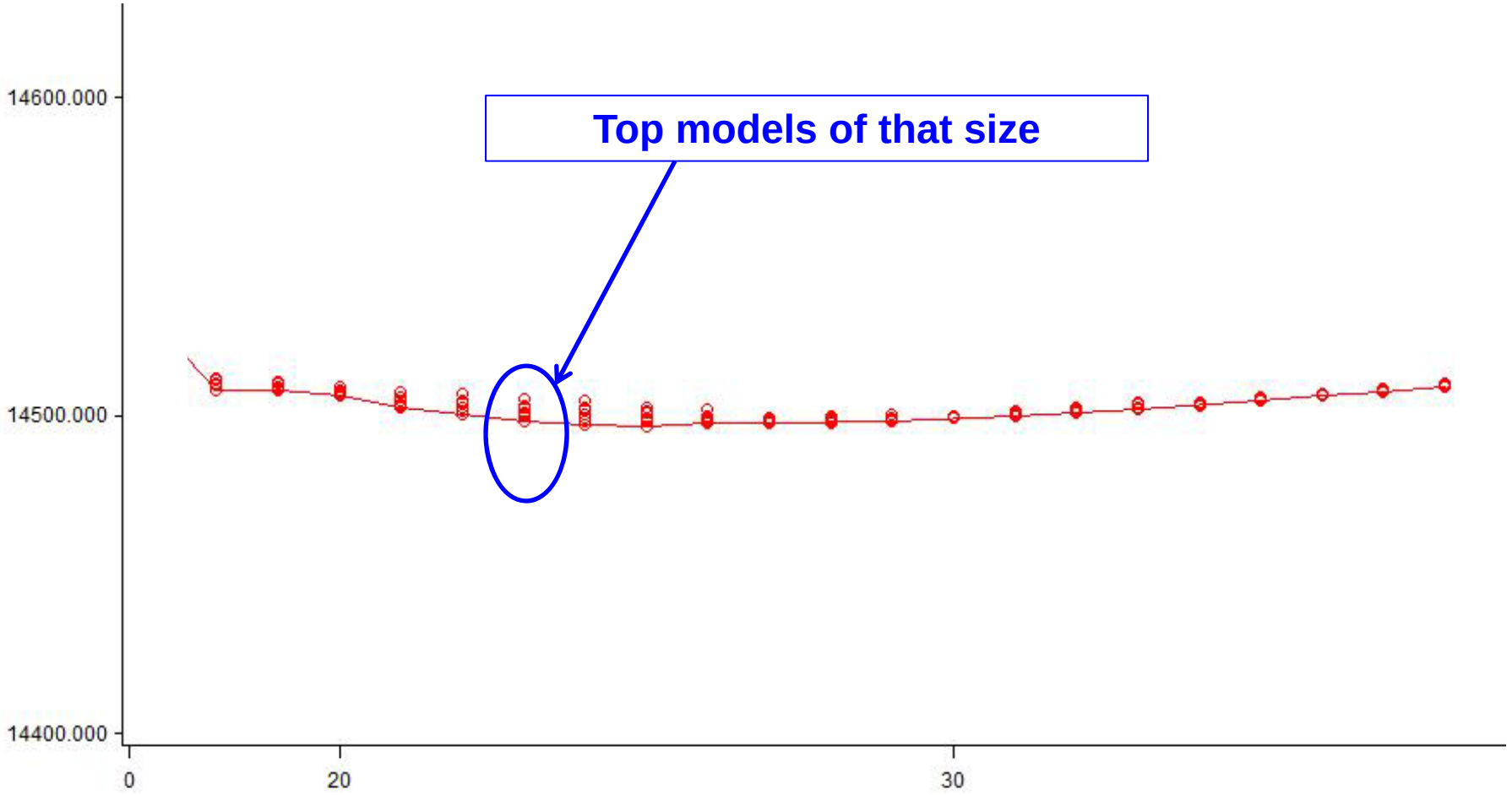




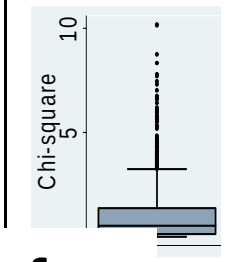
Number of predictors

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





Machine learning/causal inference

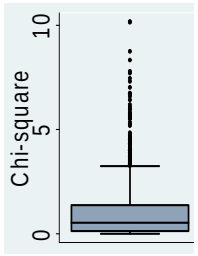


Results: About 120 models were within 10 of the model with the lowest (best) AIC. (Rule of thumb: statistically speaking AIC larger by 10 is a worse fitting model).

