Machine learning in healthcare: causal inference

Uri Shalit
IEM, Technion

Machine learning in healthcare – why now?

- Vast increase in availability of digital health data:
  Hospitals, Clinics, Kupot Cholim (HMOs)
- Personal health devices expected to generate even more data
  - Track people when healthy, not just sick
- Problems in healthcare:
  - Rising costs
  - Over diagnosis
  - Misdiagnosis
  - Mismanagement of chronic disease
- The dream is that ML will help improve health outcomes

Lecture today

- Background – ML in healthcare
- Causal inference primer

Lecture today

- Background – ML in healthcare

- Causal inference primer

ML in clinical practice - three different kinds of tasks

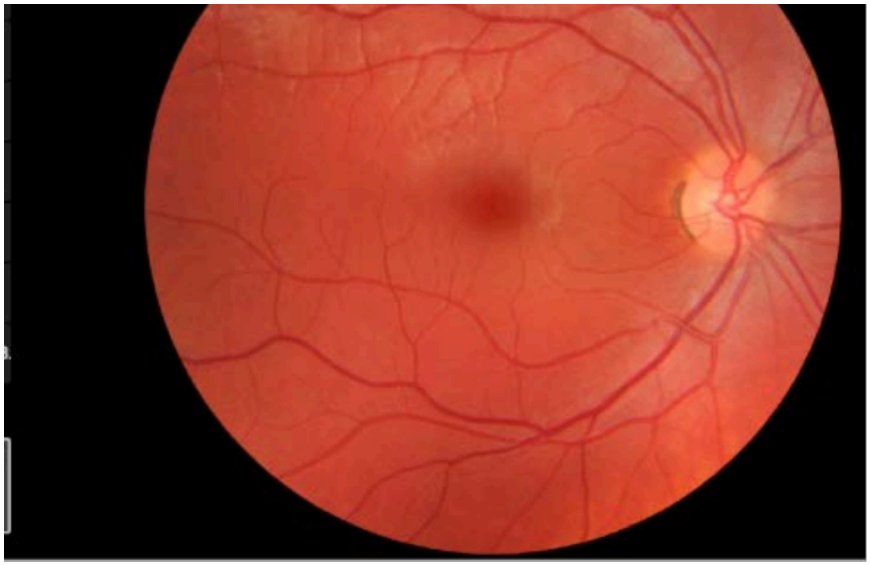
- Perceptual
- Prediction
- Counterfactual prediction

ML in clinical practice - three different kinds of tasks
Perceptual

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD



MENU

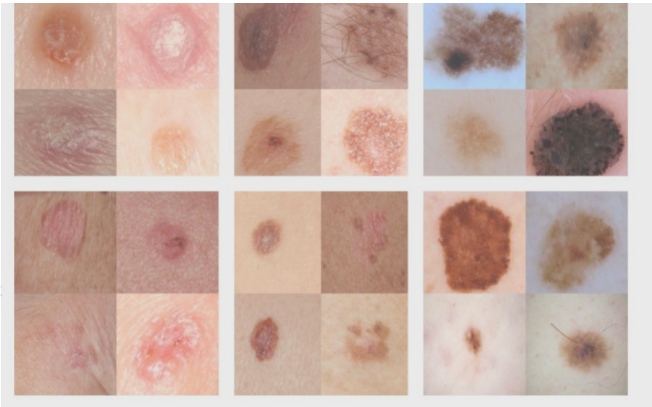
nature
International journal of science

Altmetric: 2555 Citations: 85 More detail »

Letter

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun



ML in clinical practice - three different kinds of tasks

Perceptual

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Co
Subhashini Venugopalan, MS; Kasumi Widner, MS
Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessic

MENU

nature
International journal of science

Altmetric: 2555 Citations: 85 More detail »

sification of skin
networks

in Ko, Susan M. Swetter, Helen M. Blau &

Goal: identify visual patterns as well as the best doctors

ML in clinical practice - three different kinds of tasks

Perceptual

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Co
Subhashini Venugopalan, MS; Kasumi Widner, MS
Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessic

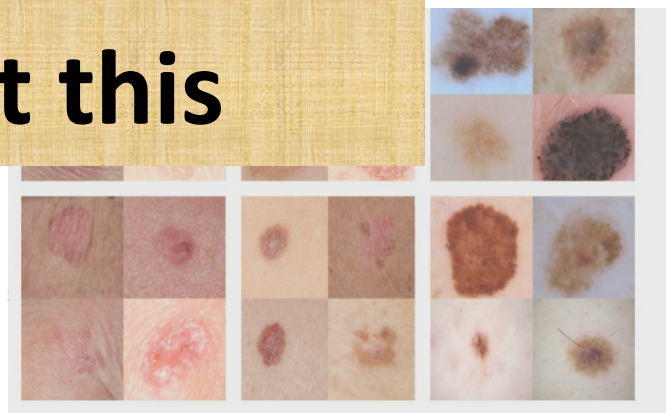
MENU

nature
International journal of science

Altmetric: 2555 Citations: 85 More detail »

sification of skin
networks

in Ko, Susan M. Swetter, Helen M. Blau &

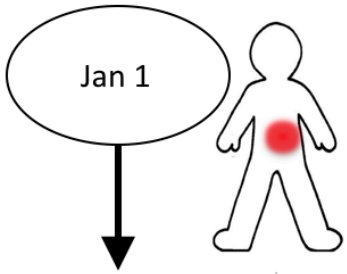


Goal: identify visual patterns as well as the best doctors

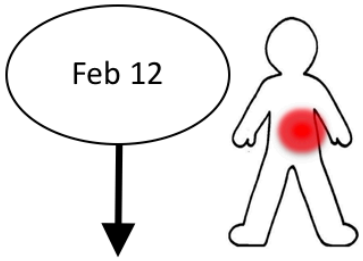
Deep nets excel at this

ML in clinical practice - three different kinds of tasks
Prediction

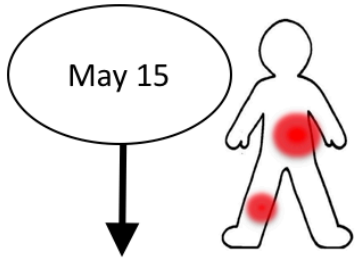
- Given a patient's medical record, use 1,000,000 other patients' records to predict what will her future be



- Blood pressure = 130
- WBC count = $6*10^9$/L
- Temperature = 98°F
- A1c = 6.6%
- Precancerous cells = 10^4
- # flu viruses = 10^6
- Thickness of heart artery plaque = 3mm



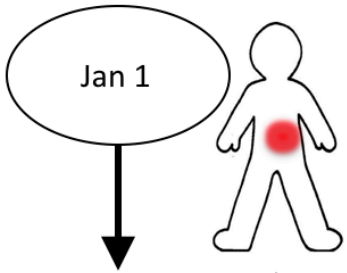
- Blood pressure = 135
- WBC count = $5.8*10^9$/L
- Temperature = 99°F
- A1c = 7.1%
- Precancerous cells = 10^4
- # flu viruses = 10^6
- Thickness of heart artery plaque = 3mm



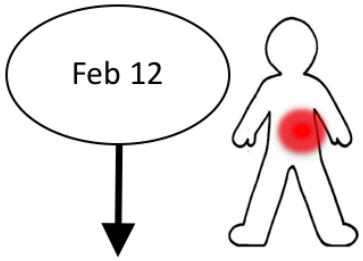
- Blood pressure = 150
- WBC count = $6.8*10^9$/L
- Temperature = 98°F
- A1c = 7.7%
- Precancerous cells = 10^4
- # flu viruses = 10^7
- Thickness of heart artery plaque = 3.5mm

ML in clinical practice - three different kinds of tasks
Prediction

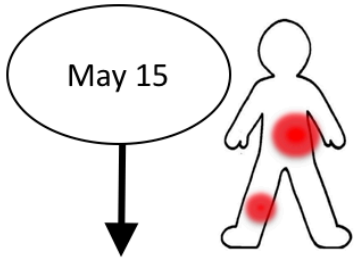
- Given a patient's medical record, use 1,000,000 other patients records to predict what will her future be ***under current practice***



- Blood pressure = 130
- WBC count = $6*10^9$/L
- Temperature = 98°F
- A1c = 6.6%
- Precancerous cells = 10^4
- # flu viruses = 10^6
- Thickness of heart artery plaque = 3mm



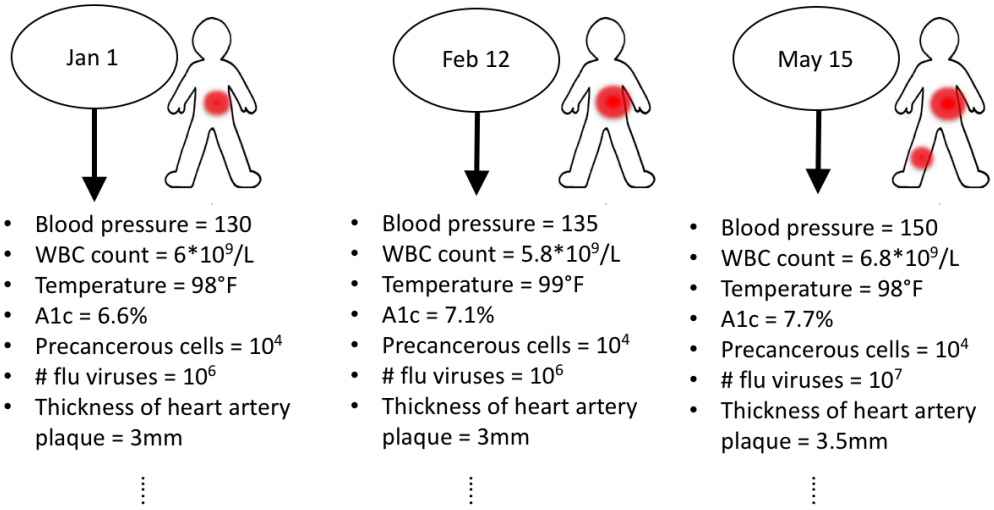
- Blood pressure = 135
- WBC count = $5.8*10^9$/L
- Temperature = 99°F
- A1c = 7.1%
- Precancerous cells = 10^4
- # flu viruses = 10^6
- Thickness of heart artery plaque = 3mm



