(I²DB)

# CAUSAL MACHINE LEARNING FOR REAL-WORLD EVIDENCE GENERATION

Linying Zhang, PhD, MS

Assistant Professor of Biostatistics
Institute for Informatics, Data Science, and Biostatistics (I²DB)
Washington University School of Medicine in St Louis

CSE521S Wireless Sensor Networks
Nov 28, 2023

# My journey

To advance science and medicine with data



BA in Biochemistry & Molecular Biology (2014)



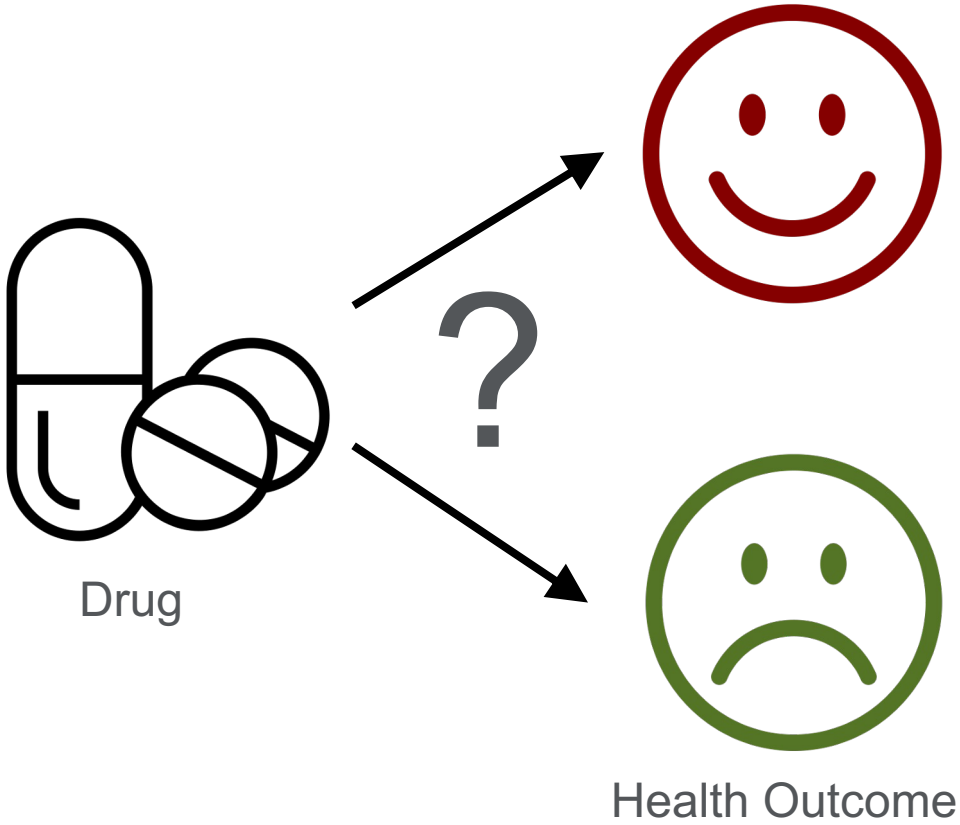
MS in Computational Biology (2018)



PhD in Biomedical Informatics (Clinical informatics track, 2023)

The goal is to integrate causal modeling with machine learning to generate reliable evidence from observational healthcare data.

# Is the treatment effective/safe?



(I²DB)

# Why large-scale analysis is needed in healthcare

All health outcomes of interest

All drugs



Figure credit: Patrick Ryan

Washington University School of Medicine in St. Louis

# Patient-level estimation requires big(ger) data

2 million patients seem excessive or unnecessary?

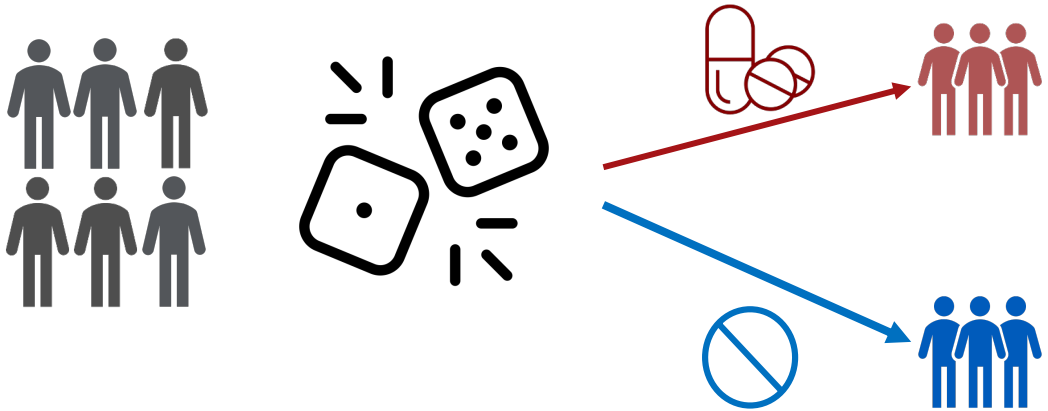
- Imagine a provider wants to compare her patient with other patients with the same gender (50%), in the same 10-year age group (10%), and with the same comorbidity of Type 2 diabetes (5%)
- Imagine the patient is concerned about the risk of ketoacidosis (0.5%) associated with two alternative treatments they are considering
- With 2 million patients, you'd only expect to observe 25 similar patients with the event, and would only be powered to observe a relative risk > 2.0

Aggregated data across a health system of 1,000 providers may contain 2,000,000 patients.

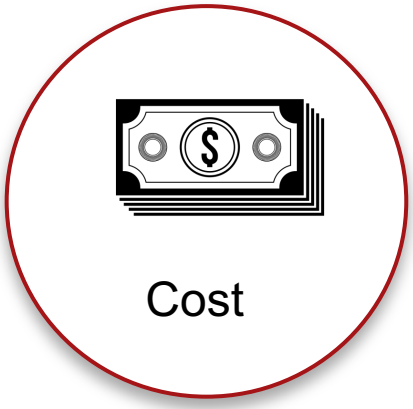
Slide credit: George Hripcsak

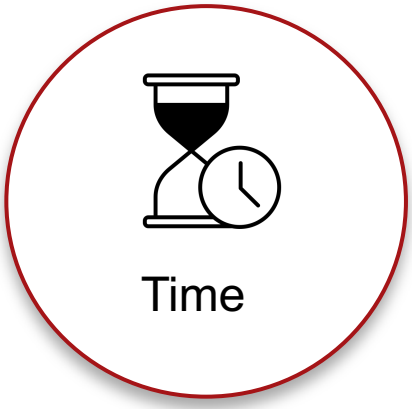
# The gold standard is randomized controlled trials

- Randomized controlled trials (RCTs) are prospective studies that measure the efficacy and safety of an intervention.
- RCT is considered as the gold-standard for establishing cause-and-effect relationships.
- Randomization balances participants characteristics and reduces confounding bias.



# RCTs can have their drawbacks










RCT is not scalable for large-scale evidence generation.

# Alternatively, there is observational data

- Subjects **observed** in their natural settings
  - Often using data collected for other purposes
  - **Real-world evidence (RWE)**
- Versus experimental
  - Randomized clinical trials (RCTs)
- Electronic health record (EHR) data
  - BJC HealthCare EHR data warehouse (2M)
- Other sources
  - Administrative claims data (e.g., IBM Marketscan (10M's))
  - Mobile sensors, Census, social media, imaging…

# Real-world data in healthcare



Electronic Health Records (EHRs)

## Structured Data



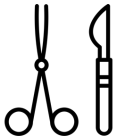
Demographics



Diagnoses



Drugs



Procedures



Labs/Vital signs

## Unstructured Data



Medical Images



Clinical Notes



Genomics



Wearables

# US National EHR data, per year

| | |
|---|---|
| 1,000,000,000 | visit notes |
| 35,000,000 | admit notes, discharge sum. |
| 46,000,000 | procedure notes |
| 3,000,000,000 | prescriptions |
| 1,000,000,000 | laboratory tests |
| >50,000,000,000 | facts |

Slide credit: George Hripcsak

# Can we trust observational studies?



EDITORIAL

Expression of Concern: Mehra MR et al. Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621.

Eric J. Rubin, M.D., Ph.D.

Article Metrics

3 References 12 Citing Articles

June 18, 2020
N Engl J Med
DOI: 10.1056/N

Related Articles

ORIGINAL
Cardiovascular

On May 1, 2020, we published "Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19,"[1] a study of the effect of preexisting treatment with angiotensin-converting enzyme (ACE) inhibitors and angiotensin-receptor blockers (ARBs) on Covid-19. This retrospective study used data drawn from an international database that

THE LANCET

Volume 395, Issue 10240, 13–19 June 2020, Page e102

Comment

Expression of concern: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis

The Lancet Editors

Show more

Share Cite

https://doi.org/10.1016/S0140-6736(20)31290-3 Get rights and content

Two retracted COVID-19 papers

# How reproducible is RWE at large-scale?

nature communications

July 2021

Article

RWE VS RWE

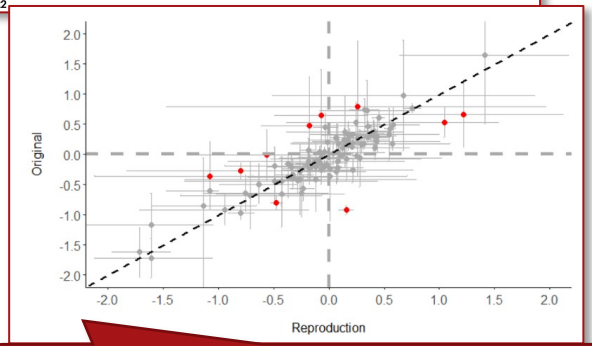
https://doi.org/10.1038/s41467-022-32310-3

**Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions**

Received: 30 August 2021

Accepted: 26 July 2022

Shirley V. Wang[1,2], Sushama Kattinakere Sreedhara[1], Sebastian Schneeweiss[1,2] & REPEAT Initiative*




"to reproduce 150 **observational** studies …effect sizes were positively correlated (Pearson's correlation = 0.85)"

# How about RWE vs RCT?

nature communications

Article

RWE VS RWE

July 2021

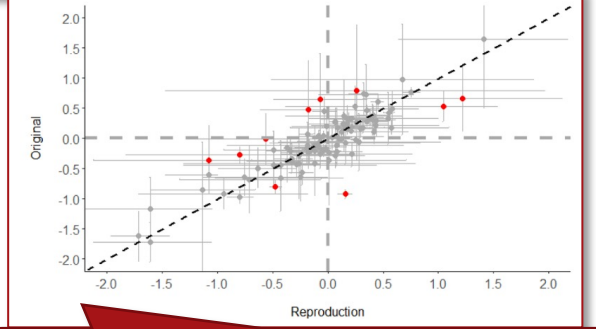
https://doi.org/10.1038/s41467-022-32310-3

## Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions

Received: 30 August 2021

Accepted: 26 July 2022

Shirley V. Wang[1,2], Sushama Kattinakere Sreedhara[1], Sebastian Schneeweiss[1,2] & REPEAT Initiative*




"to reproduce 150 **observational** studies …effect sizes were positively correlated (Pearson's correlation = 0.85)"

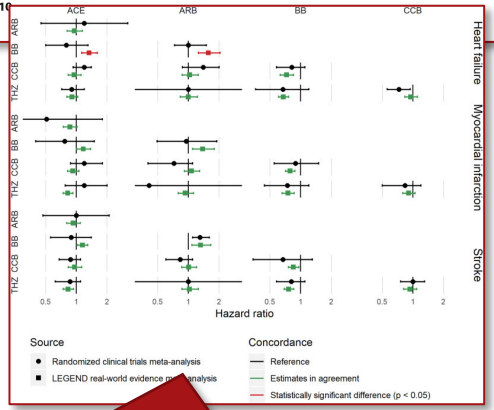
Research and Applications

RWE VS RCT

June 2020

## Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study

Martijn J Schuemie[1,2] Patrick B Ryan,[1,3] Nicole Pratt,[4] RuiJun Chen[3,5] Seng Chan You,[6] Harlan M Krumholz,[7] David Madigan,[8] George Hripcsak,[3,9] and Marc A Suchard[2,10]




28/30 effect sizes were in agreement between RWE and RCT.