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:

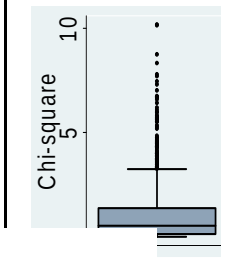
19 variable model: BLAGE WTK HGT WGTPDB HGTDB PULSE BESFEV BPAAIT CESD PACE400 PACE20 GRIP SEMITND TANDSTD COMP FALL RDW ALBUMIN STPCK

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

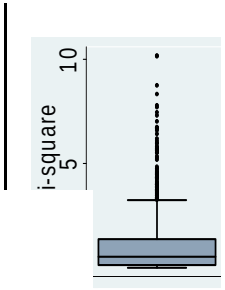
Red indicates not in the other model

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



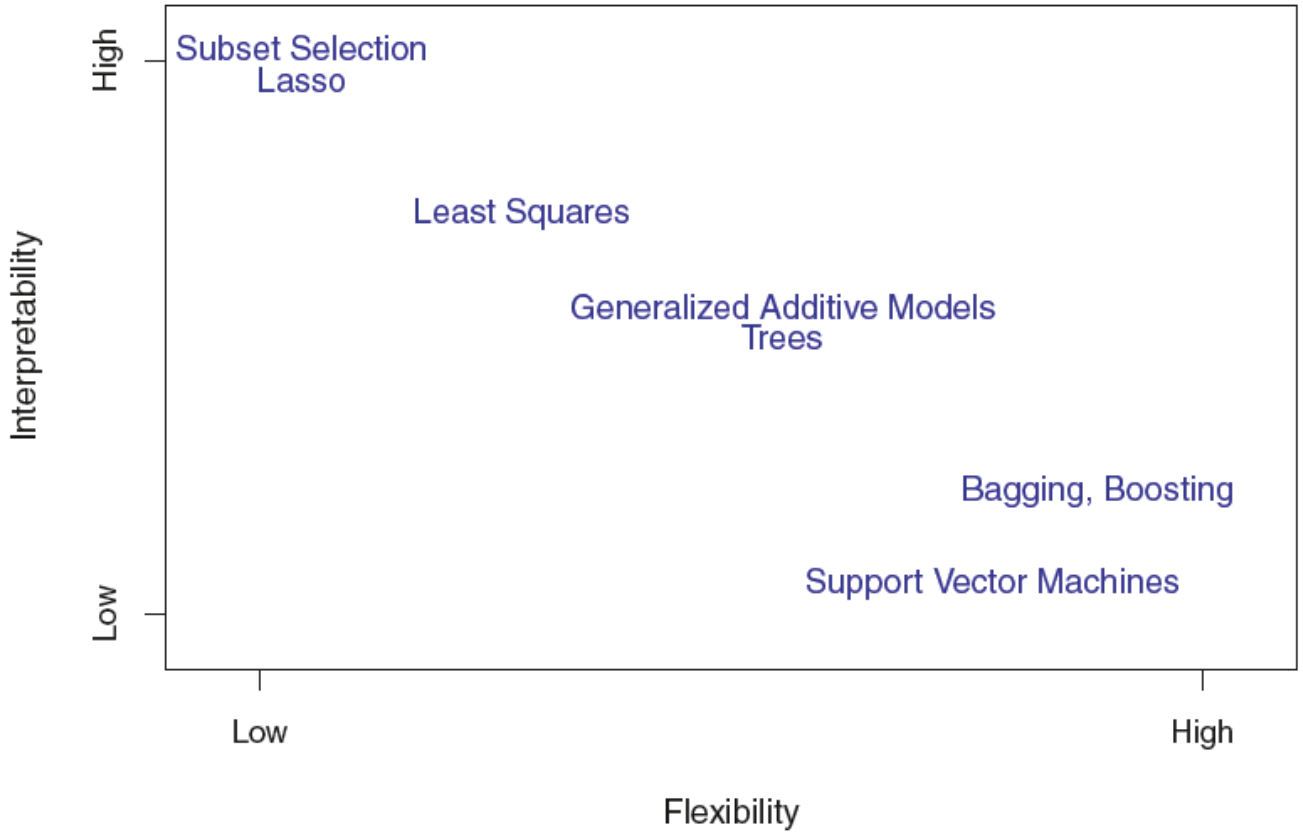
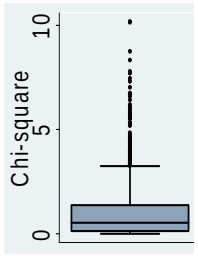


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

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

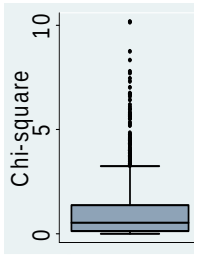
Outline



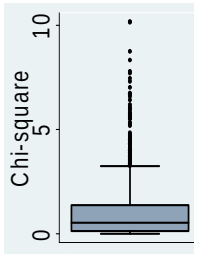
- Introduction: prediction versus causation
- Threats to causal conclusions
- Causal methods in (very) brief
- Interpretability of machine learning methods
- Some uses of big data and machine learning in causal inference
- Summary

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



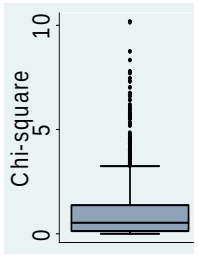
One use of big data for causal inference:



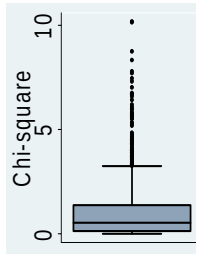
- Pragmatic randomized clinical trials
- Lecture on Monday by Mark Pletcher.