- Blood pressure = 150
- WBC count = $6.8*10^9$/L
- Temperature = 98°F
- A1c = 7.7%
- Precancerous cells = 10^4
- # flu viruses = 10^7
- Thickness of heart artery plaque = 3.5mm

ML in clinical practice - three different kinds of tasks
Prediction

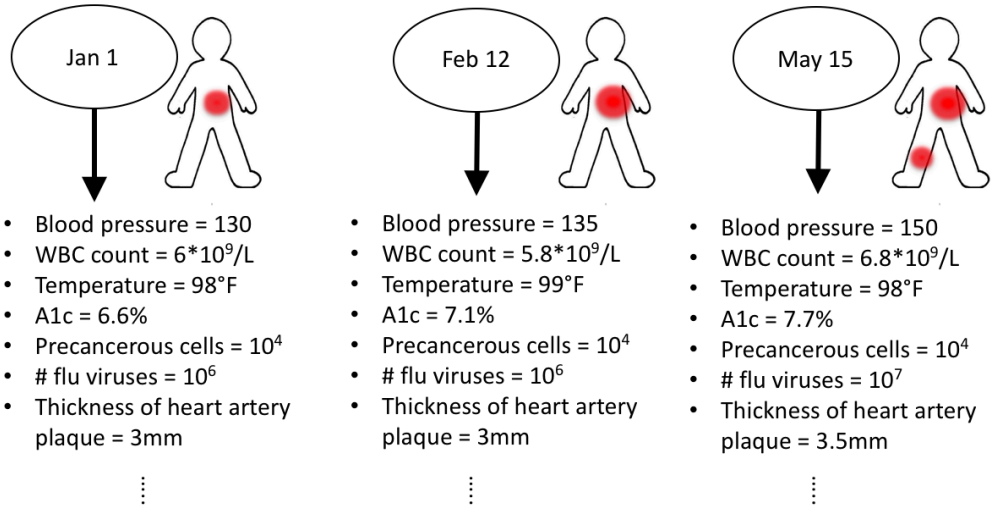
- Given a patient's medical record, use 1,000,000 other patients records to predict what will her future be ***under current practice***
- Classic work: patient risk scores widely used in medicine such as Framingham score, APACHE score and many others. Typically use logistic regression with 3-10 features
- New work: use ML with 1000s of features. Starting to be deployed in certain hospitals.



ML in clinical practice - three different kinds of tasks
Prediction

- Given a patient's medical record, use 1,000,000 other patients records to predict what will her future be ***under current practice***
- Classic work: patient risk scores widely used in medicine such as Framingham score, APACHE score and many others. Typically use logistic regression with 3-10 features
- New work: use ML with 1000s of features. Starting to be deployed in certain hospitals.

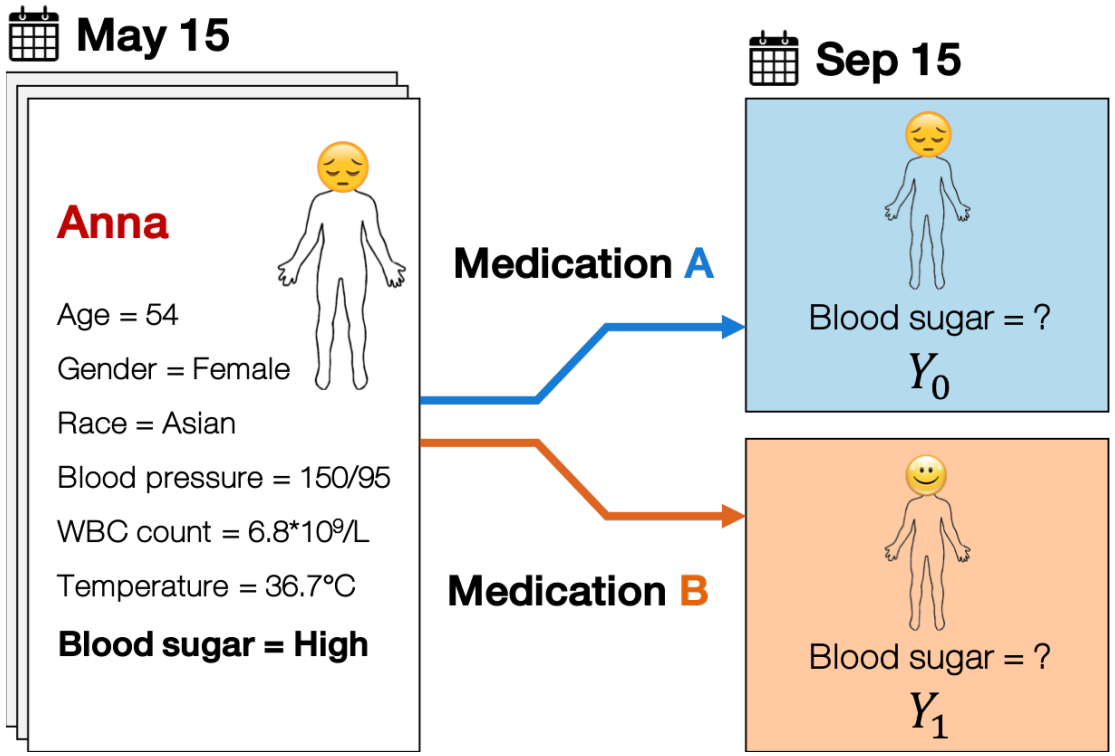
New risk score can change current practice invalidating iid assumptions



ML in clinical practice - three different kinds of tasks
Counterfactual prediction

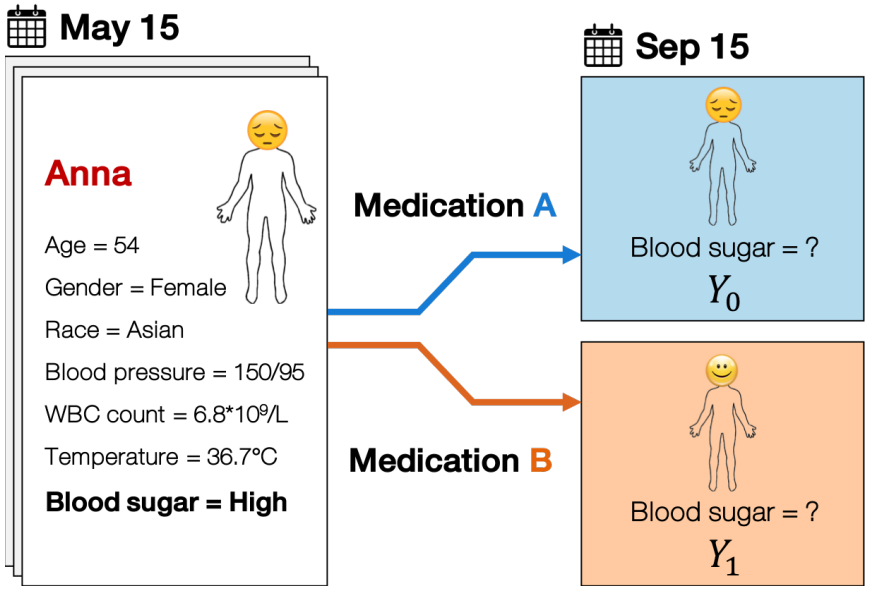
- How to best treat a patient? Might be different from current practice

ML in clinical practice - three different kinds of tasks
Counterfactual prediction



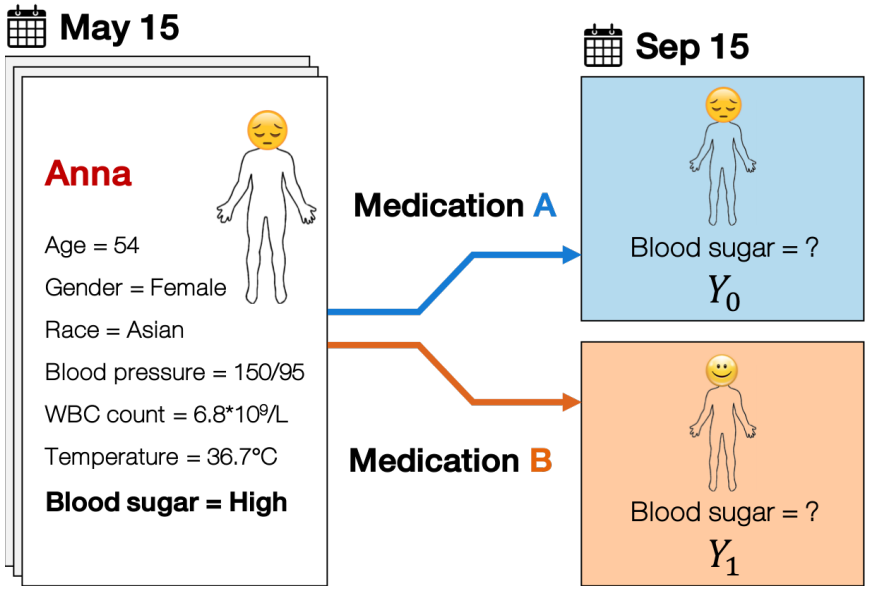
ML in clinical practice - three different kinds of tasks
Counterfactual prediction