Selected publications

# My contributions

## Prediction

Post-bariatric surgery outcomes [Richter, Albert, **Zhang** et al, *Front. Physiol*. 2022]

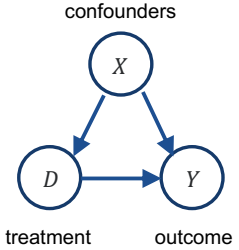
Hospitalization of COVID-19+ [Song, **Zhang** et al, *JAMIA* 2021]

Pressure injury [Song, Kang, **Zhang**, et al, *JAMIA* 2020]

## Estimation (Drug Effect)

The medical deconfounder [**Zhang** et al, *MLHC* 2019]

Large-scale propensity score [**Zhang** et al, *JBI* 2022]




## Estimation (Health Equity)

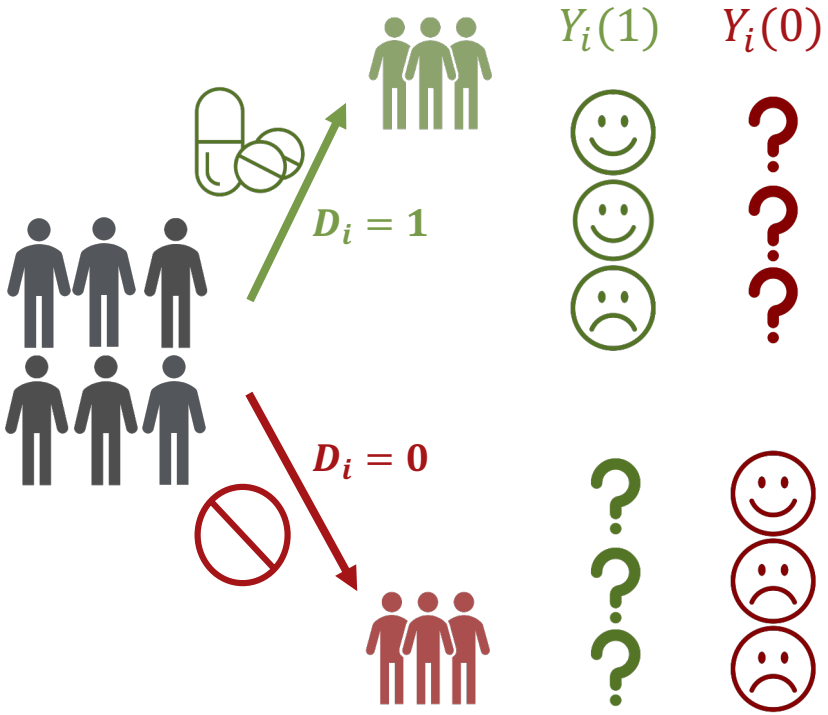
Algorithmic bias: eGFR in nephrology [**Zhang** et al, *medRxiv* 2021]

Treatment allocation bias [**Zhang** et al, *arXiv* 2022, **Zhang** et al, AMIA Symp 2023]



# Overview of Causal Inference

# Problem formulation



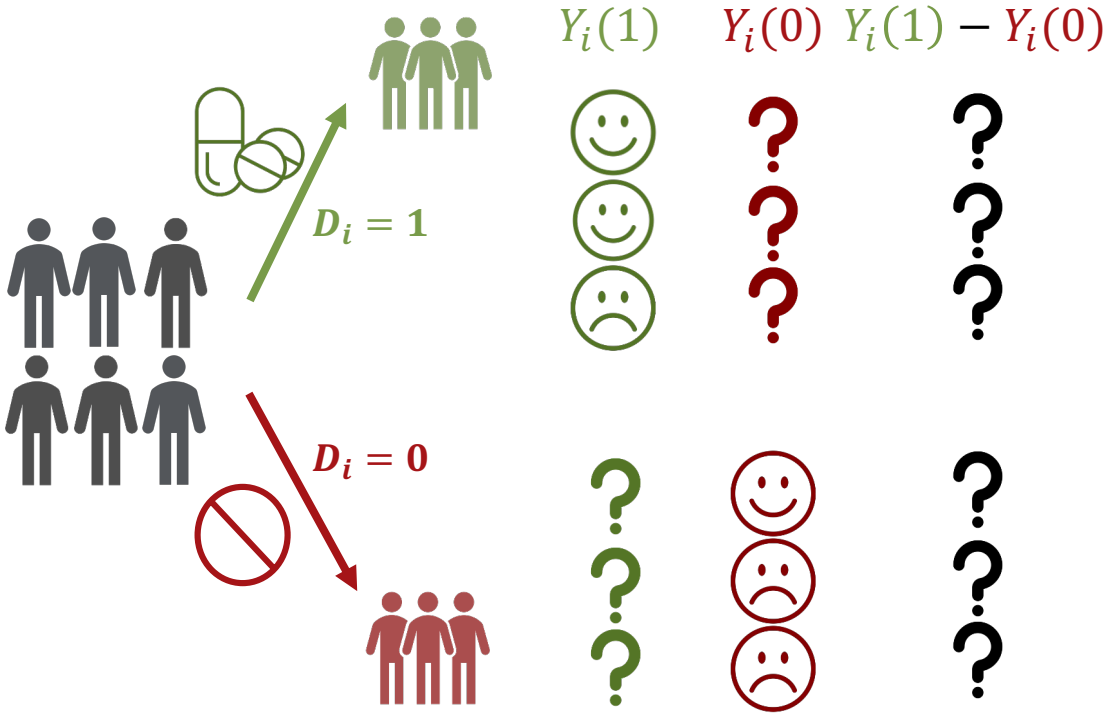



How effective is the treatment on average?

The average treatment effect (ATE) is $\mathbb{E}[Y_i(1) - Y_i(0)]$.



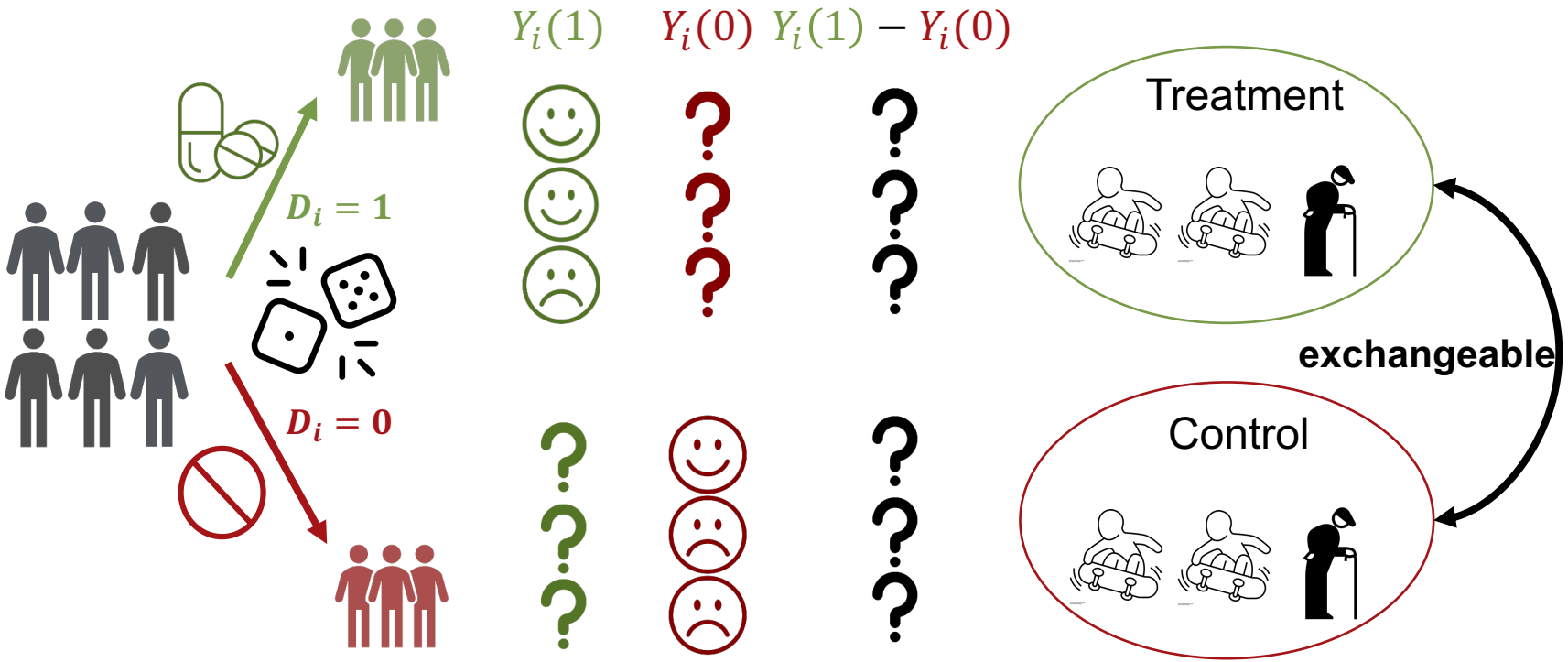
# The fundamental problem of causal inference



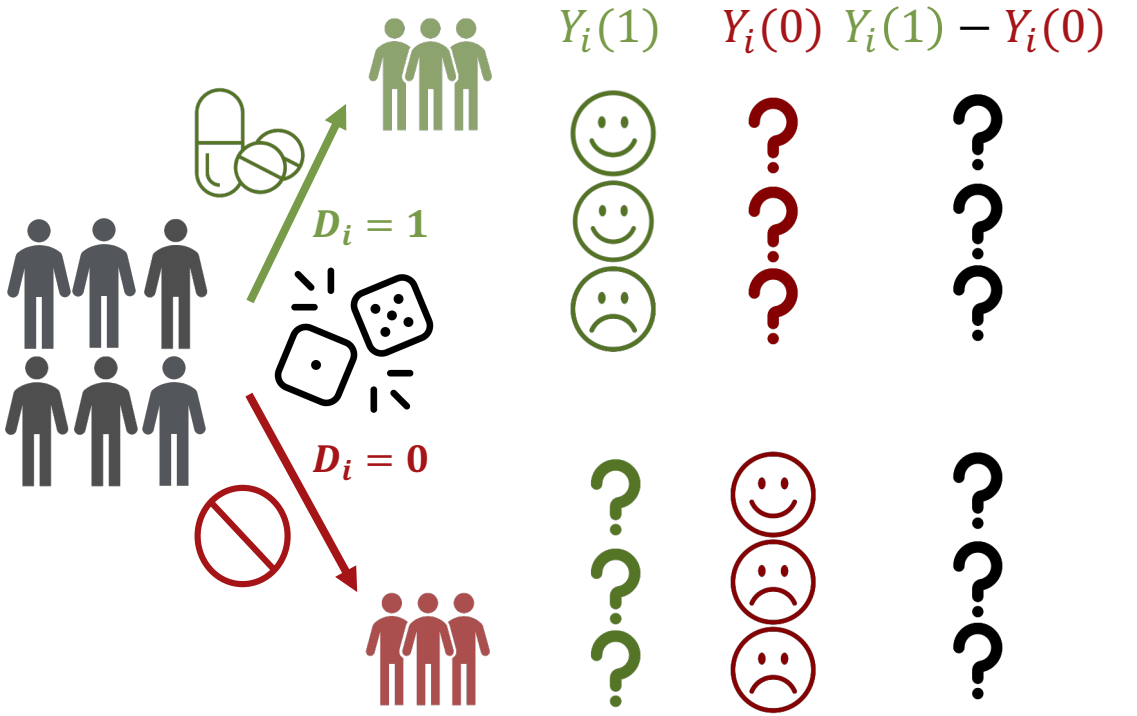

$Y_i(1) - Y_i(0)$
is **never** observed!