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 on Monday)
  - **Propensity scores**
  - Potential outcomes estimation
  - Subgroup analyses



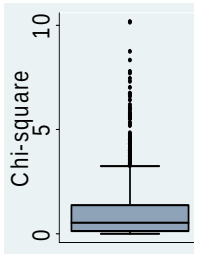
Propensity scores:

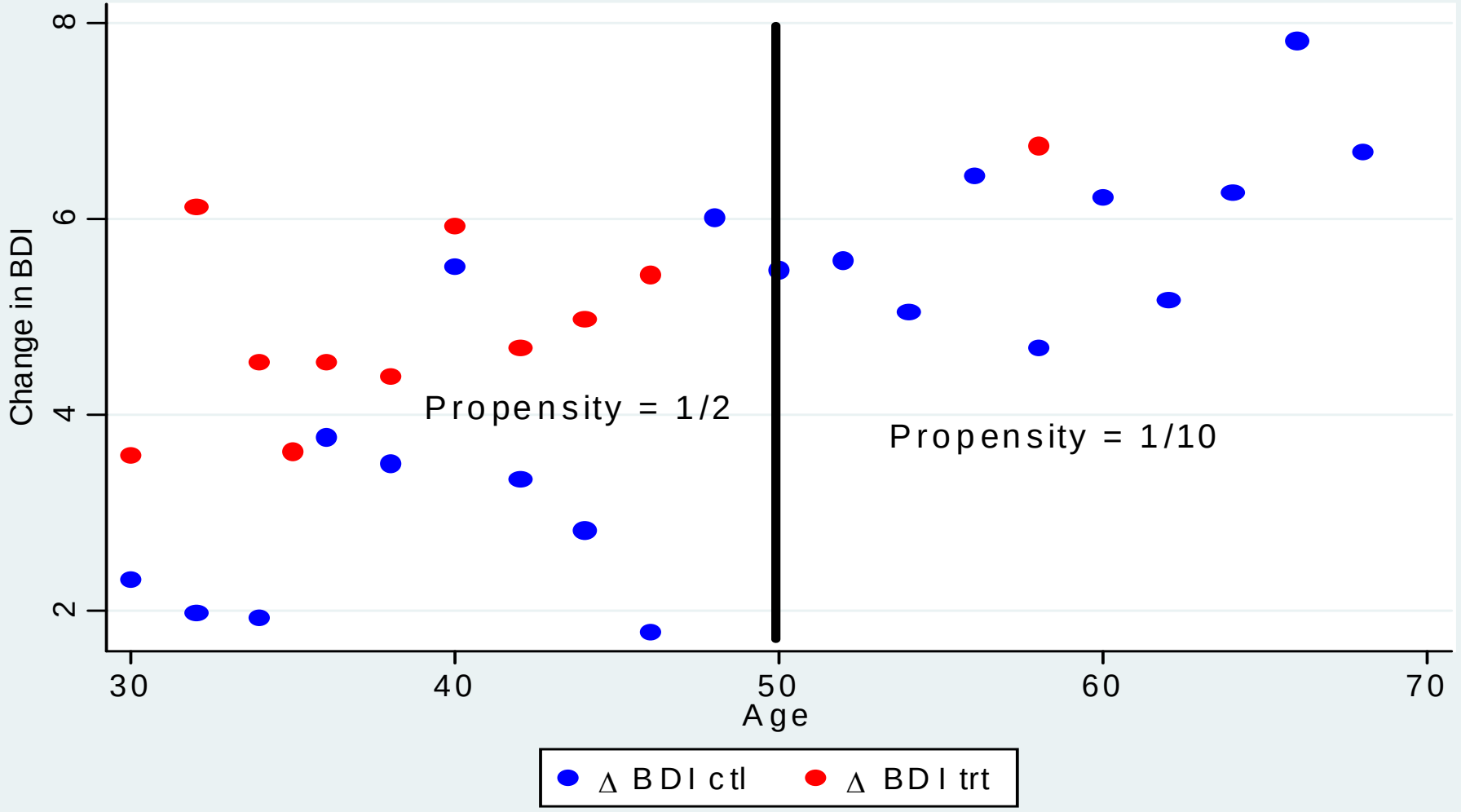


Define prop(x) be the probability of being on treatment as a function of x, all the variables that determine treatment.

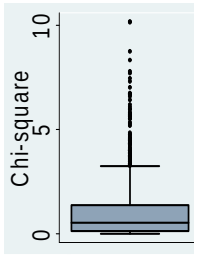
In our depression example, suppose temporarily that the probability of selecting treatment only depends on age.

Propensity scores: example



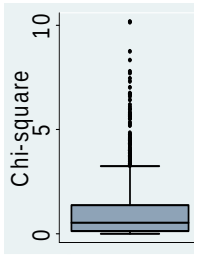


Propensity scores: theory



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 all values of the predictors are completely balanced between the two groups (for a given value of the propensity score).*
2. Therefore you only need to adjust for prop(x) to balance the groups.

Propensity scores:



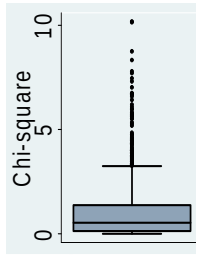
<u>Adjusting for propensity score categories and using a linear regression model</u>:

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

Contrast previous result:

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

Propensity scores: practical issues



Often divide propensity scores into quintiles in order to adjust.

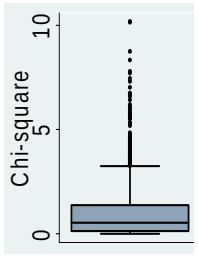
What if not all the variables that determine treatment are measured? Or included correctly in the model?

Suggests being more inclusive with both predictors and interactions.

And to handle continuous predictors with flexible functional forms.