- Algorithmically personalized medicine



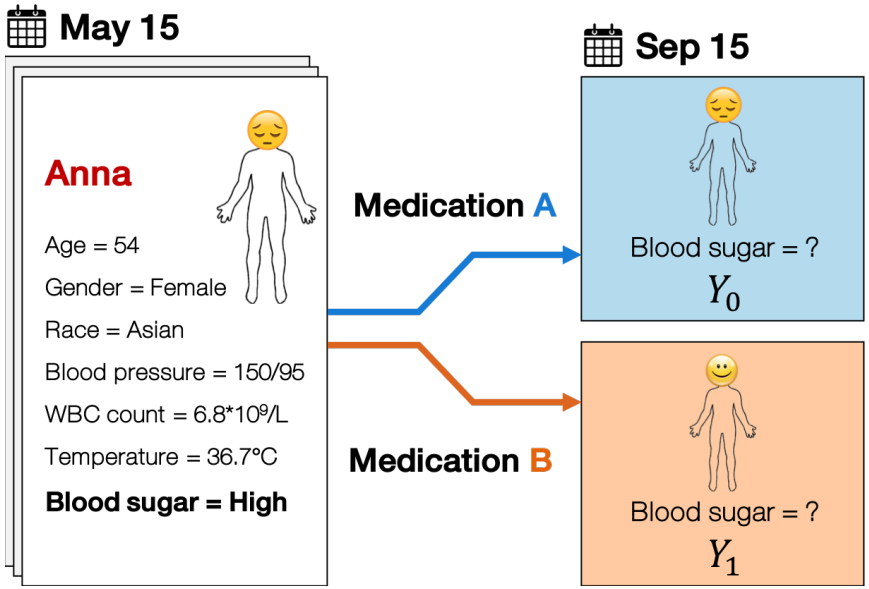
ML in clinical practice - three different kinds of tasks
Counterfactual prediction

- Algorithmically personalized medicine
- In general the difficulty scales from:
  1. People who get A vs. B are completely different →impossible
  2. People who get A vs. B are different because of unknown factors →very hard
  3. People who get A vs. B are different in ways we can reliably model → doable
  4. People get A or B at random → easier



ML in clinical practice - three different kinds of tasks
Counterfactual prediction

- Algorithmically personalized medicine
- In general the difficulty scales from:
  1. People who get A vs. B are completely different →impossible
  2. **People who get A vs. B are different because of unknown factors →very hard**
  3. **People who get A vs. B are different in ways we can reliably model → doable**
  4. **People get A or B at random → easier**



Talk today

- Background – ML in healthcare

- Causal inference primer

Talk today

- Background – ML in healthcare

- Causal inference primer

Counterfactuals as causal inference (Pearl, Rubin)

- A definition of “X causes Y”:

  *if **X** wouldn’t have happened, **Y** wouldn’t have happened*

- Probabilistic version (Pearl, 2009)
  Y** would be **y** had **X** been **x** in situation **U = u

  *“Patient **u**’s blood sugar **Y** would be **y=6.3%** had she received medication **x**”*

Causation from correlation

Causation from correlation

Predict causal effects

on a per-**individual** basis

from **observational** data

in **high** dimensions

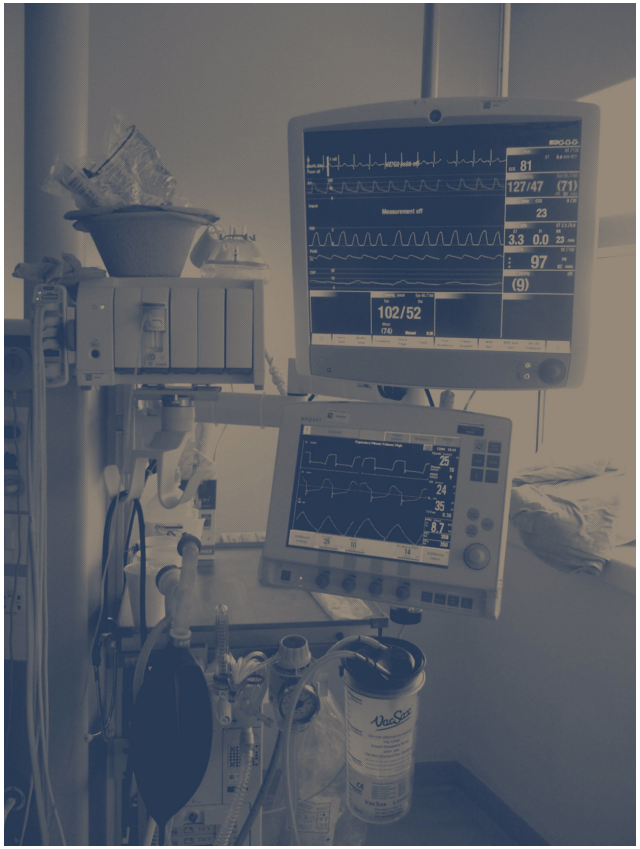
Causation from correlation: when experiments are difficult or impossible

- Healthcare
- Economics: who should get job training? Should we raise minimum wage?
- Ad-placement
- Marketing: which campaigns work? Who should we give coupons to?
- Education: which curriculum works better? personalized education

“Easy”: Prediction

- **Predict diabetes condition 1 year from now** for patient using records of past patients (lab tests, prescriptions, etc.)

- We can predict pretty well (AUC 0.8, Razavian et al. 2015)



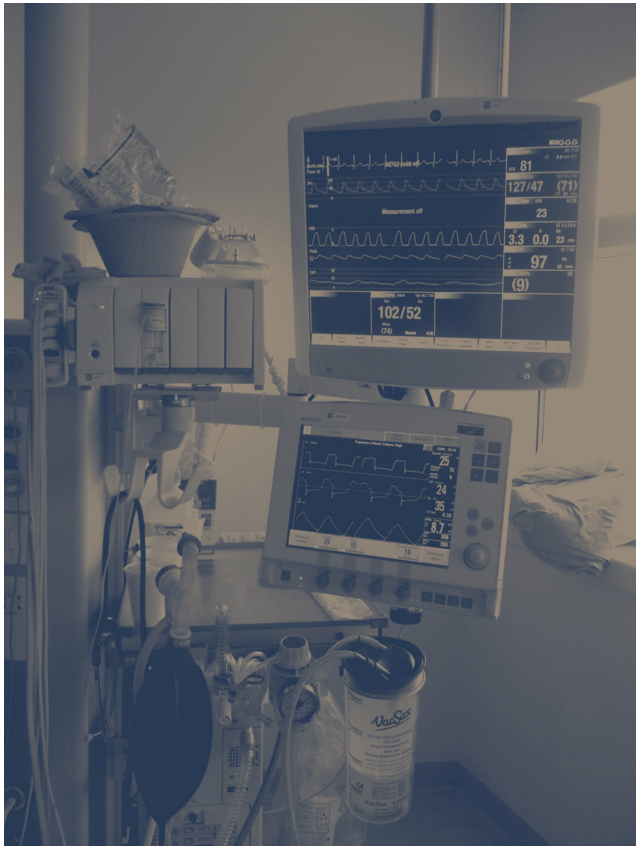
“Easy”: Prediction

- **Predict diabetes condition 1 year from now** for patient using records of
(lab tests,
- We can pr
pretty well
(AUC 0.8, Razavian et al. 2015)

But is prediction what we really want?

Prediction vs. action

- NSAID treatment (e.g. aspirin, ibuprofen) is strong predictor of diabetes onset
- Does this imply I shouldn't take aspirin for fear of diabetes?
- No! The fact that taking NSAIDs is a good predictor of diabetes doesn't mean NSAIDs cause diabetes



What are we doing wrong?

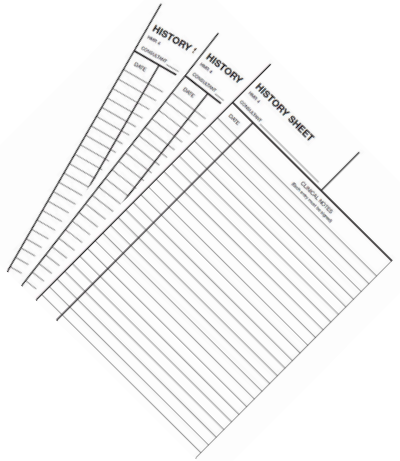
- Taking aspirin is an ***action***
- To take good actions we must understand **causality**
  "Does aspirin cause diabetes?"

- **Predictions** ab
  "Would this pc *hey taken aspirin?"*
- Mistakes can be extremely costly!