# If the treatment is randomized …



# If the treatment is randomized …



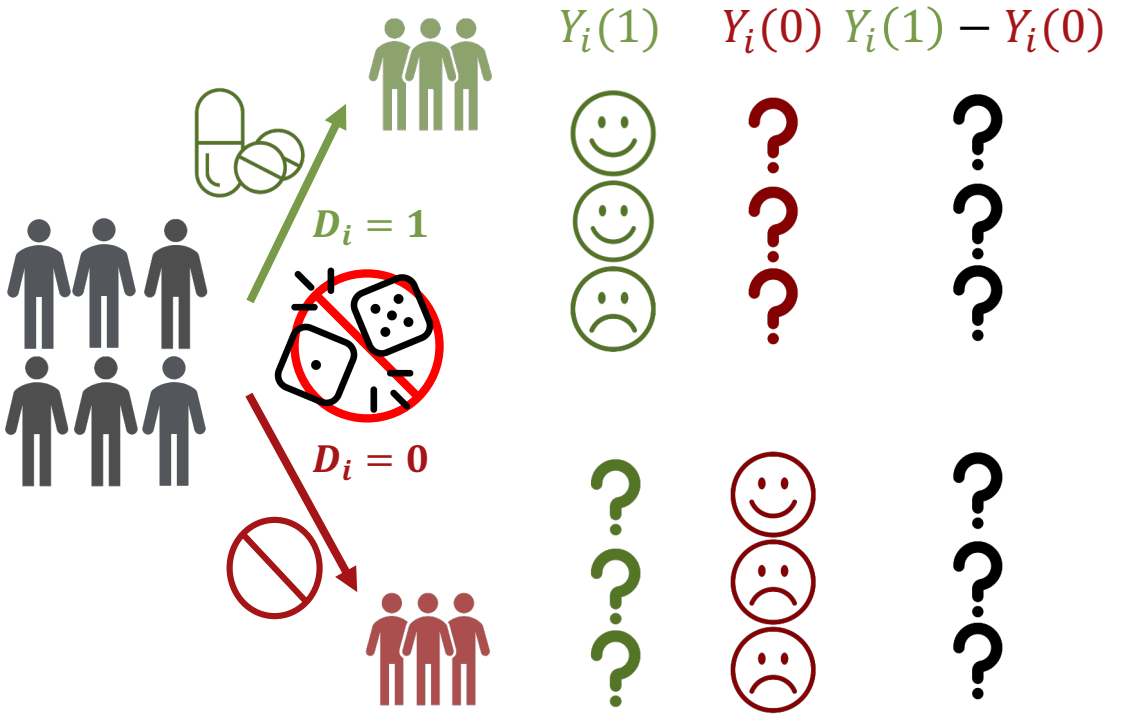

$$\mathbb{E}[Y(1) - Y(0)]$$

$$= \checkmark$$

$$\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$$



# If the treatment is NOT randomized …




$$\mathbb{E}[Y(1) - Y(0)]$$

$$\neq$$

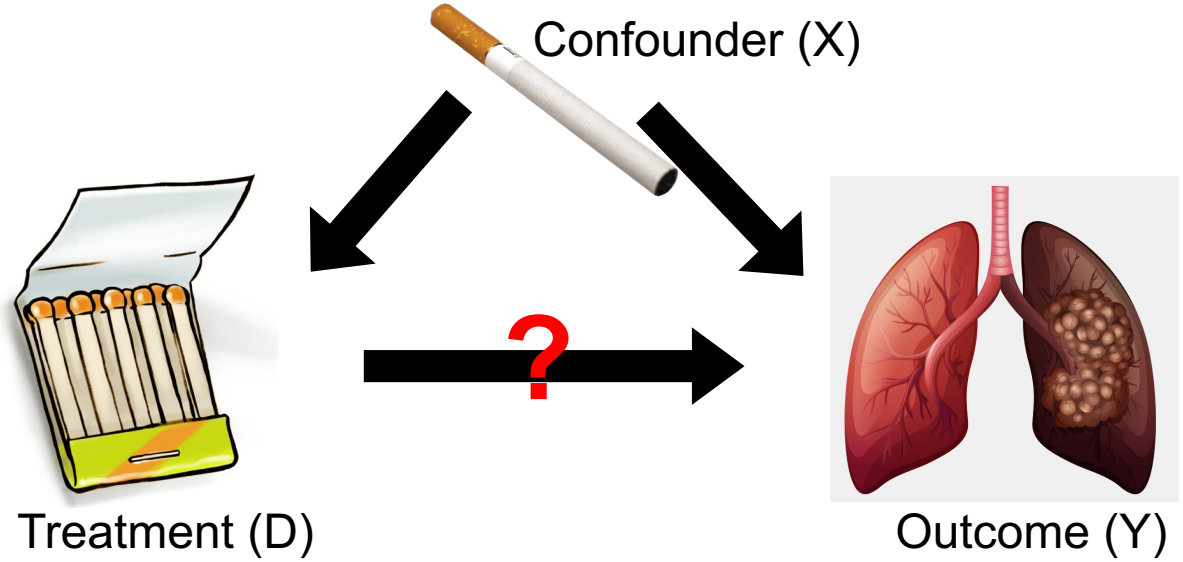
STOP

$$\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$$



# Confounding bias

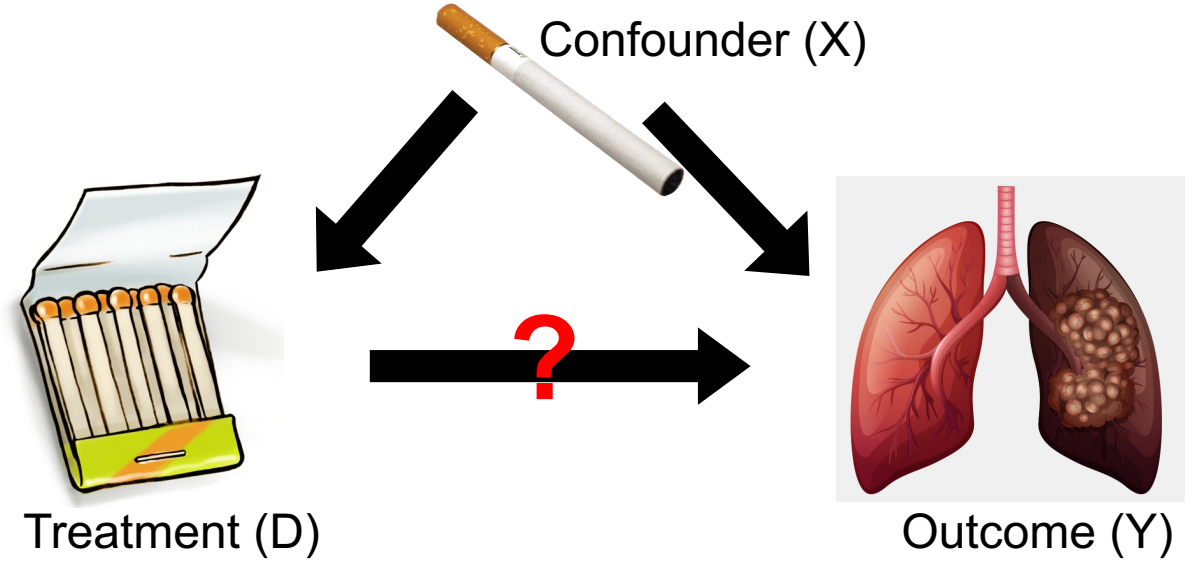
Confounding remains one of the major challenges to causal inference with observational data.



# Adjusting for confounders in causal effect estimation

$$\mathbb{E}[Y(1) - Y(0)] \neq \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]$$

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\mathbb{E}[Y \mid X = x, D = 1]] - \mathbb{E}_X[\mathbb{E}[Y \mid X = x, D = 0]]$$



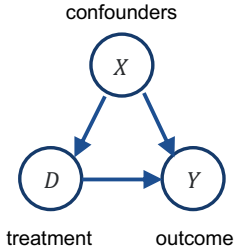
# Part I: Reliability of treatment effect estimation

# Part I: Reliability of treatment effect estimation

## Estimation (Drug Effect)

**The medical deconfounder [Zhang et al, *MLHC* 2019]**

Large-scale propensity score [Zhang et al, JBI 2022]




## Estimation (Health Equity)

Algorithmic bias: eGFR in nephrology [**Zhang** et al, *medRxiv* 2021]

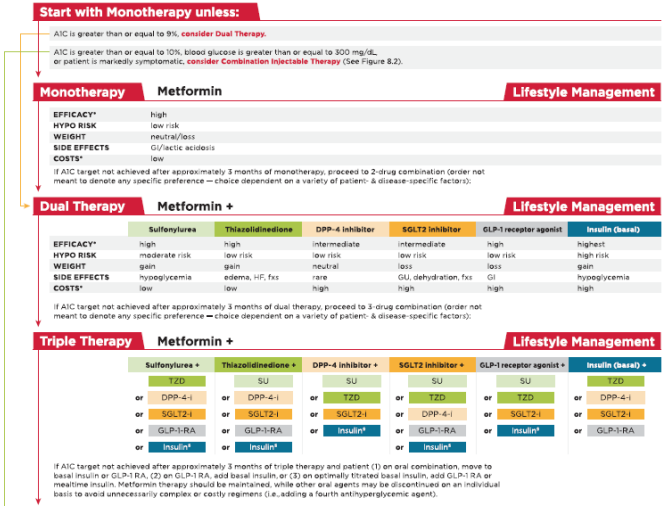
Treatment allocation bias [**Zhang** et al, *arXiv* 2022]



# Concurrent usage of multiple medications

Concurrent usage of medications is common in certain patient populations.

- Patients with **multimorbidities** [Aggarwal et al. 2020]
  - Older adults [Sharma et al. 2016]
- Patients with **chronic pain** [Medicare Payment Advisory Committee. 2014]
  - Co-prescription of opioids, NSAID (anti-inflammatory), and other CNS depressants.
- Patients with **advanced cancer** [LeBlanc et al. 2015]
- Patients with **type 2 diabetes** [Johnson et al. 2017]
  - Combination therapy is advocated by clinical guidelines if monotherapy fails to meet glycemic target.




Antihyperglycemic therapy in type 2 diabetes: general recommendations. Figure reprinted from Johnson *et al,* 2017.

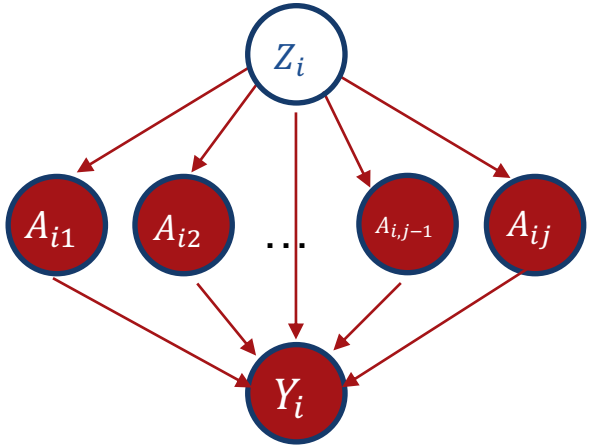
# Multiple causal inference

| | Drug 1 | Drug 2 | Drug 3 | Drug 4 | ... | Blood sugar level |
|---|---|---|---|---|---|---|
|  | 1 | 1 | 0 | 0 | … | 120 |
|  | 0 | 1 | 1 | 0 | … | 130 |
|  | 0 | 1 | 1 | 1 | … | 200 |
| … | … | … | … | … | … | … | |

Multiple causal inference seeks to estimate the causal effects of multiple causes.

- Data: $\{(\boldsymbol{a}_1, y_1), \ldots, (\boldsymbol{a}_n, y_n)\}$, where $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{im})$
- Goal: Estimate the average treatment effect of $\boldsymbol{a}$: ATE = $\mathbb{E}[Y(\boldsymbol{a}) - Y(\boldsymbol{a}')]$

# The deconfounder (Wang and Blei 2019)




For the *i*-th patient,

$A_i$: causes

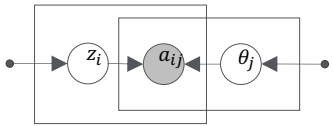
$Y_i$ : outcome

$Z_i$ : unmeasured confounder

- **Unmeasured confounders** affect both the causes and the outcome.
- Unmeasured confounders lead to **biased** effect estimates if not adjusted.
- The **multiplicity** of causes helps us construct a variable that contains unmeasured confounders.
- Bayesian probabilistic models produces **representations** of treatments that capture **unmeasured confounder.**

# The deconfounder (Wang and Blei 2019)

MODEL ASSIGNED CAUSES




ESTIMATE SUBSTITUTE CONFOUNDERS

$$\{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n\}$$

$$\hat{z}_i = \mathbb{E}[Z_i | A_i = a_i]$$

ESTIMATE CAUSAL EFFECTS

$$\mathbf{E}[Y(a)] = \mathbf{E}\left[\mathbf{E}\left[Y \middle| A = a, \hat{z}\right]\right]$$

- Find, fit, and check a **factor model** of the causes.
- Use the factor model to form **substitute confounders**.
- Use the substitute confounders in downstream **causal inference**.

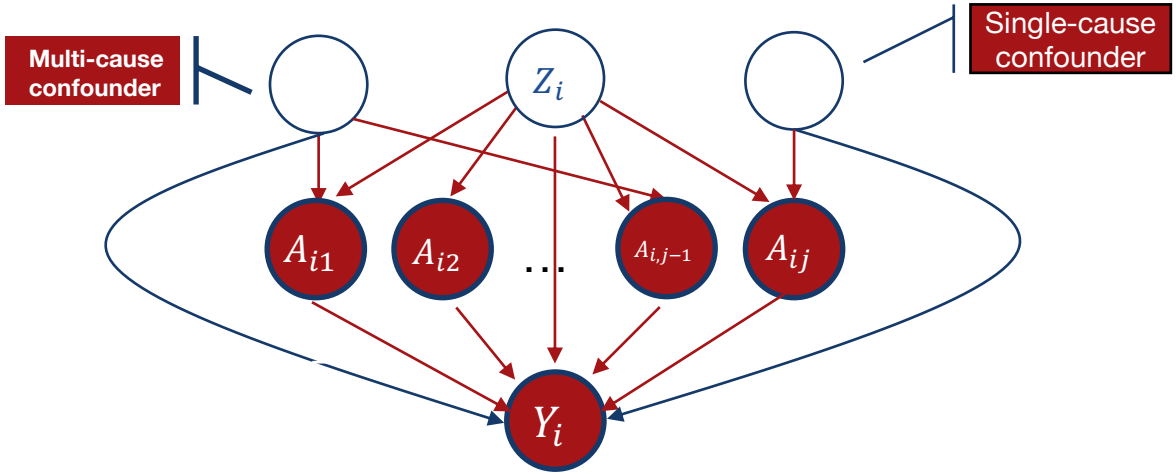
# Intuition:
# Leveraging co-prescription pattern to infer unmeasured confounders

| | Severity | Drug 1 | Drug 2 | Drug 3 | Drug 4 | … | Blood sugar level |
|---|---|---|---|---|---|---|---|
|  | Low | 1 | 1 | 0 | 0 | … | 120 |
|  | Median | 0 | 1 | 1 | 0 | … | 130 |
|  | High | 0 | 1 | 1 | 1 | … | 200 |
| … | … | … | … | … | … | … | … |

Confounder $Z_{n\times 1}$ Treatments $A_{n\times m}$ Outcome $Y_{n\times 1}$

- For example, we may be able to infer how sick a patient is by looking at what medications the patient takes.

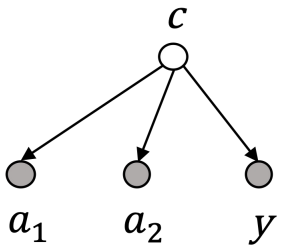
# The deconfounder: assumptions



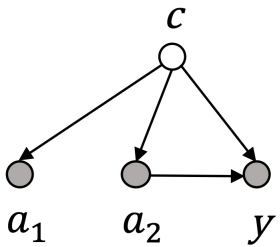

- The deconfounder assumes no unmeasured **single-cause confounders**.
- Assumes the medications do not cause each other.