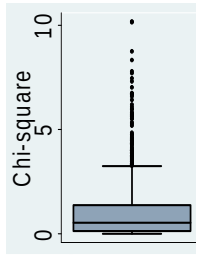
So something that *is* easier with large databases and machine learning.

Propensity score estimation using IBM SPSS Modeler



- 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 Type Node to set Treatment as the Target variable temporarily).
- The propensity scores are these predicted probabilities.

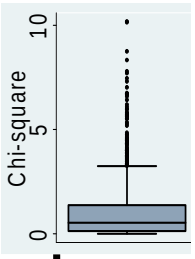
Propensity score estimation using IBM SPSS Modeler



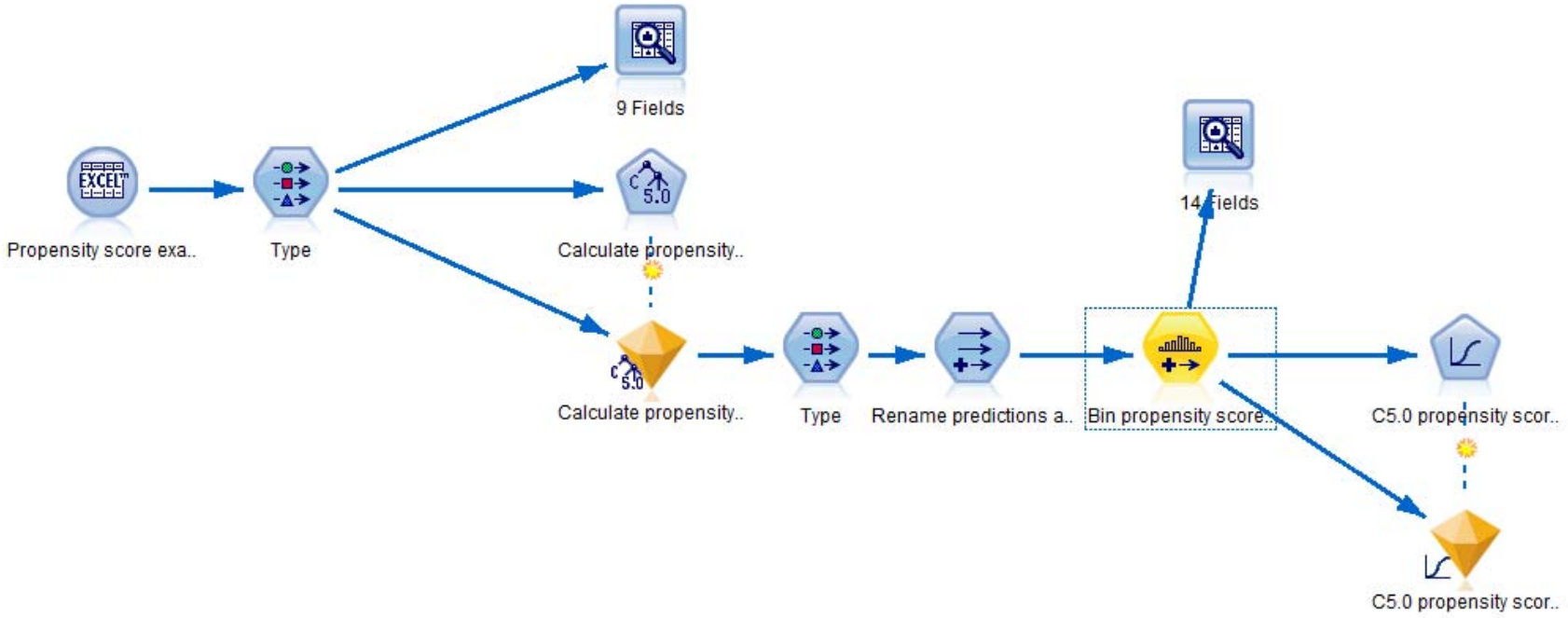
Use the predicted probabilities in a standard statistical analysis:

- For example, bin the propensity scores into five equal categories using a Bin Node (bin method = Tiles (equal count), quintiles) and then use a logistic regression node (for a binary outcome) or a linear regression model node (for a numeric outcome).
- The only predictors are the propensity score bins (as a nominal variable) and treatment.

Propensity score example

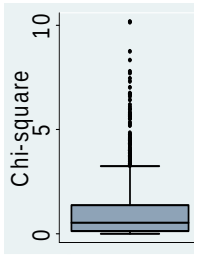


I have posted an example Modeler stream and dataset on the website.