Causal inference is about learning the effects of actions

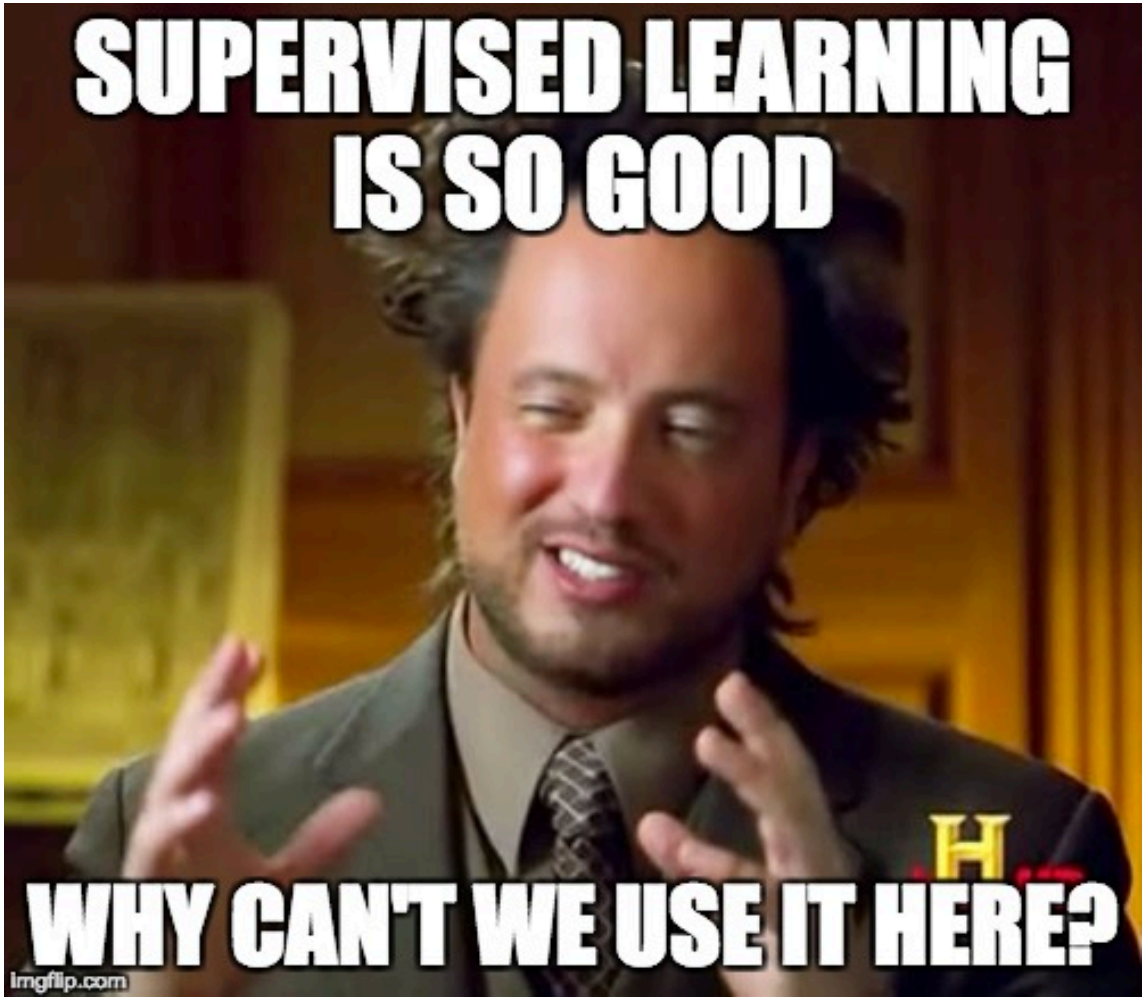
When supervised learning isn't enough

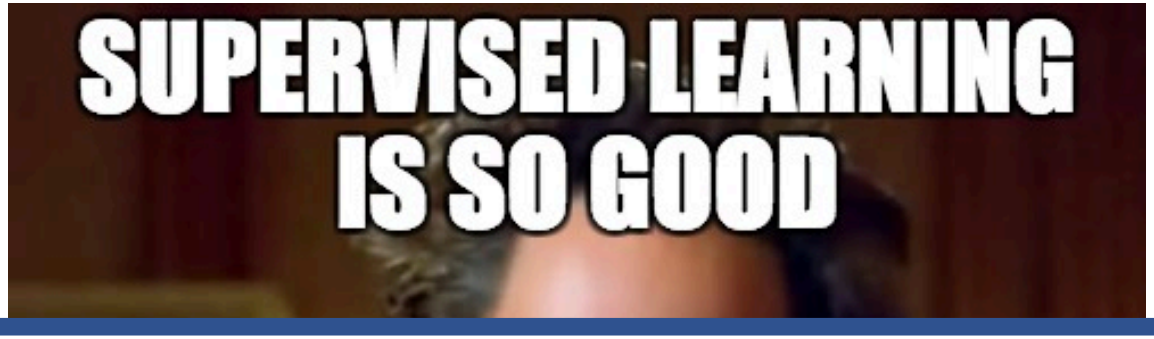
- Dataset of 10,000,000 patients
- Medications, blood tests, past diagnoses, doctors' notes, demographics, genetic testing



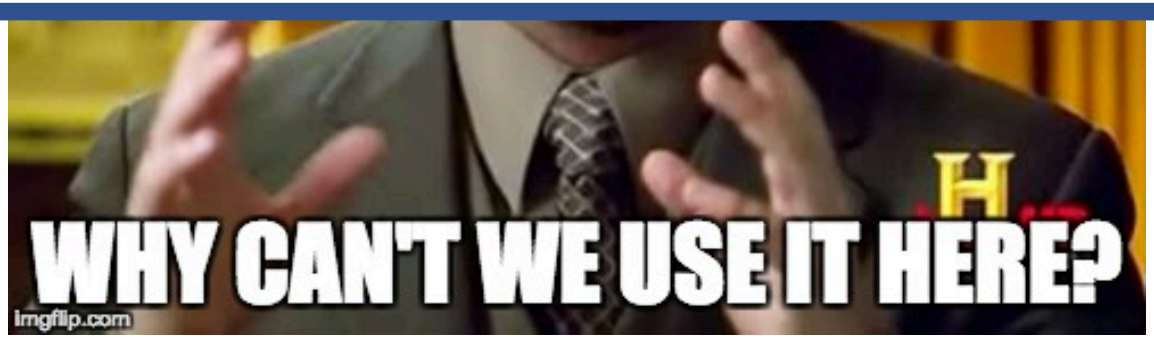
When supervised learning isn't enough

- Patient "Anna" comes in with hypertenstion
  - Asian, 54, history of diabetes, blood pressure 150/95, …
- Which medication will better lower her blood pressure:
  - Calcium channel blocker (A)
  - ACE inhibitor (B)
- I have data from 10,000,000 other patients – surely that can help!


SUPERVISED LEARNING
IS SO GOOD
WHY CAN'T WE USE IT HERE?
imgflip.com



Why can’t we just learn the connection between patient, treatment and outcome with supervised learning (say a neural net)?



When supervised learning isn't enough

- Patient "Anna" comes in with diabetes
  - Asian, 54, history of hypertension, blood pressure 150/95, …
- Which medication will better lower her blood pressure:
  - DPP4 (A)
  - SGLT-2 (B)
- I have data from 10,000,000 other patients – surely that can help!

When supervised learning isn't enough

- Patient "Anna" comes in with diabetes
  - Asian, 54, history of hypertension, blood pressure 150/95, …
- Which medication will better lower her blood pressure:
  - DPP4 (A)
  - SGLT-2 (B)
- I have data from 10,000,000 other patients – surely that can help!
- Build a regression model from patient features to blood pressure

When supervised learning isn't enough

- Build regression model from patient features to blood sugar

- Input:

Anna's features

A

Output:

predicted blood sugar

–

Anna's features

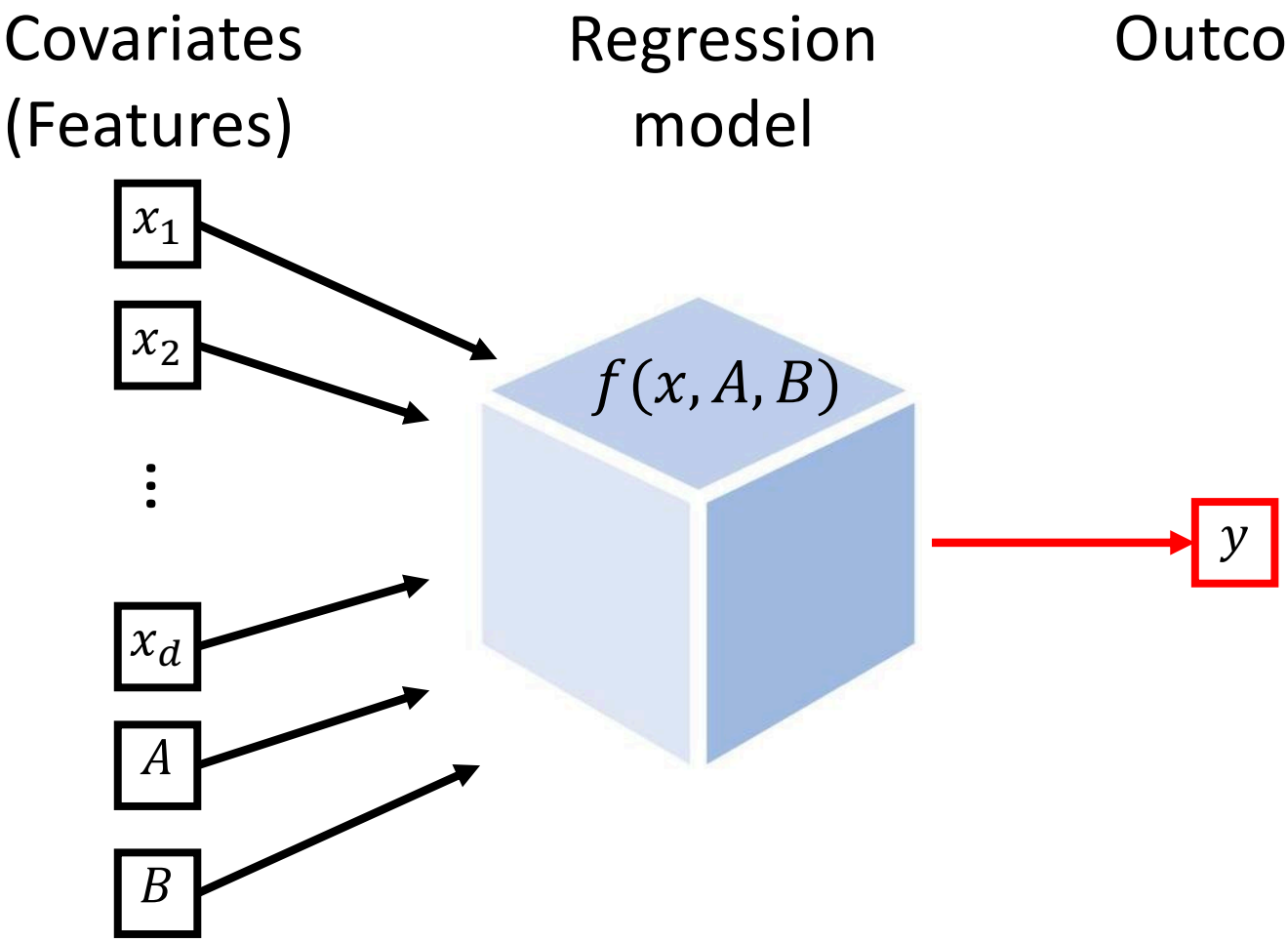
B

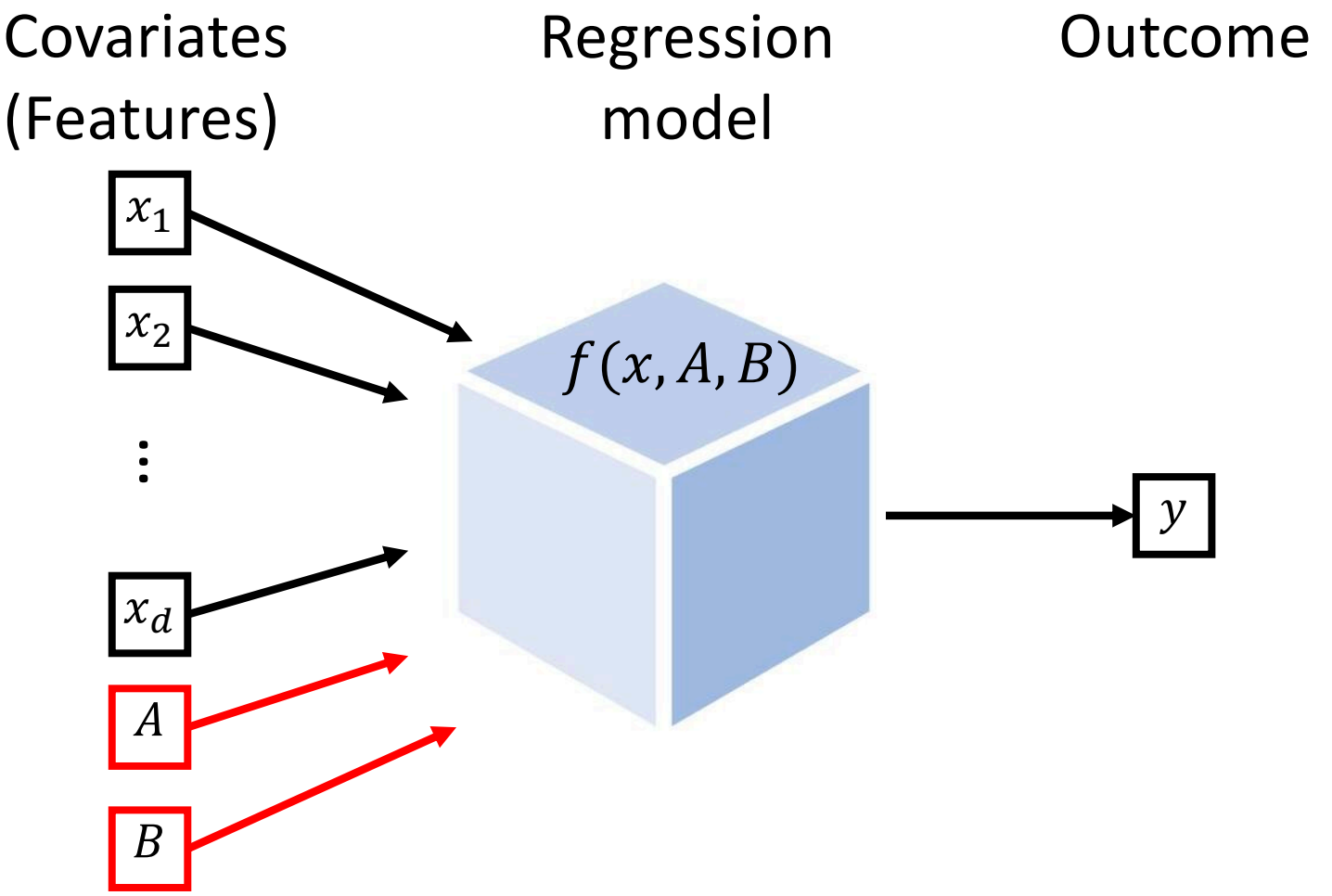
predicted blood sugar

=

?

- Compare


Covariates
(Features)
Regression
model
Outcome
x_1
x_2
x_d
A
B
$f(x, A, B)$
y


Covariates
(Features)
Regression
model
Outcome
x_1
x_2
x_d
A
B
$f(x, A, B)$
y

When supervised learning isn't enough

- This is not a classic supervised learning problem
- Our model was optimized to predict outcome, not to differentiate the influence of ***A*** vs. ***B***
- What if our high-dimensional model threw away the feature of medication ***A***/***B***?
- Hidden confounding:
  Maybe using ***B*** is *worse* than ***A***, but rich patients usually take ***B*** and richer people also have better health outcomes.
  If we don't know whether a patient is rich or not,
  we might conclude ***B*** is better

When supervised learning isn't enough

- Hidden confounding:
  Maybe using ***B*** is *worse* than ***A***, but…

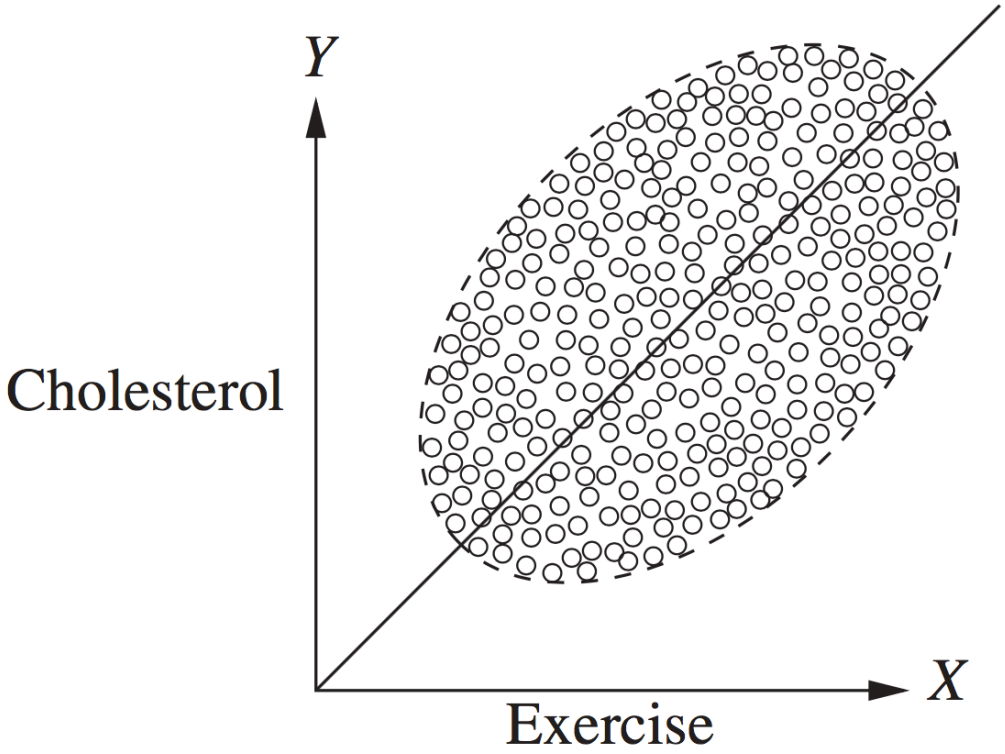
When supervised learning isn't enough

- Hidden confounding:
  Maybe using ***B*** is *worse* than ***A***, but…
  rich patients usually take ***B*** …

When supervised learning isn't enough

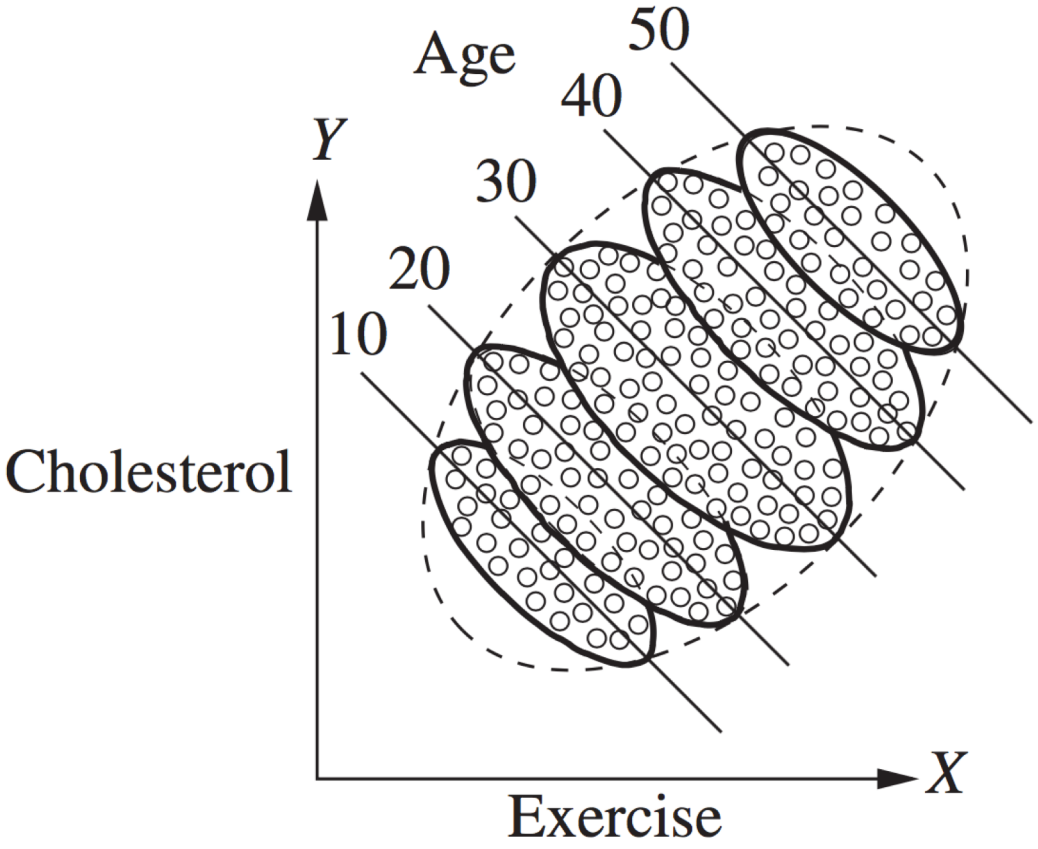
- Hidden confounding:
  Maybe using ***B*** is *worse* than ***A***, but…
  rich patients usually take ***B*** …
  therefore patients who took ***B*** exhibit better health outcomes.
- For prediction purposes this is the right thing to do:
  B is telling you that the patient is probably richer and will probably be healthier
- But for optimal action this is bad

Does exercise raise your cholesterol?