# Simulation setup

- The goal is to check whether the deconfounder is able to identify the true cause and its effect size.
- Simulate two scenarios: 1) no real cause; 2) one real cause
- Three models are compared:
  - Unadjusted
  - Oracle (adjusting for the true confounder)
  - Medical deconfounder (Probabilistic PCA as the factor model)



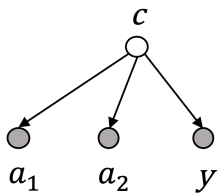

**Scenario 1**
**no real cause**




**Scenario 2**
**one real cause**

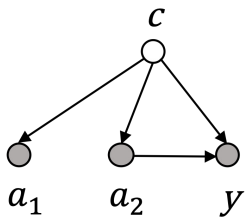
# Simulation results

**Scenario 1**
**(no real cause)**




| | medication 1 | | medication 2 | |
|---|---|---|---|---|
| | **coef (std err)** | ***p*-value** | **coef (std err)** | ***p*-value** |
| **Truth** | 0.000 | 1 | 0.000 | 1 |
| **Oracle** | 0.025 (0.039) | 0.522 | -0.022 (0.040) | 0.594 |
| **Unadjusted** | 0.125 (0.042) | 0.003 | 0.141 (0.041) | 0.001 |
| **Med. Dcf.** | 0.020 (0.081) | **0.803** | 0.052 (0.071) | **0.461** |

**Scenario 2**
**(one real cause)**

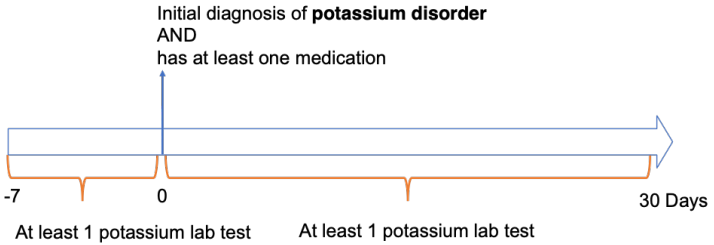



| | medication 1 | | medication 2 | |
|---|---|---|---|---|
| | **coef (std err)** | ***p*-value** | **coef (std err)** | ***p*-value** |
| **Truth** | 0.000 | 1 | 0.300 | 0 |
| **Oracle** | 0.058 (0.038) | 0.132 | 0.329 (0.039) | 0.000 |
| **Unadjusted** | 0.181 (0.040) | 0.000 | 0.469 (0.040) | 0.000 |
| **Med. Dcf.** | 0.069 (0.063) | **0.272** | 0.333 (0.072) | **0.000** |

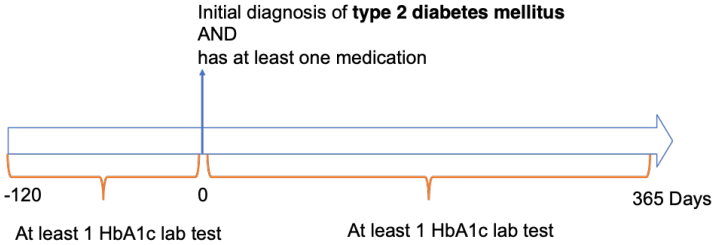
The medical deconfounder is able to identify the true cause and estimate the effect unbiasedly.

# Empirical studies with EHRs

- **Potassium disorder cohort**:
  - Index date: earliest diagnosis of hyperkalemia or hypokalemia
  - Features: all medications (at ingredient level) on index date
  - Outcome: change in serum potassium before and after treatment.

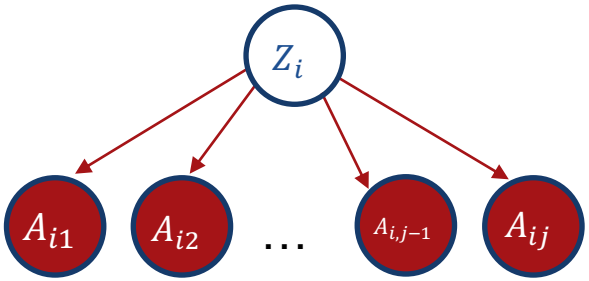


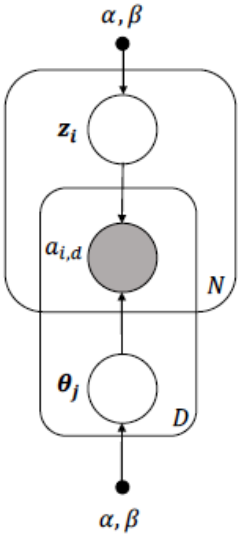

- **Type 2 diabetes mellitus cohort**:
  - Index date: earliest diagnosis of type 2 diabetes
  - Features: all medications (at ingredient level) on index date
  - Outcome: change in hemoglobin A1c before and after treatment.



# Poisson factorization as the factor model

Example: Poisson factorization [Gopalan et al. 2015]



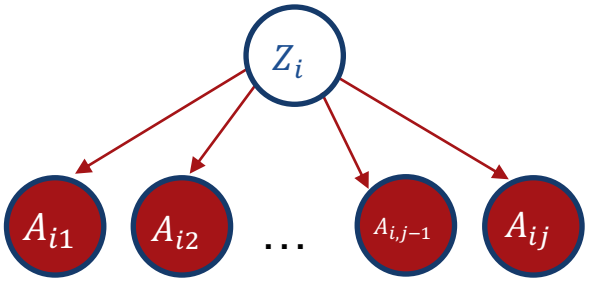



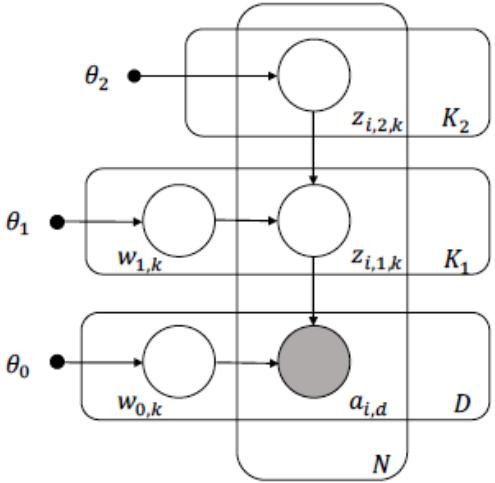

$$Z_i \sim \text{Gamma}(\alpha, \beta),$$
$$\theta_j \sim \text{Gamma}(\alpha, \beta),$$
$$A_{ij} \mid Z_i \sim \text{Poisson}(Z_i^T \theta_j).$$

- A Bayesian form of non-negative matrix factorization [Lee & Seung 1999]
- Provides a generative model of the assigned causes $A_{ij}$.

# Deep exponential family as the factor model

Example: Deep exponential family [Ranganath et al. 2015]




$$
\begin{aligned}
W_{l,k} &\sim \text{Gamma}(\alpha, \beta), \\
Z_{i,2,k} &\sim \text{Gamma}(\alpha, \beta), \\
Z_{i,1,k} \,|\, Z_{i,2,k} &\sim \text{Gamma}(\alpha, g(W_{1,k} Z_{i,2,k})), \\
A_{i,d} \,|\, Z_{i,1} &\sim \text{Poisson}(g(W_0 Z_{i,1})).
\end{aligned}
$$

- Inspired by the hidden structures used in deep neural networks.
- Provides a generative model of the assigned causes $A_{ij}$.

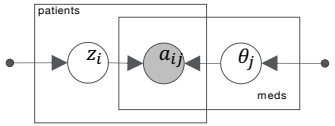
# Empirical study workflow




Extract EHRs from the CUMC database

## Factor model

Fit and check a factor model




Estimate the substitute confounder

$$\hat{Z}_i \triangleq \mathbb{E}_Z \left[ Z_i \mid \mathbf{A}_i, \hat{\theta} \right]$$

## Outcome model

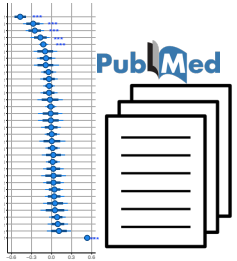
Fit a regression model

$$Y_i \sim \mathcal{N}\left(\sum_{j=1}^{D} \beta_j A_{ij} + \sum_{k=1}^{K} \gamma_k \hat{Z}_{ik}, \sigma^2\right)$$

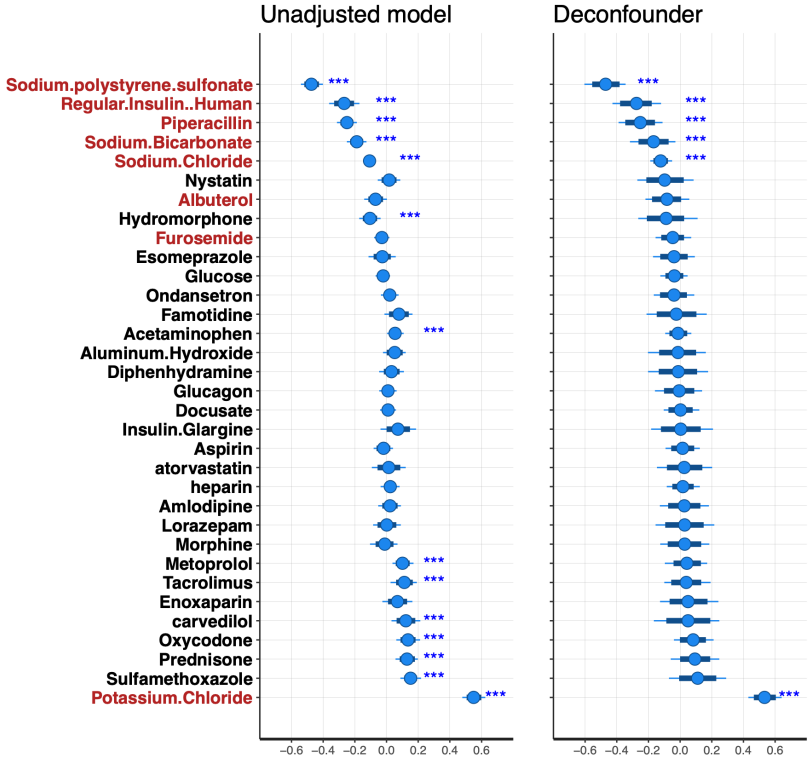
Estimate causal effects

$$\mathbb{E}\left[Y_i(\mathbf{a})\right] = \mathbb{E}_Z\left[\mathbb{E}_Y\left[Y_i \mid Z_i, \mathbf{A}_i = \mathbf{a}\right]\right]$$

Compare the results to findings in the medical literature



# Treatment effects on serum potassium level




Apply the medical deconfounder to the potassium disorder cohort.

- The unadjusted model

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D} \beta_j A_{ij}, \sigma^2)$$

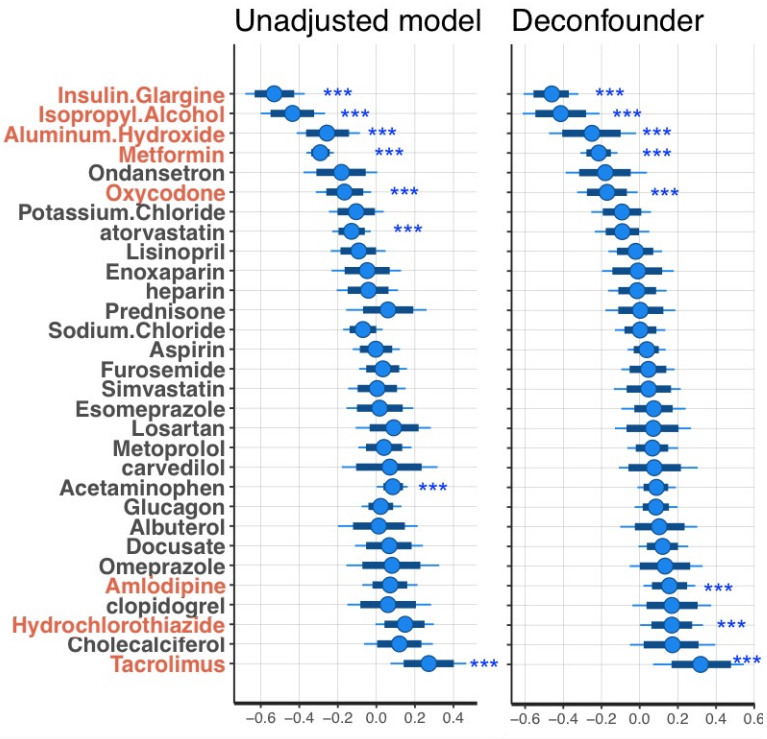
- The medical deconfounder

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D} \beta_j A_{ij} + \sum_{k=1}^{K} \gamma_k \hat{Z}_{ik}, \sigma^2)$$

**the substitute confounder**

Red:
Causal drugs based on literature review.

# Treatment effects on hemoglobin A1c




Apply the medical deconfounder to the type 2 diabetes cohort.

- The unadjusted model

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D} \beta_j A_{ij}, \sigma^2)$$

- The medical deconfounder

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D} \beta_j A_{ij} + \sum_{k=1}^{K} \gamma_k \hat{Z}_{ik}, \sigma^2)$$

**the substitute confounder**

Red:
Causal drugs based on literature review.

# Contributions

- Demonstrate a novel use of probabilistic models for inferring unmeasured confounders from medication co-prescription pattern.
- Deep exponential family with sparse priors for latent variables appeared to better capture correlation among medications.
- Extend treatment effect estimation from binary treatment to multiple treatments.

# Limitations

- Assumes that the medications don't cause each other.
  - Factor model does not capture the distribution of causally-dependent causes.
- Assumes no single-cause confounder.
  - The conditional independence of factor models only guarantees $Z_i$ is multi-cause.
  - If $Z_i$ is just a covariate that does not affect the outcome, conditioning on it increases the variance of the causal estimate.
- Need an infinite number of medications to estimate $Z_i$ with full certainty (pinpoint $Z_i$).
- The quality of multi-medication phenotype may be poor.
  - The duration of on-treatment is not well-captured in EHR.
  - Health care process is fragmented.

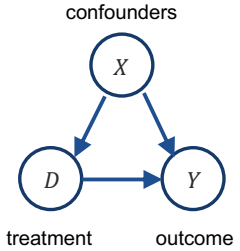
Zhang L, Wang Y, Ostropolets A, Mulgrave JJ, Blei DM, Hripcsak G. The medical deconfounder: assessing treatment effects with electronic health records. *MLHC*. 2019.