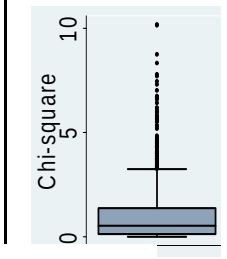


Outline

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



Regression estimation

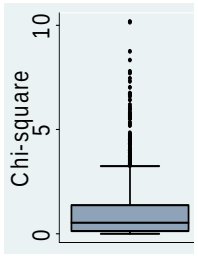


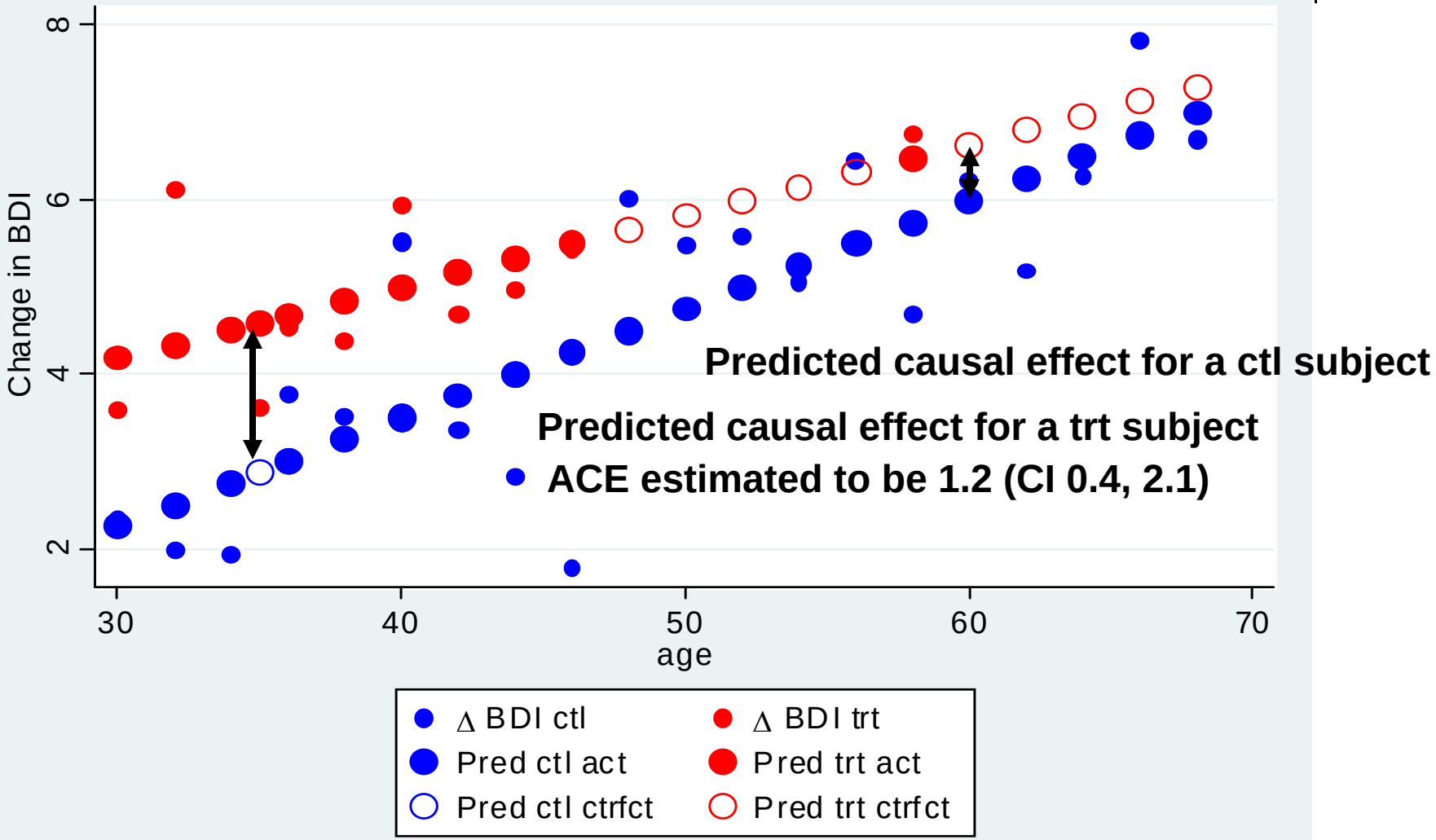
In the depression example, we used a linear regression model. We can use that to get an estimate of the causal effect for each person. Then we could calculate the average causal effect.

Also called *G-estimation* in the causal literature and *marginal estimates* in economic statistics.

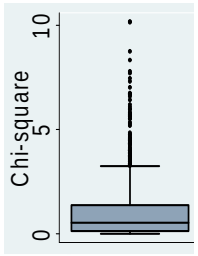
Can get marginal estimates for subpopulations, e.g., causal effect in users of the intervention or in younger participants.

Regression estimation





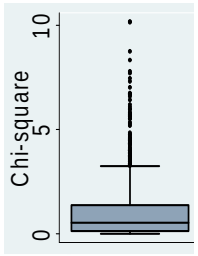
Potential outcomes estimation using IBM SPSS Modeler



- 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	

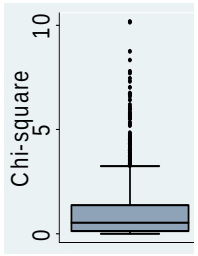
Potential outcomes estimation using IBM SPSS Modeler



- Want to end up with a predicted outcome under treatment and under control:

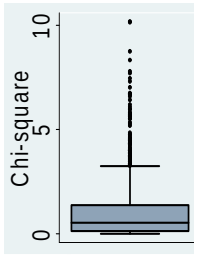
Patient_ID	Sex	Treatment	Systolic_BP	Pred_under_Trt	Pred_und_Ctl
1	F	1	120	122	135
2	M	1	135	137	142
3	F	0	142	137	140

Potential outcomes estimation using IBM SPSS Modeler



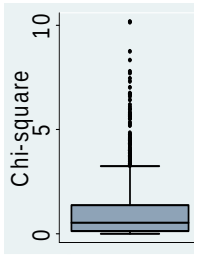
- Read the data using a Source Node.
- Read in a duplicate version of the data using another Source Node. In this part of the stream, switch the coding of treatment and control and recode the outcome variable to missing.
- *Append* the duplicate data to the original data. So now the dataset is twice as large as the original and has a copy of every person under control and under treatment (though only one of them has the outcome – the other is missing).