Pearl,
Glymor,
Jewell, 2016

Does exercise raise your cholesterol?



Pearl, Glymor, Jewell, 2016

How can we determine the existence and size of causal effects?

Randomized controlled trials

- Give each patients medication A or medication B **completely at random**
- Observe which group has better outcomes
- And we're done!
- There is a formal theory proving why this is right

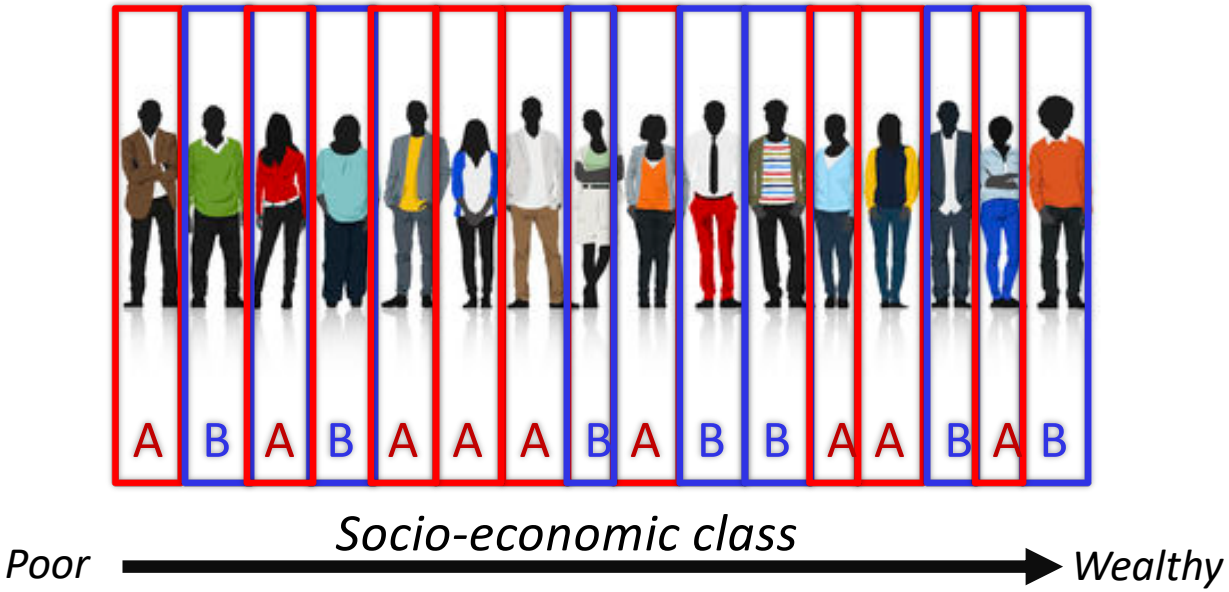
- Does inhaling Asbestos cause cancer?
- Does decreasing the interest rate reinvigorate the economy?
- We have a budget for **one new** anti-diabetic drug experiment. Can we use past health records of 100,000 diabetics to guide us?

Randomized trials vs. observational studies

treatment

A **or** B

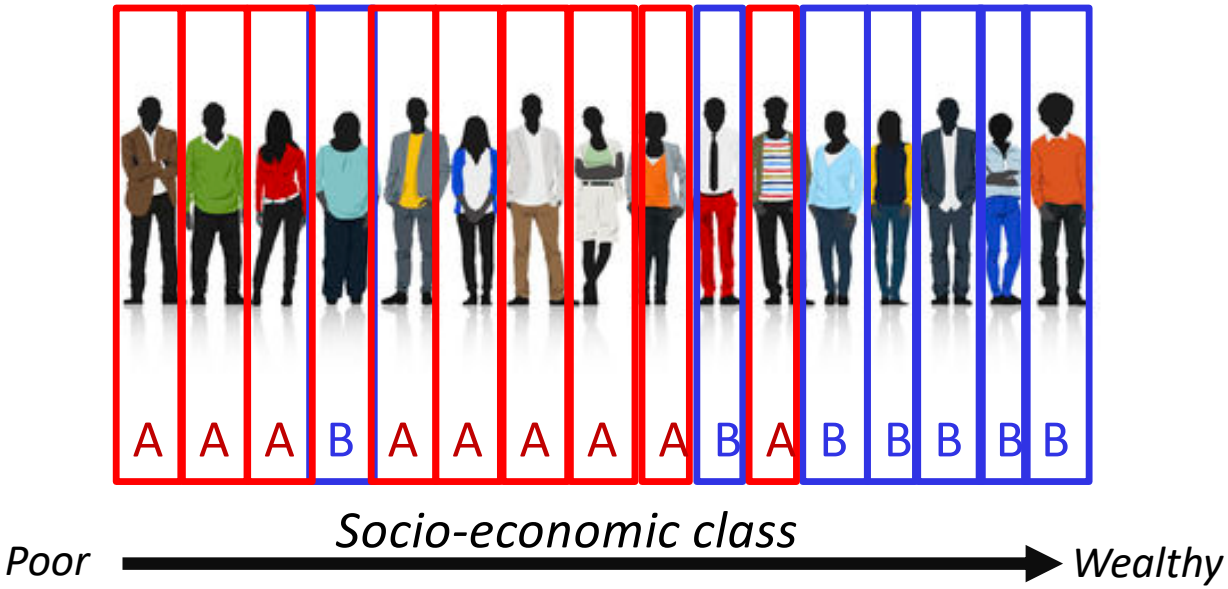
Randomized controlled trial (RCT)



treatment

A **or** B

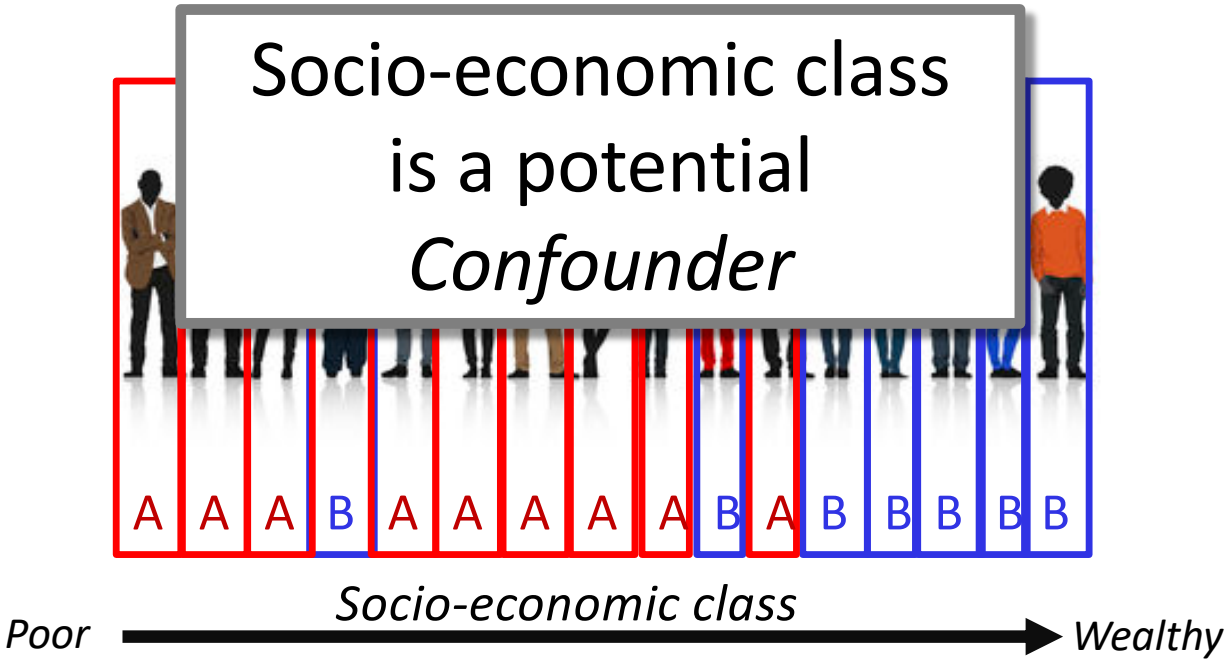
Observational study



treatment

A **or** B

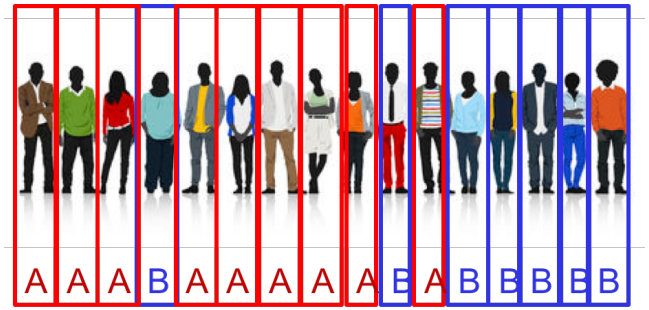
Observational study

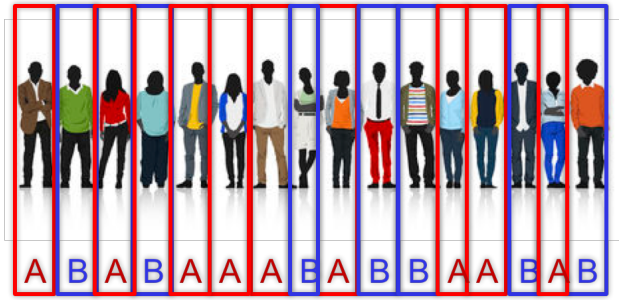


treatment

A **or** B

In many fields randomized studies are the gold standard for causal inference, but…