# Part I: Reliability of treatment effect estimation

## Estimation (Drug Effect)

The medical deconfounder [Zhang et al, MLHC 2019]

**Large-scale propensity score [Zhang et al, JBI 2022]**



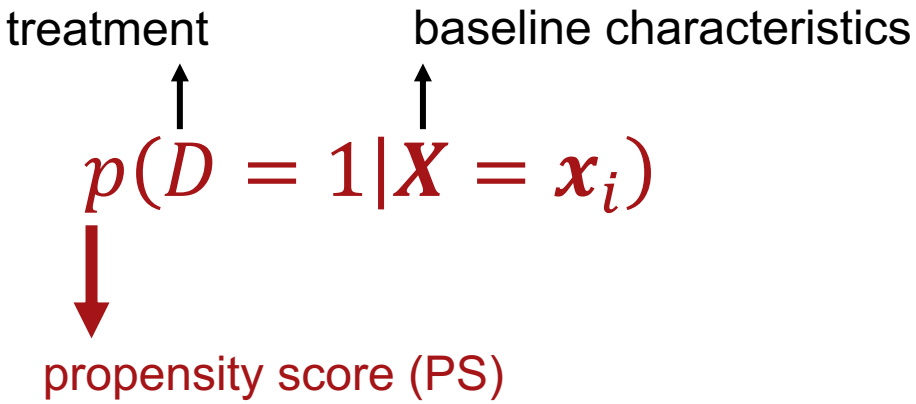

## Estimation (Health Equity)

Algorithmic bias: eGFR in nephrology [**Zhang** et al, *medRxiv* 2021]

Treatment allocation bias [**Zhang** et al, *arXiv* 2022]



Propensity score is a popular method to adjust for confounding.




Propensity score is the probability of a patient receiving the treatment given the patient's baseline characteristics.

# Propensity scores can be estimated from data

$$\text{logit}(\hat{p}) = \hat{\beta}_0 + \underbrace{\hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_j x_{ij}}_{\text{“confounders”}}$$

$\hat{p}$: propensity score

- In practice, propensity score is often estimated with a logistic regression.
- Treatment status is regressed on measured covariates that are believed to be confounders.

# Using propensity scores to remove confounding bias

Propensity scores can be used in various ways to remove confounding.

- Matching
- Stratification
- Inverse probability treatment weighting (IPTW)

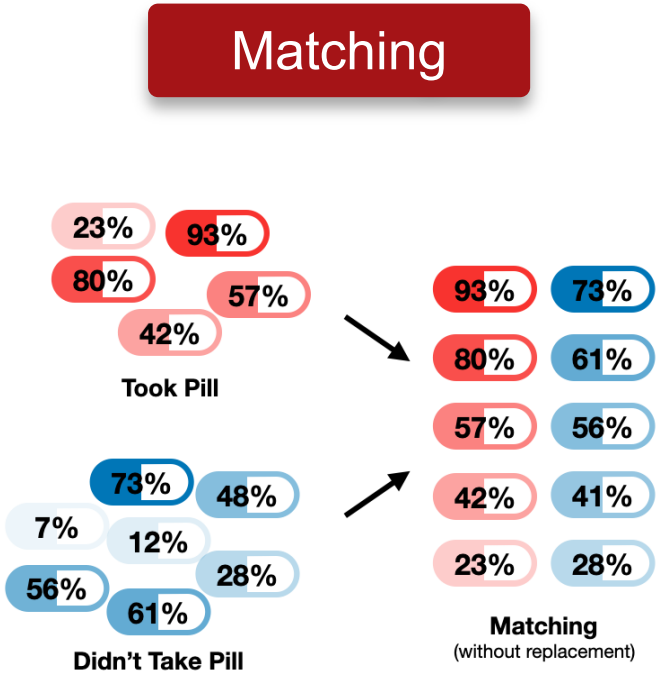
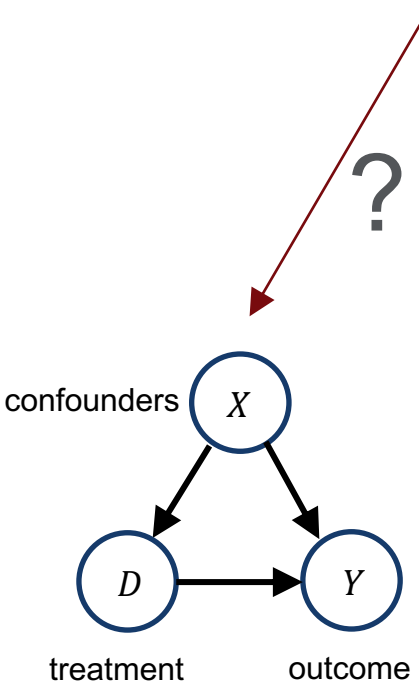



Image by Shawhin Talebi.

# How are confounders selected?




EHR database

Demographics | Diagnoses | Drugs | Procedures | Labs

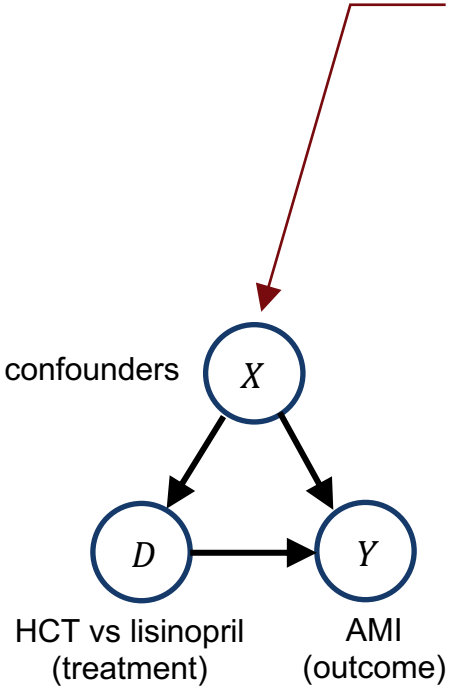
Domain Knowledge

Literature Review

Empirical association

(I²DB)

# How well do observational studies agree on confounders?




Hicks 2018: age, sex, year of cohort entry, body mass index, smoking status, alcohol related disorders (including alcoholism, alcoholic cirrhosis of the liver, alcoholic hepatitis, and hepatic failure), and history of lung diseases (including pneumonia, tuberculosis, and chronic obstructive pulmonary disease), duration of HTN Rx, statin use, #drugs

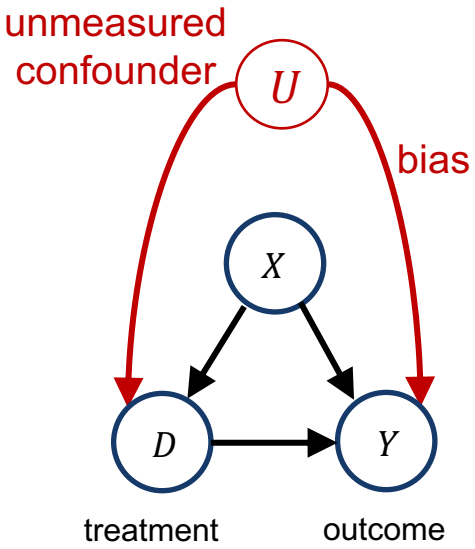
Ku 2018: age, sex, race, income status, baseline HF, baseline myocardial infarction, baseline peripheral artery disease, baseline stroke, baseline eGFR, baseline proteinuria, and time-dependent covariates including diabetes mellitus, obesity, systolic blood pressure, statin use, aspirin use, diuretic use, and concurrent use of other antihypertensive agents for the outcome of HF

Magid 2010: age, gender, days on thiazide prior to 2nd agent start, # of visits prior to thiazide, Mean Systolic BP, Mean Diastolic BP, Chronic Obstructive Pulmonary Disease, Hyperlipidemia, Cancer, Dementia, Chronic liver disease, Depression

Hasvold 2014: age, gender, elevated blood glucose, overweight and low socio-economic status are known risk factors for diabetes, High cholesterol and hypertension are additionally known risk factors for CVD

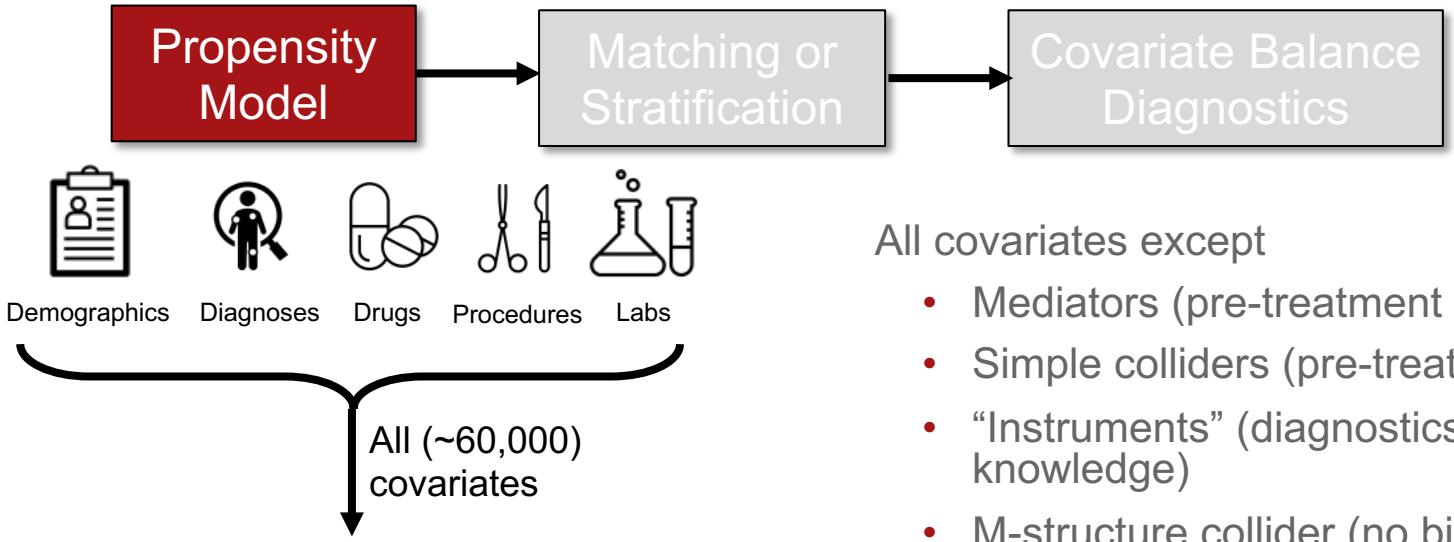
Blue: good agreement. Green: partial agreement. Red: lack of agreement

# Unmeasured confounder leads to biased estimates




- **Poor agreement** among manually selected confounders.
  - Confounding structure is rarely known.
- Missing a confounder will lead to **biased effect estimates**.
- Poor agreement on confounders across studies **weakens the reliability** of evidence from real-world data.

# Large-Scale Propensity Score (LSPS)



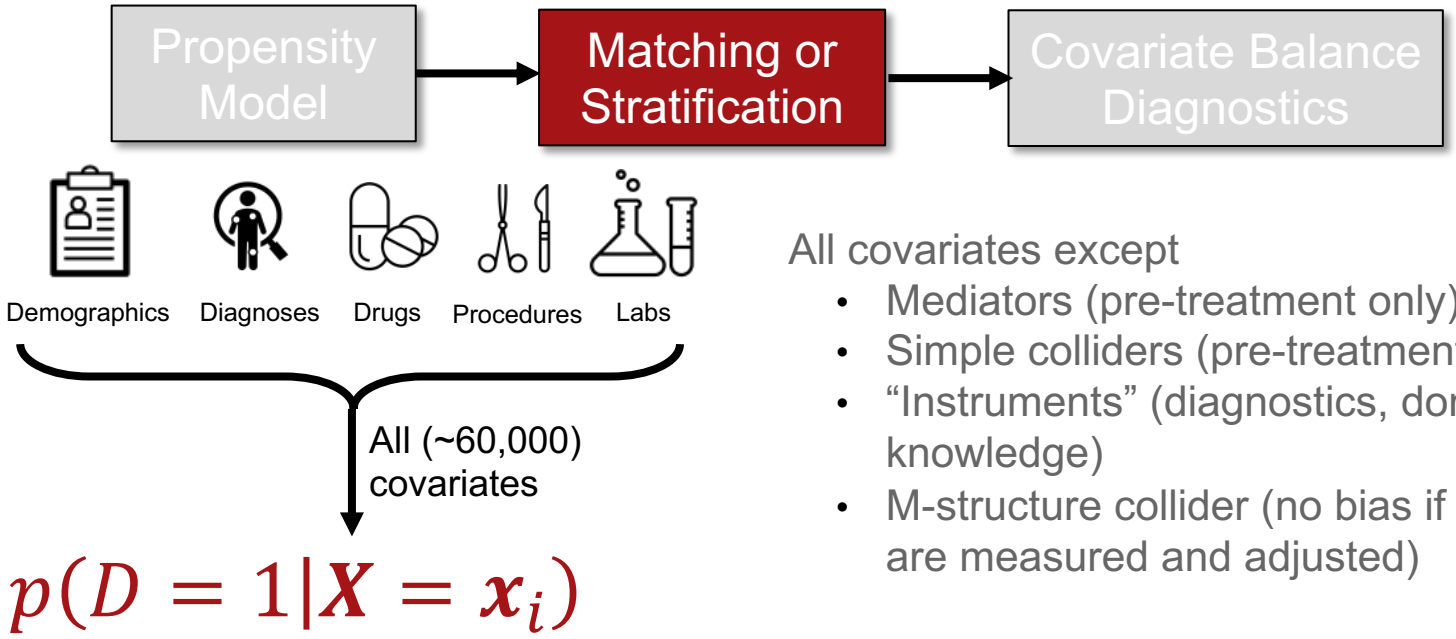

All covariates except

- Mediators (pre-treatment only)
- Simple colliders (pre-treatment only)
- "Instruments" (diagnostics, domain knowledge)
- M-structure collider (no bias if its causes are measured and adjusted)

LSPS includes (nearly) **all covariates** in the propensity model to adjust for confounding.