Potential outcomes estimation using IBM SPSS Modeler



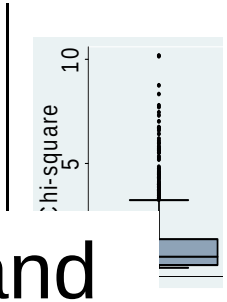
- Separate the data into data for the treatment group and control groups.
- Fit your favorite machine learning method to predict the outcome in the treatment group and do the same for the control group. This will give predictions for the data that were missing previously.
- *Merge* the datasets together so everyone has a predicted outcome under the treatment and under the control conditions.

Potential outcomes estimation using IBM SPSS Modeler

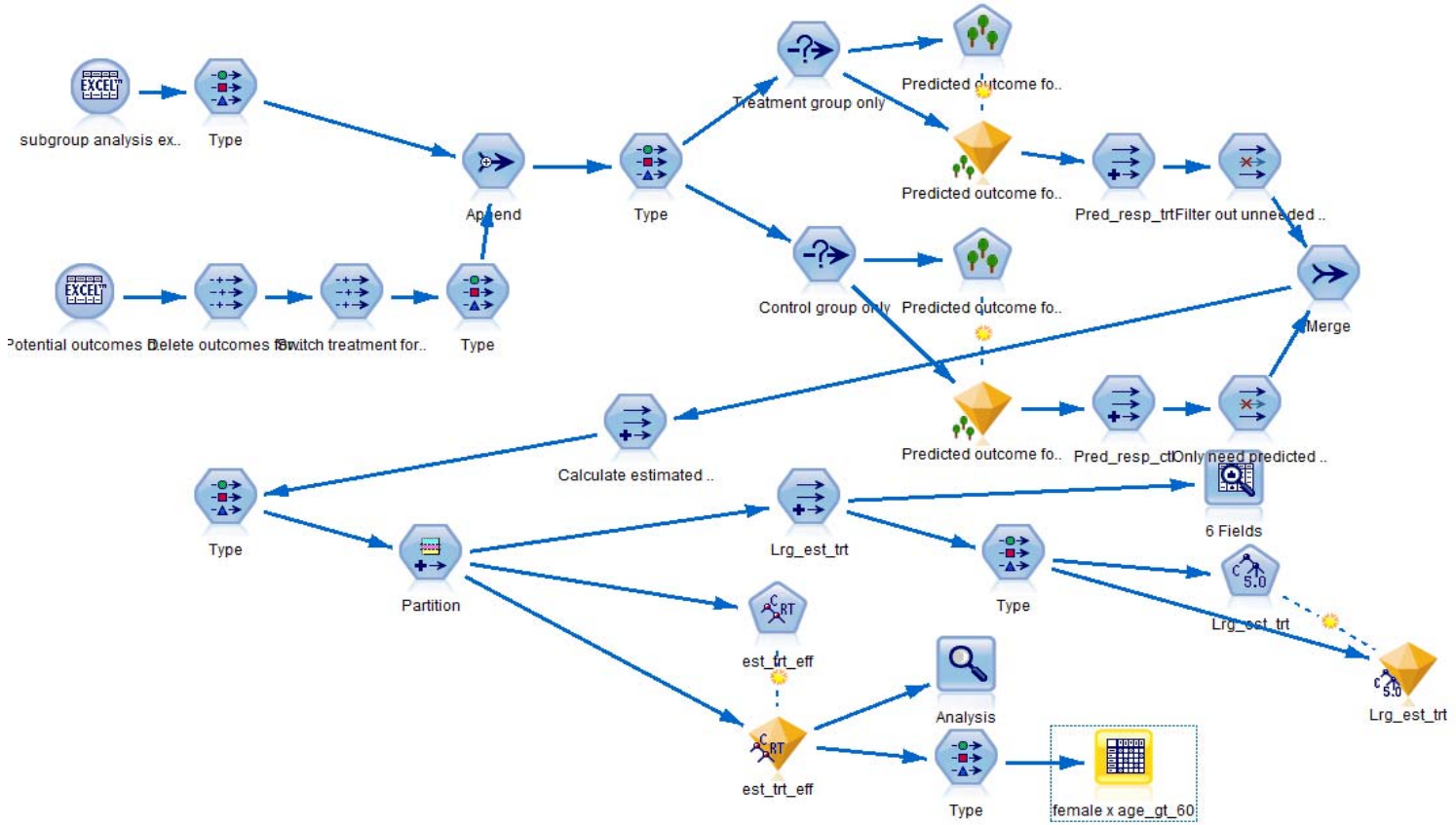


- Average all the differences between the under-treatment and under-control conditions to estimate the average causal effect (ACE).
- Or the ACE in different subgroups.

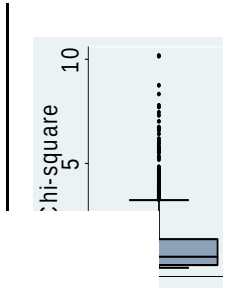
Potential outcomes estimation



I have posted an example Modeler stream and dataset on the website illustrating the idea for this and the next section (subgroup analysis).