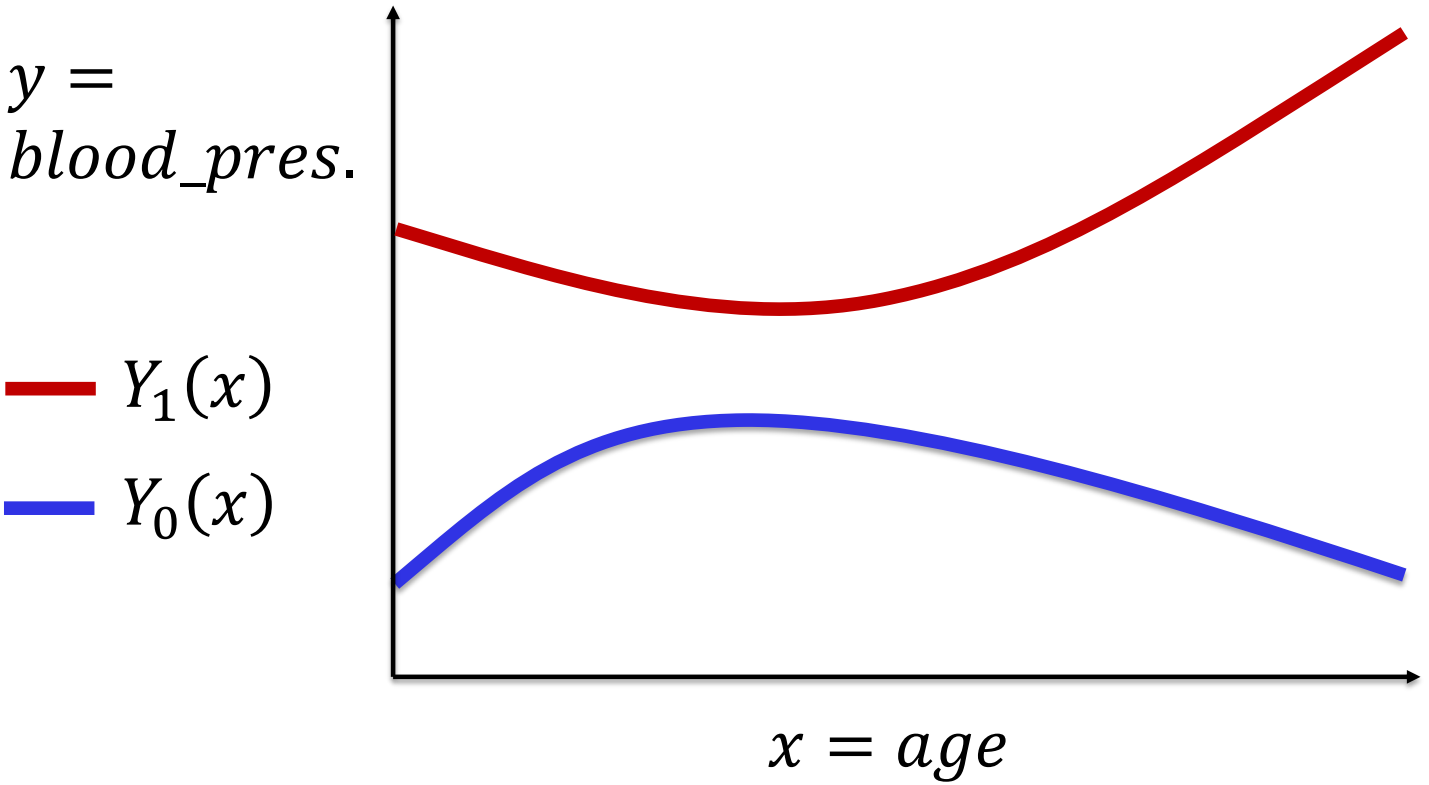




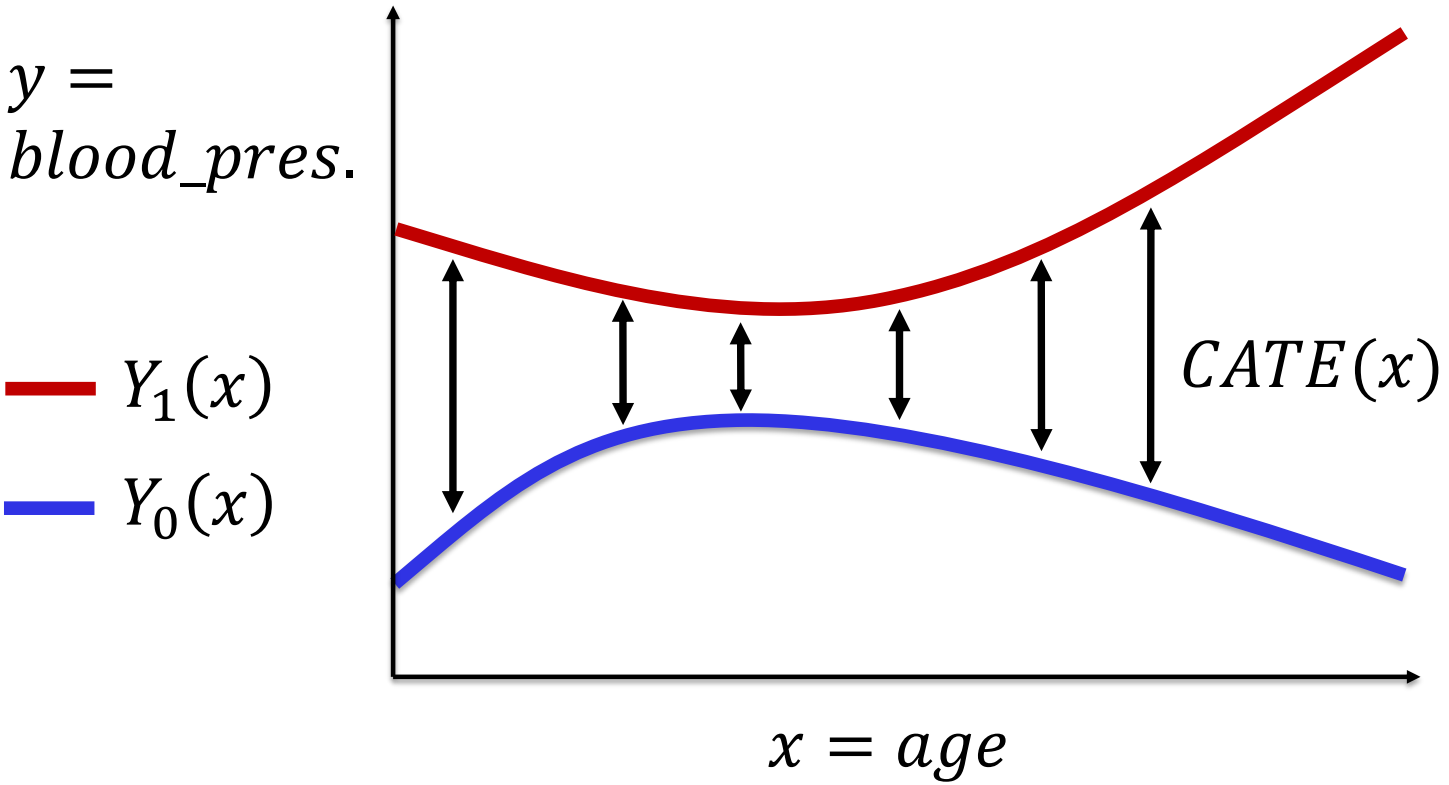
Even randomized controlled trials have flaws

- Study population might not represent true population
- Recruiting is hard
- People might drop out of study
- Study in one company/hospital/state/country could fail to generalize to others