Tian Y et al. *Int J Epidemiol*. 2018.
Hripcsak G et al. *JAMA IM* 2020.

# Large-Scale Propensity Score (LSPS)




All covariates except

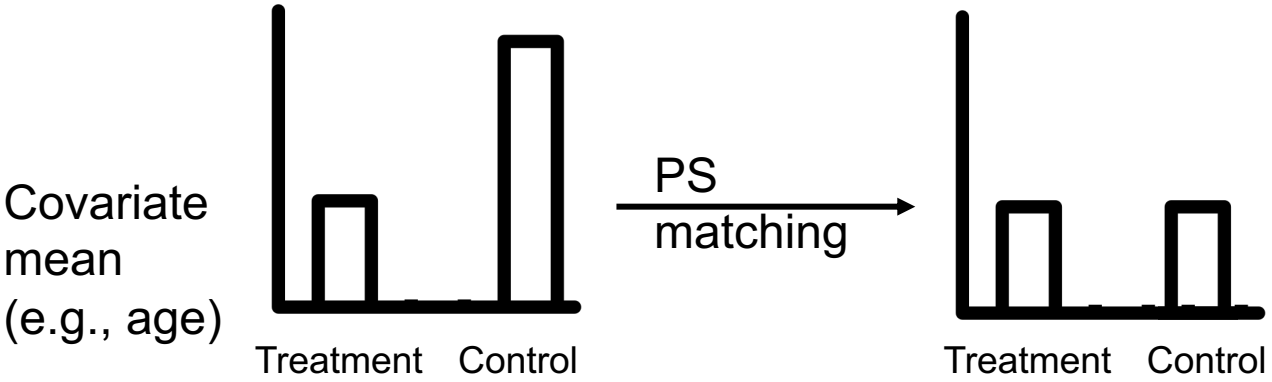
- Mediators (pre-treatment only)
- Simple colliders (pre-treatment only)
- "Instruments" (diagnostics, domain knowledge)
- M-structure collider (no bias if its causes are measured and adjusted)

LSPS includes (nearly) **all covariates** in the propensity model to adjust for confounding.

Tian Y et al. *Int J Epidemiol*. 2018.
Hripcsak G et al. *JAMA IM* 2020.

# Large-Scale Propensity Score (LSPS)

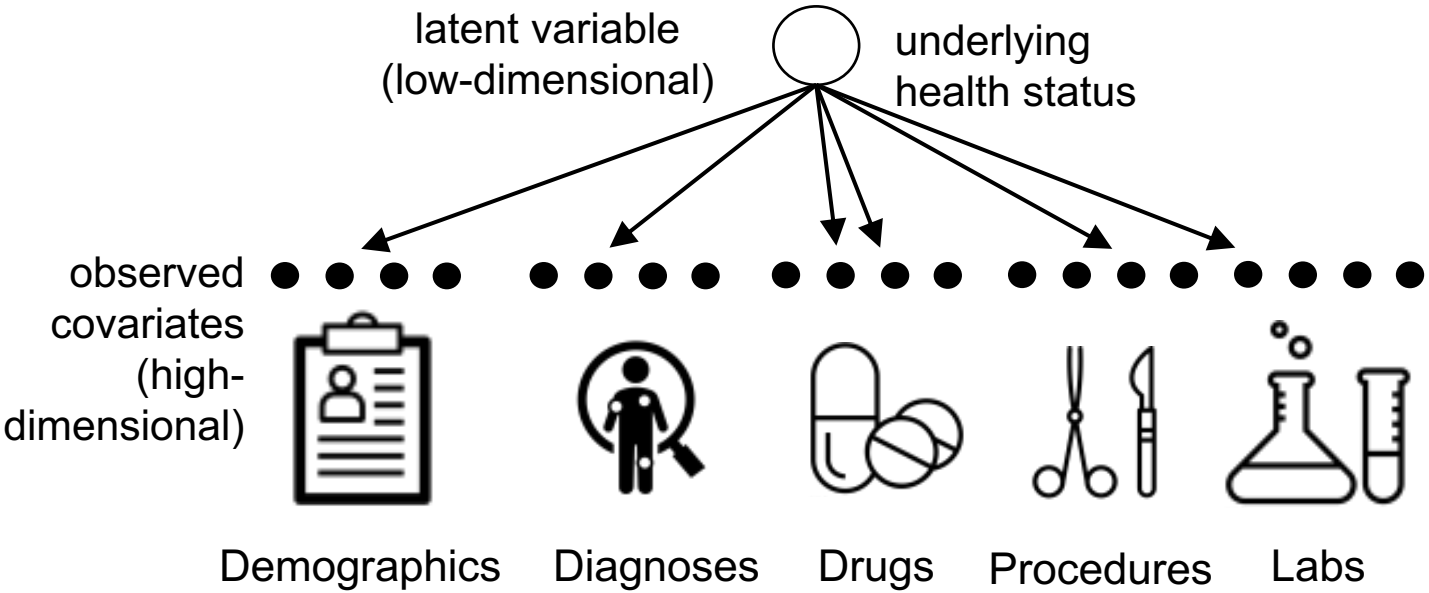


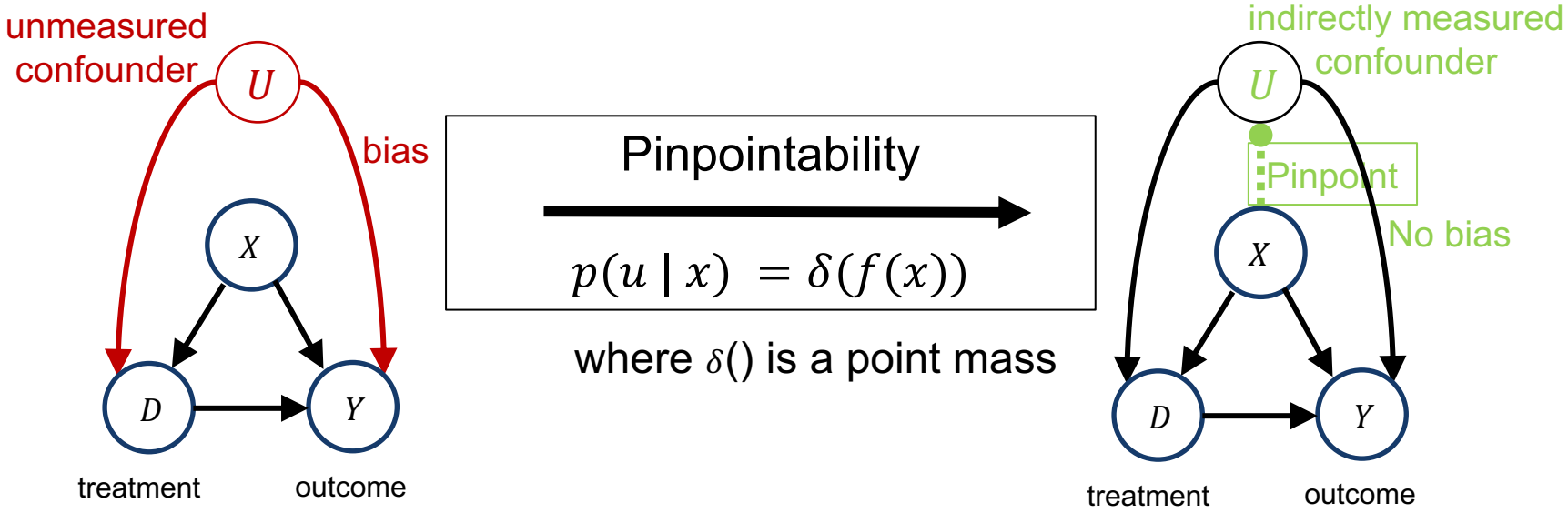

LSPS is observed to **balance on unmeasured (hold-out) confounders.**

# Intuition: Covariates in EHRs are highly correlated

An unmeasured confounder is likely correlated with other measured covariates through shared latent variables (e.g., underlying health status)



# LSPS adjusts for unmeasured confounder under pinpointability




Assumption 1 & Theorem 1 in [**Zhang** et al, *JBI* 2022]

# When the pinpointability assumption is relaxed…

The unmeasured confounder is only partially recovered

$$p(u \mid x) = f(x) + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$\sigma^2$ ↑ pinpointability ↓

Further assumes

$\mathbb{E}[D \mid U, X]$
$\mathbb{E}[Y \mid D, U, X]$ → both are monotonic in U

$$ATE_{true} \leq ATE_{LSPS} \leq ATE_{PS} \quad \text{or} \quad ATE_{true} \geq ATE_{LSPS} \geq ATE_{PS}$$

Theorem 2 in [**Zhang** et al, *JBI* 2022]

(I²DB)

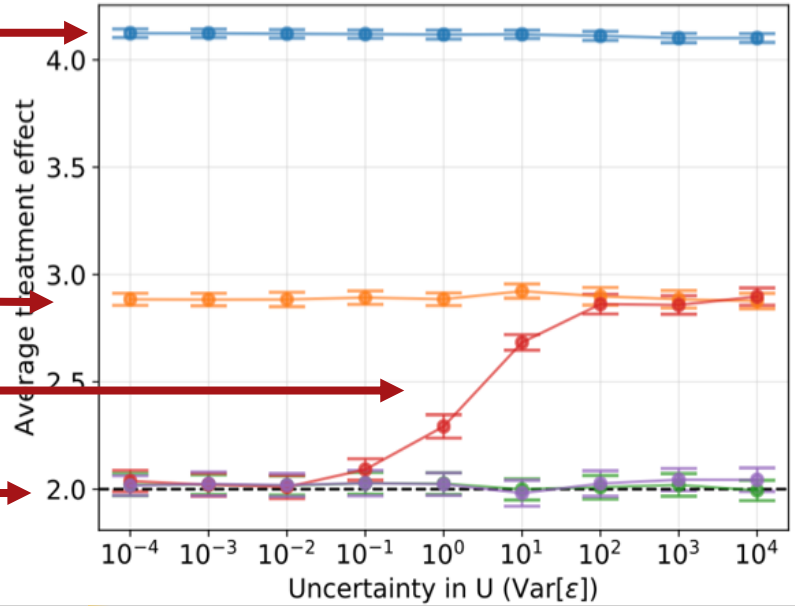
# Simulation:
# LSPS is more robust to unmeasured confounding

Unadjusted

Manual PS without U

LSPS without U

Manual PS with U
LSPS with U




Bias increases

Strong

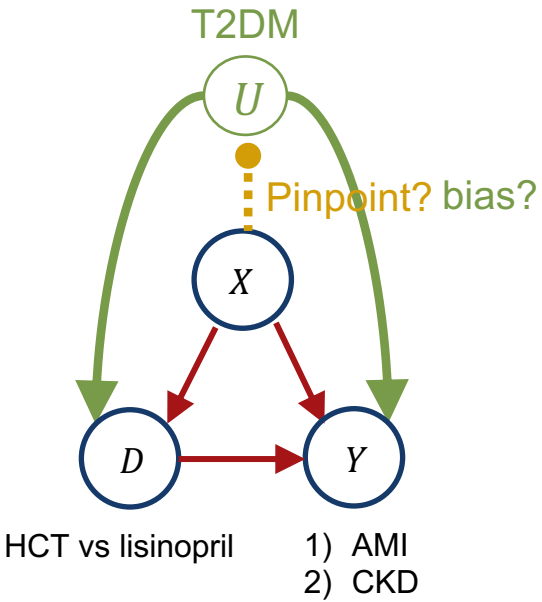
Weak

Pinpointability

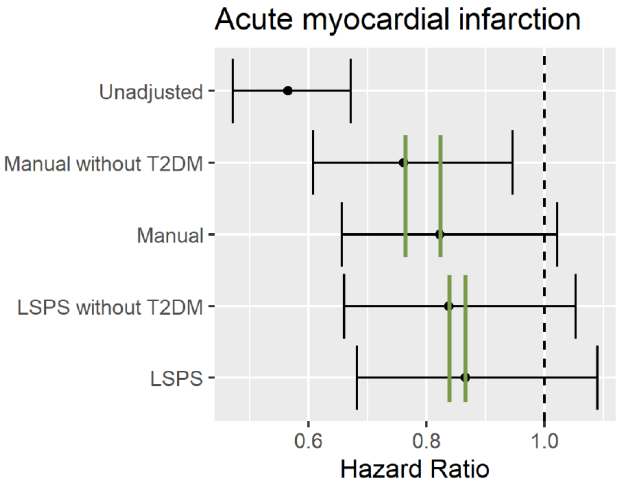
# Hypertension study with EHRs

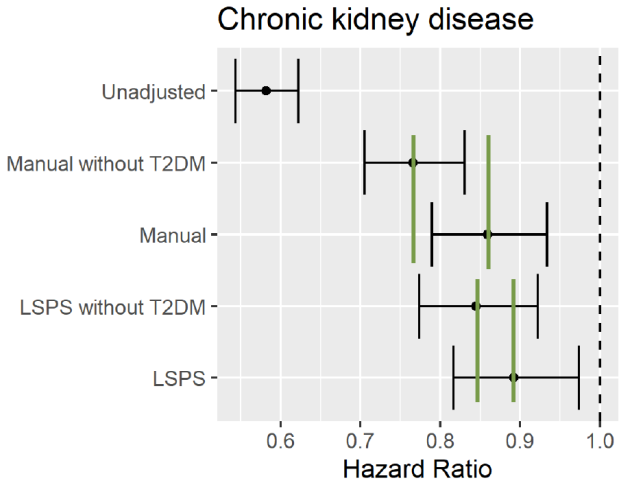
- Study design:
  - **Treatments**: hydrochlorothiazide (HCT) vs lisinopril (comparator)
  - Two clinical **outcomes**:
    - acute myocardial infarction (AMI)
    - chronic kidney disease (CKD).
  - **Unmeasured Confounder**: T2DM
    - Held out all T2DM-related conditions and medications
- Models:
  - unadjusted: no covariate was adjusted.
  - manual PS without T2DM
  - manual PS
  - LSPS without T2DM
  - LSPS

Manual:
All confounders from the literature.