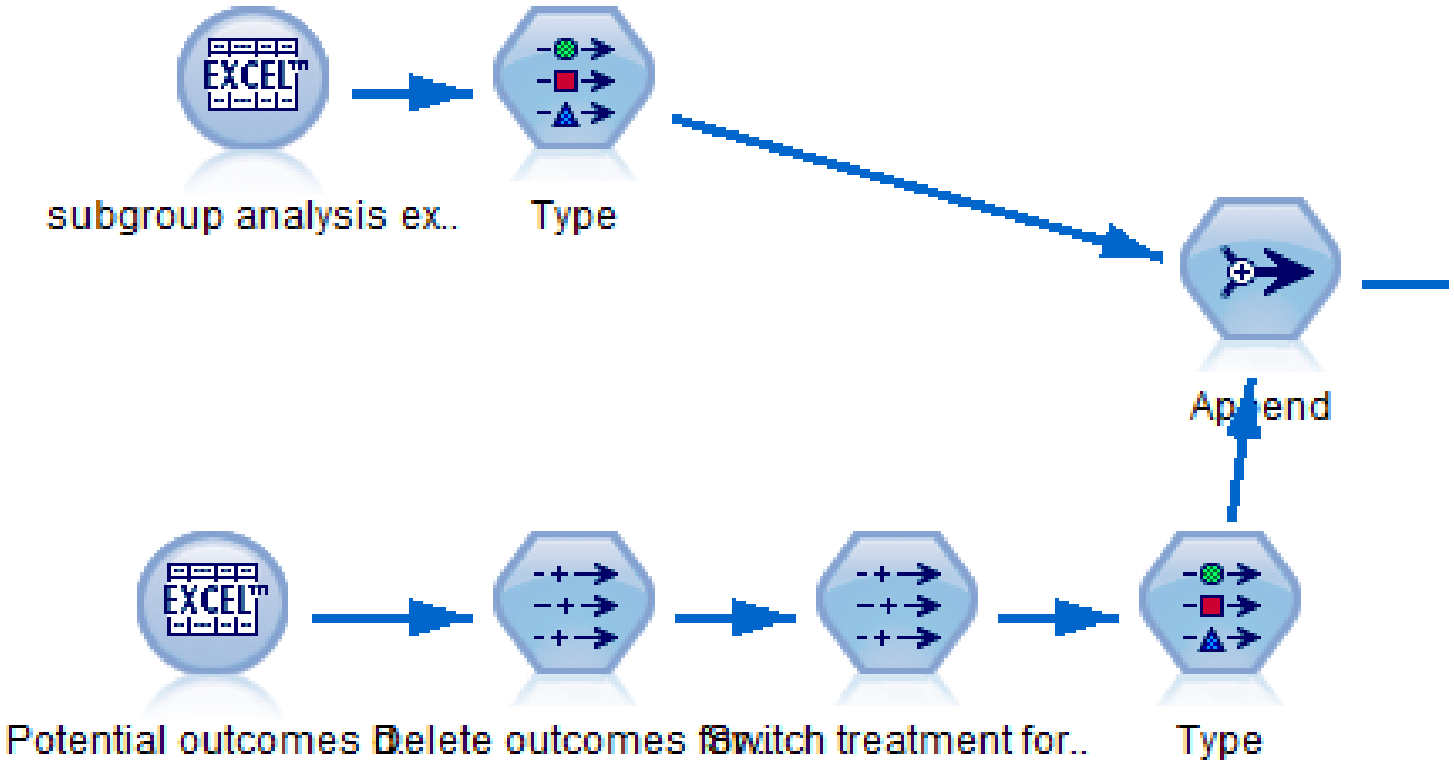


Potential outcomes estimation

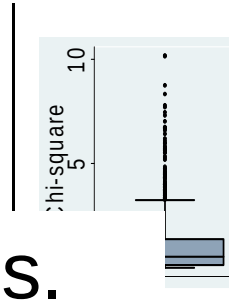


Looks complicated, but let's break it down.