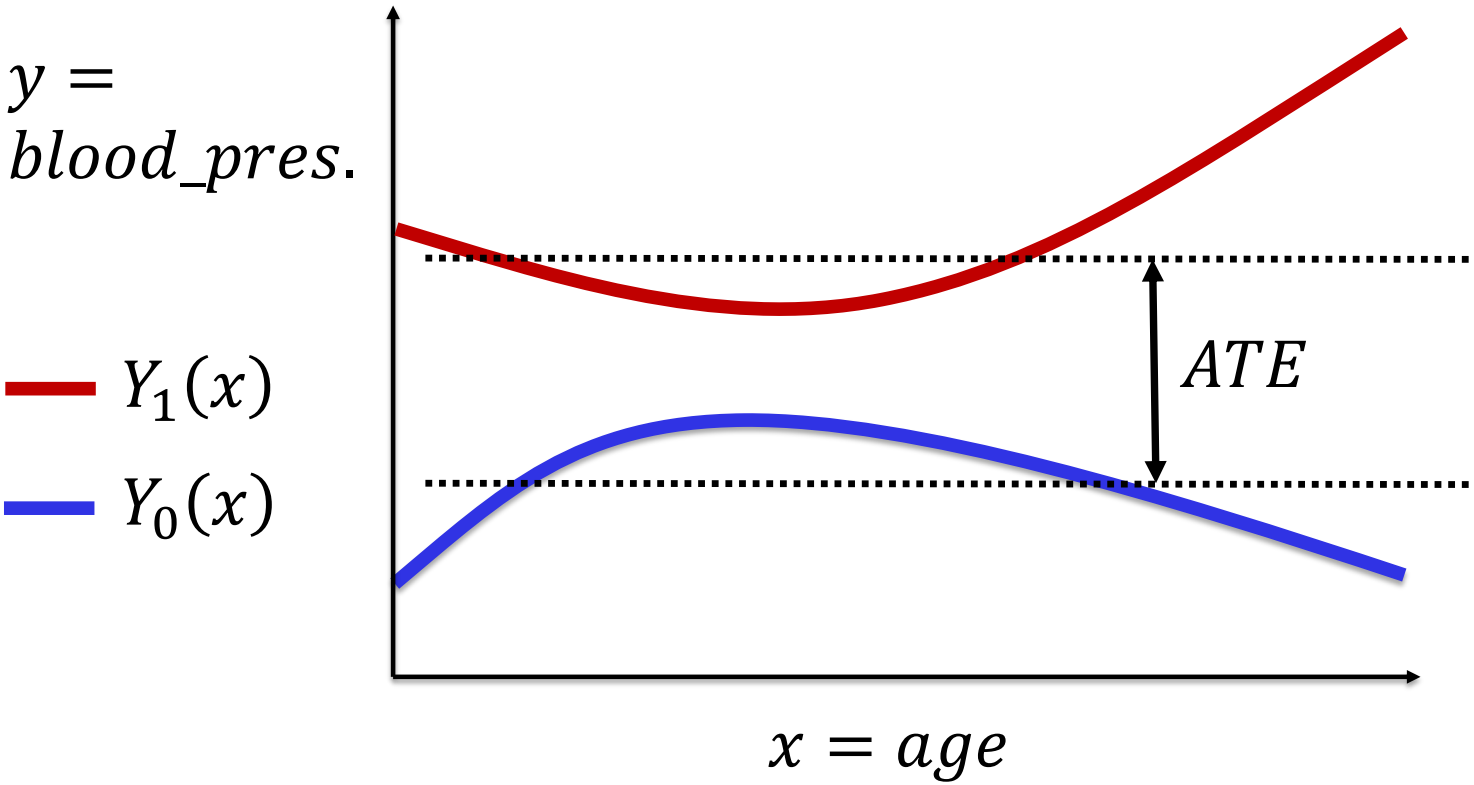
Covariate adjustment



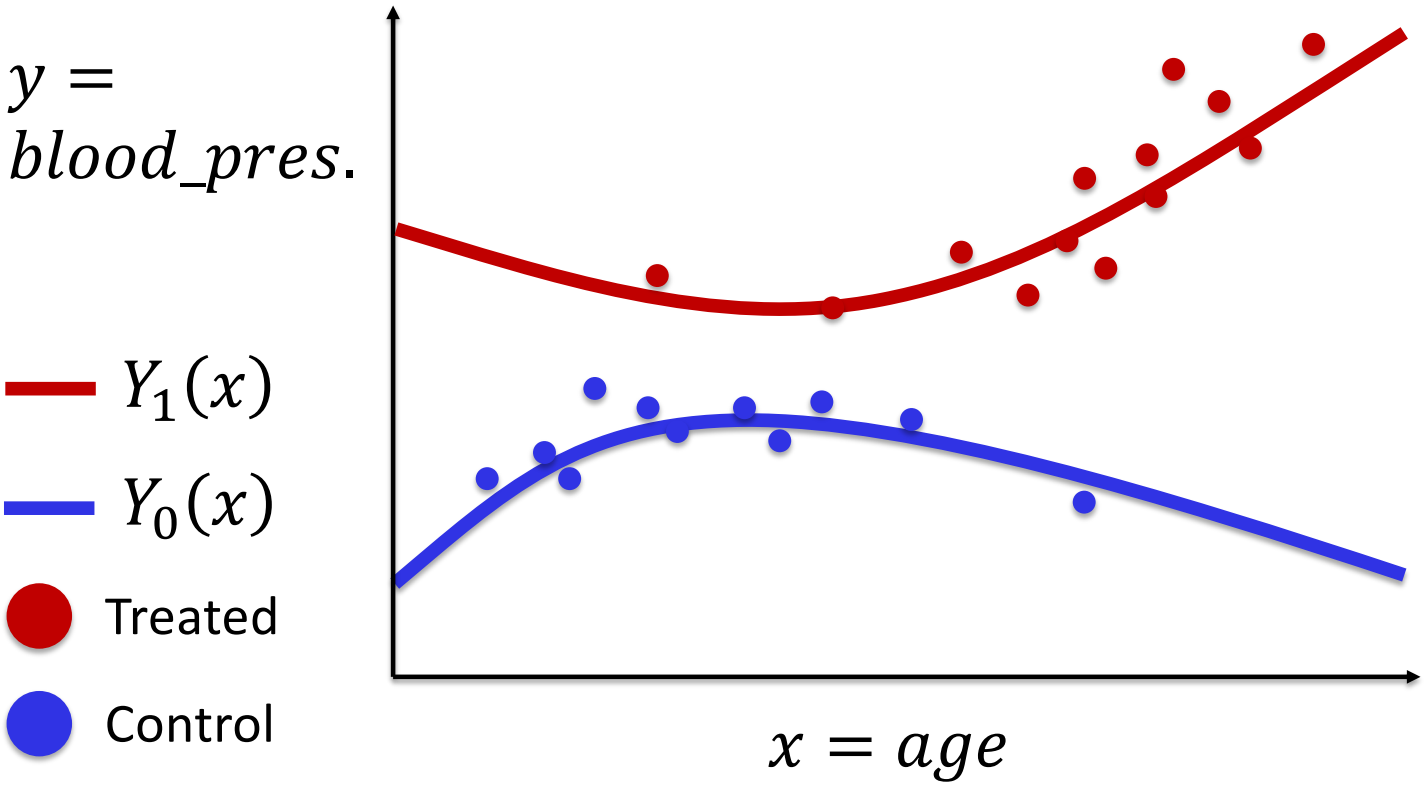
Covariate adjustment



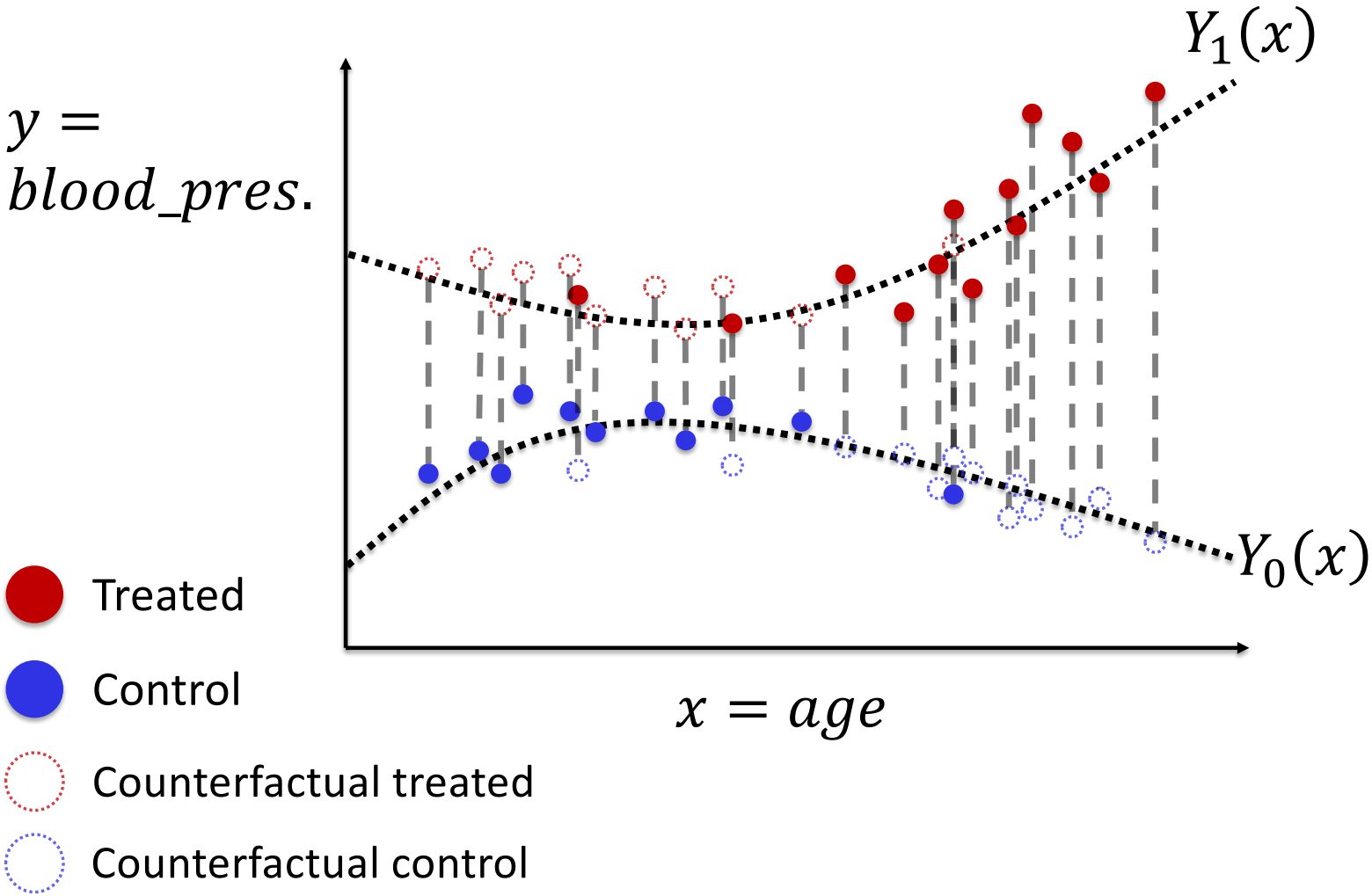
Covariate adjustment



Covariate adjustment



Covariate adjustment



Why is this not supervised learning?

- You can’t know how well you did on the counterfactuals
- NO TEST SET
- Adding certain types of variables will make your counterfactual predictions *worse* while making your observed predictions better

- Domain knowledge is provably necessary

“The Assumptions”

Sufficient conditions for causal inference to be possible:

1) *No unmeasured confounders*

2) *Common support*

Causal inference – what I want you to take away

- When using data of past actions to learn how to act, there is a risk of hidden confounding
- There is no “easy” solution
- Partial solutions:
  - Generate data with random actions
    e.g. clinical trials
  - Have a great mechanistic/physical model
    e.g. launching a spaceship to another planet
  - Use methods of causal inference, including domain knowledge
    e.g. observational study with covariate adjustment