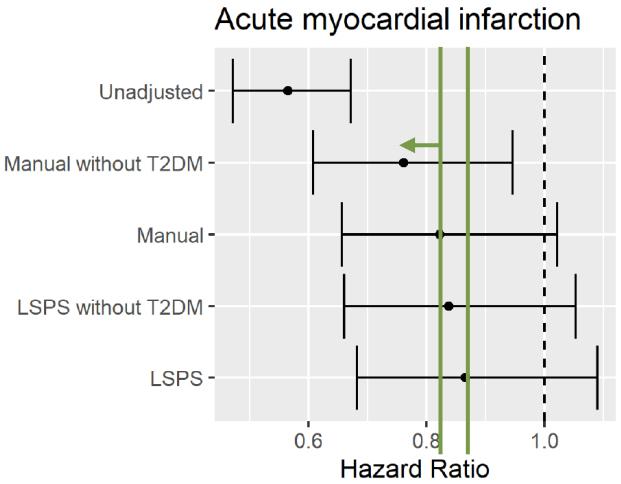


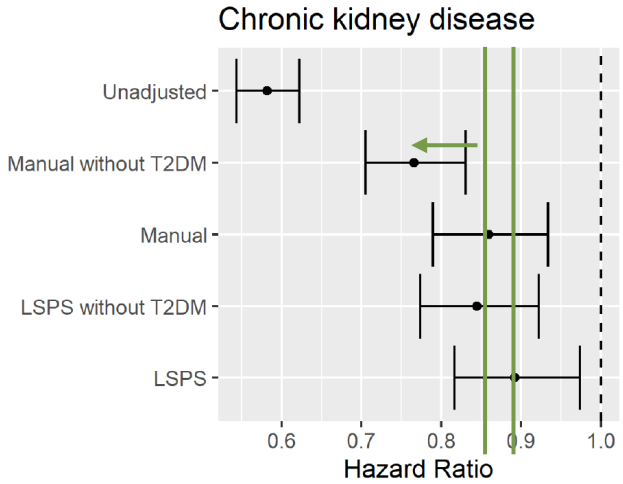
# LSPS is more robust to unmeasured confounder



# LSPS is more robust to unmeasured confounder



# Conclusions

- Under pinpointability, LSPS can produce unbiased or less biased estimates than classical PS adjustment methods.
- Pinpointability holds reasonably well when LSPS is used than traditional PS with covariate selection.
- LSPS is implemented in CohortMethod, an open-source R package
  - https://github.com/OHDSI/CohortMethod

(I²DB)

# Reliable RWE generation at large-scale

All health outcomes of interest

All drugs



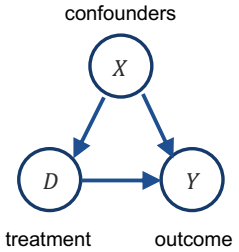
LSPS is used in **100+ clinical publications** to generate reliable real-world evidence: https://www.ohdsi.org/publications/

# Part II: Reliability of health equity assessment

# Part II: Reliability of health equity assessment

## Estimation (Drug Effect)

The medical deconfounder [**Zhang** et al, *MLHC* 2019]
Large-scale propensity score [**Zhang** et al, *JBI* 2022]



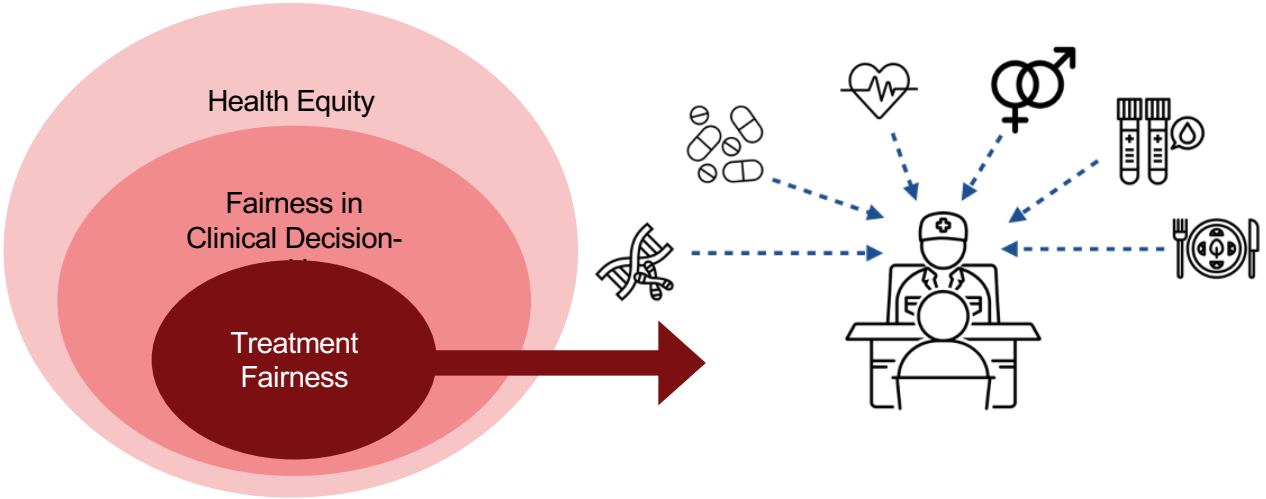

## Estimation (Health equity)

Algorithmic bias: eGFR in nephrology [**Zhang** et al, *medRxiv* 2021]
Treatment allocation bias [**Zhang** et al, *arXiv* 2022]



# Fairness in clinical decision-making

- Fairness in clinical decision-making is an important component of health equity.
- Many factors could potentially affect a treatment decision.

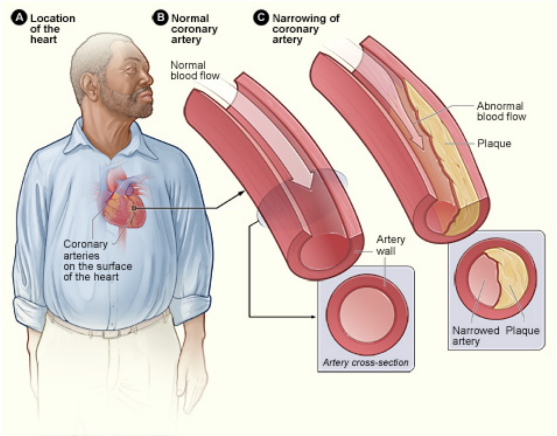



Goal: Assess fairness of treatment allocation with EHRs.

# Example: Coronary artery disease (CAD)

- **Coronary artery disease** is the most common type of heart disease.
- Patients with coronary artery disease are at higher risk of developing AMI.

Phenotype




Coronary artery disease (CAD)

Outcome

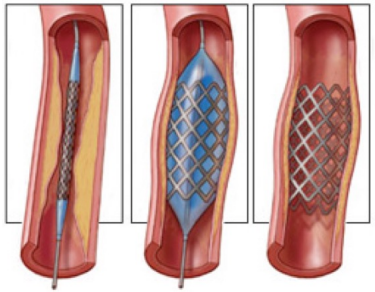


Acute myocardial infarction (AMI)

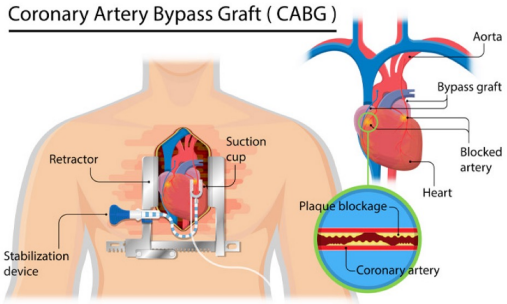
# Example: Coronary artery disease (CAD)

- **Women**, **racial and ethnic minorities**, patients without **health insurance**, and those who live in **low-income** neighborhoods may have inadequate access to revascularization.

Treatment



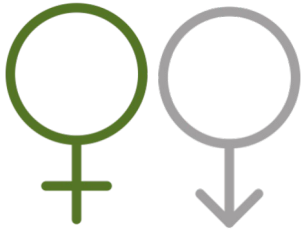
Percutaneous coronary intervention (PCI)




Coronary artery bypass graft (CABG)

Revascularization (PCI and CABG)

Sensitive attribute

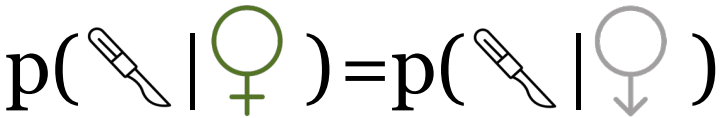


biological sex (or race)

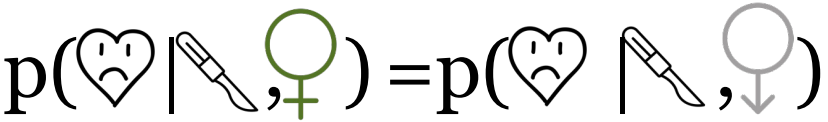
**Question: Is there sex-related bias in allocating revascularization?**

# Popular (associational) fairness definitions

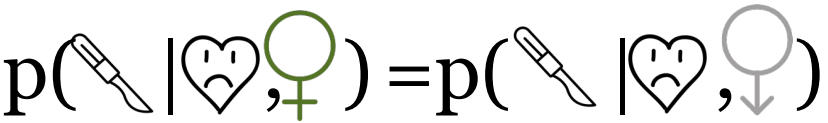
**Statistical parity**



**Calibration**

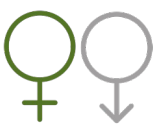


**Accuracy**



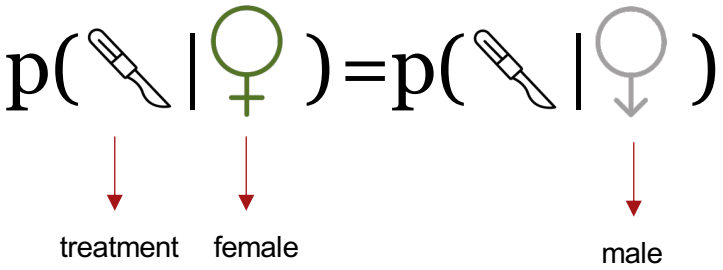
 treatment

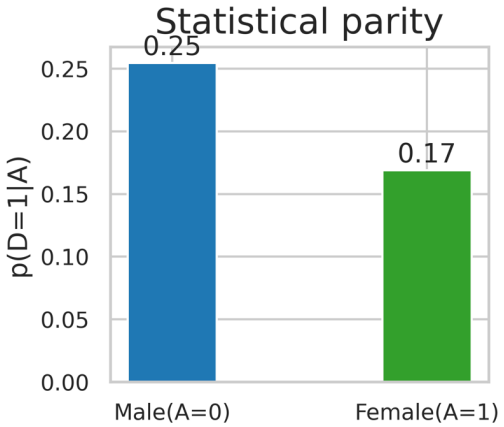
 outcome

 sensitive attribute: sex

# Statistical parity

**Question**: Is the treatment assigned at equal rate between men and women?

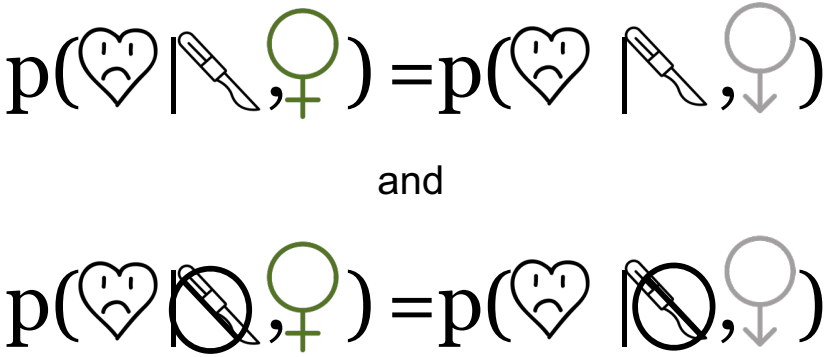


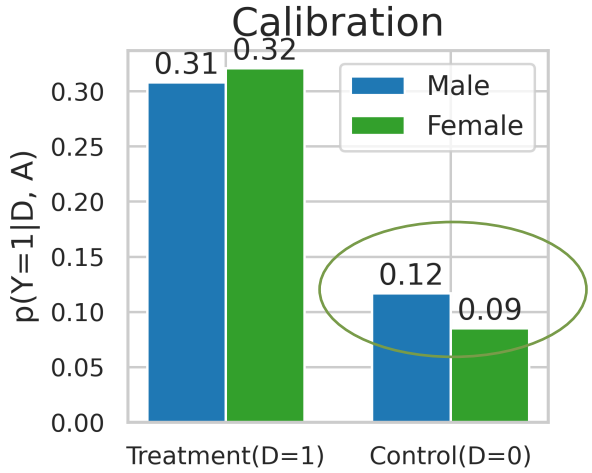




**Result**: Male patients were more likely to receive revascularization treatment than female patients. Bias against women?