1. Create duplicate records

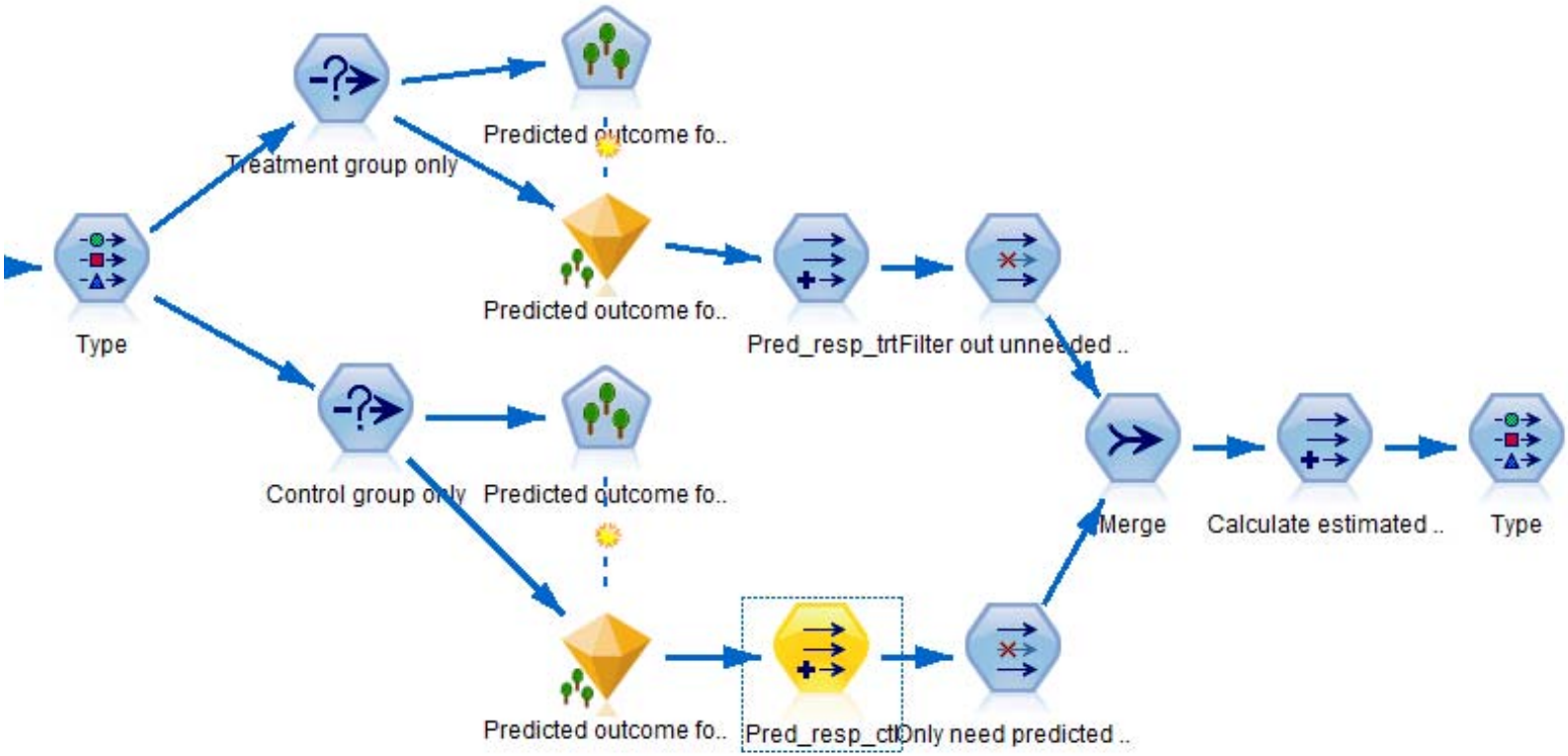


Potential outcomes estimation



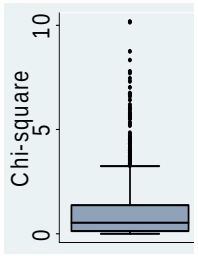
Looks complicated, but let's look at the steps.

2. Estimate effects in treatment and control



3. (Next) Use potential outcomes to get ACE

Potential outcomes estimation using IBM SPSS Modeler

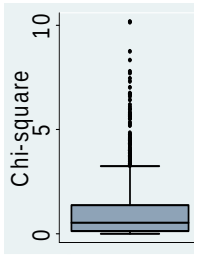


Issues

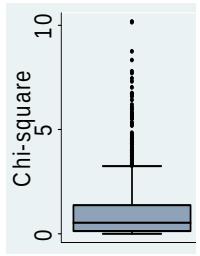
- (+) Can use flexible models in each condition to help avoid issues with incorrectly specifying the model.
- (-) Requires large sample sizes in treatment and in control groups.
- (0) Does not solve the problem of "unmeasured confounders". Not surprisingly, you can only adjust for measured quantities.

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 on Monday)
  - Propensity scores
  - Potential outcomes estimation
  - **Subgroup analyses**

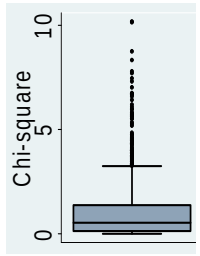


Subgroup effects and precision medicine



- 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

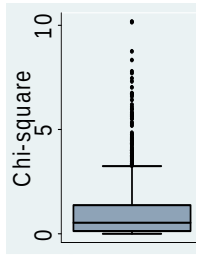


- Race (5 categories), birth year (say 3), sex (2), Down's syndrome (2), chromosomal (2) or other congenital anomaly (2), birth weight (6), gestational age (say 3), Apgar score (2), direct antiglobulin test positive (2), maternal age (2), delivery mode (2), multiple birth (2), jaundice (2).
- So $5\times3\times2\times2\times2\times2\times6\times3\times2\times2\t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

  $=276{,}480$ subgroups

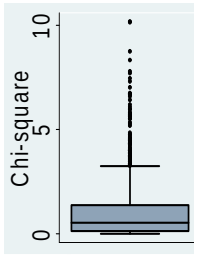
  (for 500,000 babies)

Subgroup analysis issues



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

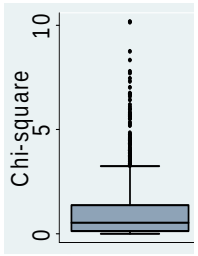
Subgroup problem – proposed solution



- 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.
- Use a recursive partitioning algorithm to find subgroups with different treatment effects (variation: code treatment effects as large/not).

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

Summary



- Distinguish questions by whether they can be answered by prediction only or require more detailed analysis.
- 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
  - Predicting potential outcomes
  - Identifying subgroups on the basis of different estimated treatment effects