# Calibration

**Question**: Does heart attack happen at equal rate between men and women, given their treatment status?




**Result:** Heart attack happened more frequently for male patients than for female patients in the control group. Maybe bias against men?

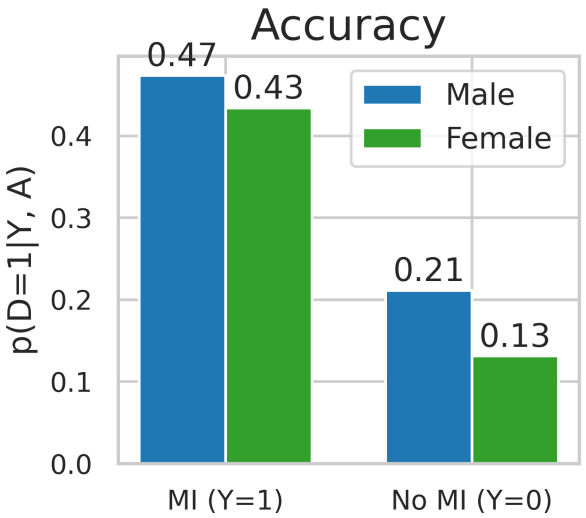
# Accuracy

**Question**: Is the treatment assigned at equal rate between men and women, given their (observed) outcome?

$$p(\text{🔪} \mid \text{☹}, \text{♀}) = p(\text{🔪} \mid \text{☹}, \text{♀})$$

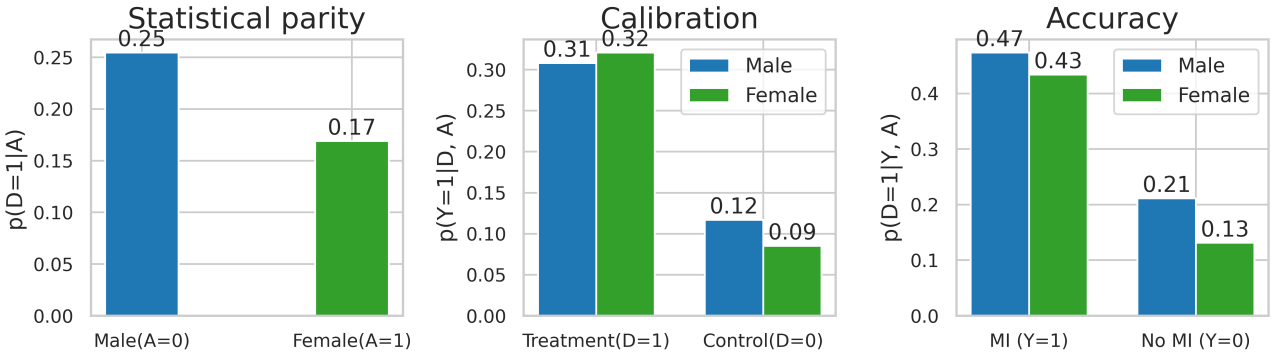
and

$$p(\text{🔪} \mid \text{☺}, \text{♀}) = p(\text{🔪} \mid \text{☺}, \text{♀})$$




**Result:** Male patients are more likely to receive the treatment than female patients, regardless of their outcome status. Bias against women?

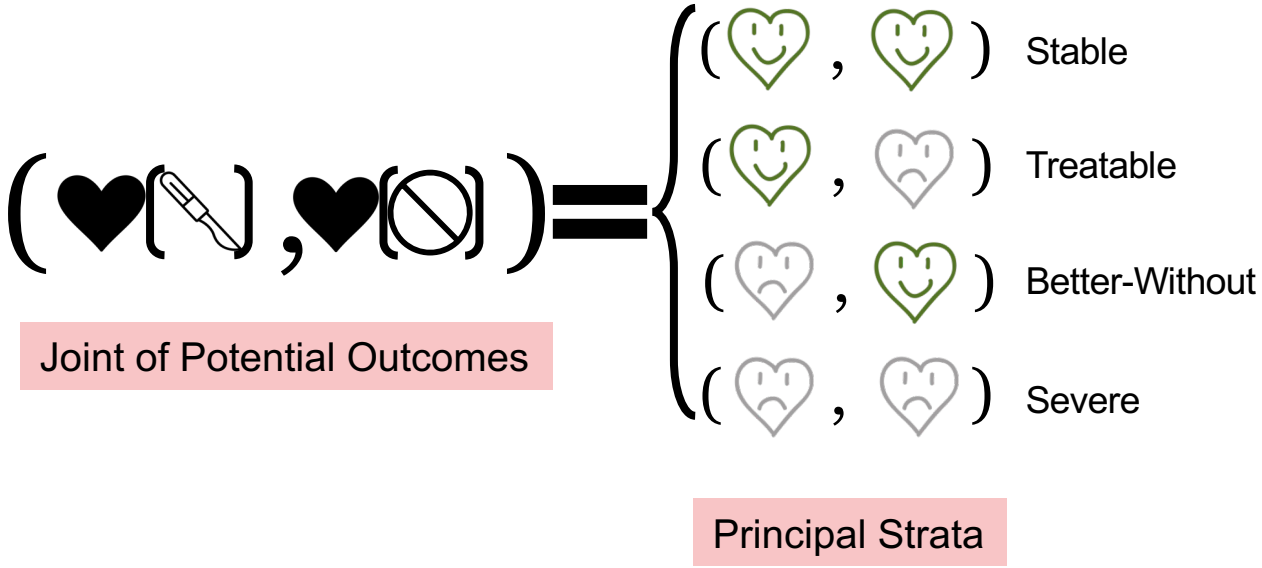
# Limitations of associational definitions




- **Conclusions about fairness differ** depending on which metric we use.
- All metrics fail to consider differences in baseline characteristics among patients.

# Principal Fairness: A Causal Fairness

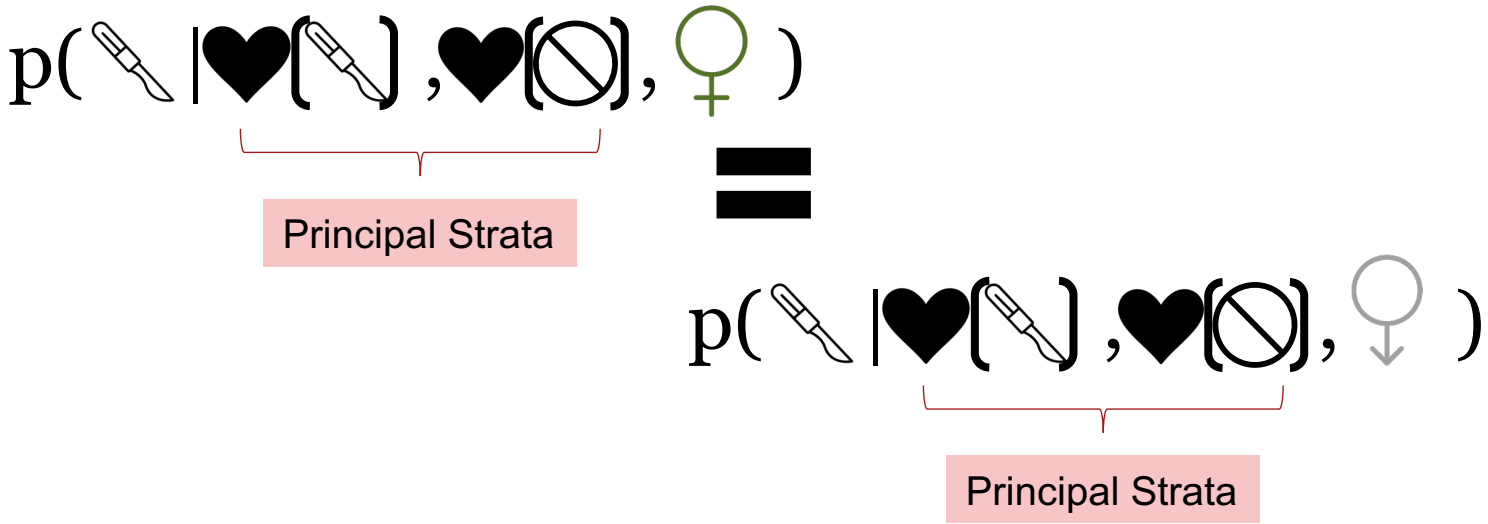
- **Idea**: Assess fairness among **patients who would benefit similarly** from the treatment.
- **Question**: How to **define patient similarity** based on treatment response?




[Imai K and Jiang Z. *Statistical Science*, 2022]

# Principal Fairness: A Causal Fairness

- A treatment satisfies principal fairness if the treatment is assigned at **equal rates** between men and women who would **benefit equally from the treatment** (i.e., patients in the same principal stratum).



# Assumptions and estimation of principal fairness

**Recall that joint potential outcomes are never observed.**

**Estimation** of principal fairness requires fitting three models:

1. Propensity model $p(D|X)$
2. Potential outcome model $p(Y(0) \mid X)$
3. Potential outcome model $p(Y(1) \mid X)$

$Y_i(1) - Y_i(0)$ is **never** observed!

All pre-treatment covariates

**Assumptions**:

1. No unmeasured confounder
2. Conditional independence of potential outcomes



# Estimation of principal fairness

ASSUMPTION 1 (Conditional independence of potential outcomes).

$$Y(1) \perp\!\!\!\perp Y(0) \mid X$$

ASSUMPTION 2 (Unconfoundedness).

$$Y_i(1), Y_i(0) \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

Our target estimand is

$$p(D \mid Y(0), Y(1), A) = \frac{p(D, Y(0), Y(1) \mid A)}{p(Y(0), Y(1) \mid A)}$$

The denominator is

$$\begin{aligned} p(Y(0), Y(1) \mid A) &= \int p(Y(0), Y(1) \mid X, A) p(X|A) dx \\ &= \int p(Y(0) \mid X, A) p(Y(1) \mid X, A) p(X|A) dx \\ &= \int p(Y \mid D = 0, X, A) p(Y \mid D = 1, X, A) p(X|A) dx \\ &\approx \frac{1}{n_a} \sum_{i:A_i=a} p(Y \mid D = 0, X, A) p(Y \mid D = 1, X, A) \end{aligned}$$

The first equality holds because of law of total probability. The second holds because of Assumption 1. The third equality holds because of Assumption 2.

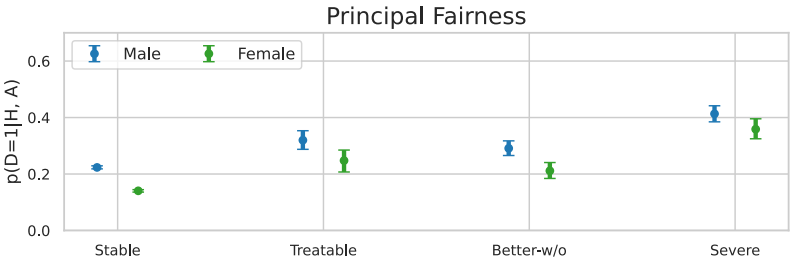
The numerator is

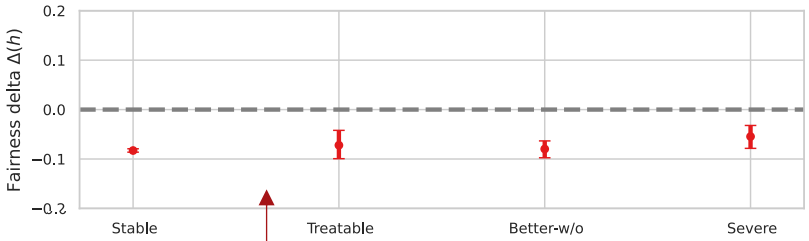
$$\begin{aligned} p(D, Y(0), Y(1) \mid A) &= \int p(D, Y(0), Y(1) \mid X, A) p(X|A) dx \\ &= \int p(D \mid X, A) p(Y(0), Y(1) \mid X, A) p(X|A) dx \\ &= \int p(D \mid X, A) p(Y(0) \mid X, A) p(Y(1) \mid X, A) p(X|A) dx \\ &\approx \frac{1}{n_a} \sum_{i:A_i=a} p(D \mid X, A) p(Y \mid D = 0, X, A) p(Y \mid D = 1, X, A) \end{aligned}$$

The first equality holds because of law of total probability. The second holds because of Assumption 2. The third equality holds because of Assumption 1.

(I²DB)

# Principal fairness (sex) with EHR data




Women were treated less often than men.

Contributions:

- Generalized the original formulation of principal fairness
- Developed an estimation algorithm for principal fairness
- Demonstrated its performance with EHR data

(I²DB)

## Prediction

Post-bariatric surgery outcomes [Richter, Albert, **Zhang** et al, *Front. Physiol*. 2022]

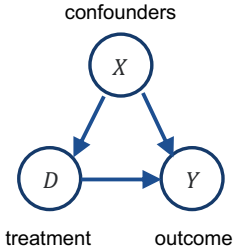
Hospitalization of COVID-19+ [Song, **Zhang** et al, *JAMIA* 2021]

Pressure injury [Song, Kang, **Zhang**, et al, *JAMIA* 2020]

## Estimation (Drug Effect)

**The medical deconfounder** [**Zhang** et al, *MLHC* 2019]

**Large-scale propensity score** [**Zhang** et al, *JBI* 2022]




## Estimation (Health Equity)

Algorithmic bias: eGFR in nephrology [**Zhang** et al, *medRxiv* 2021]

**Treatment allocation bias** [**Zhang** et al, *arXiv* 2022]



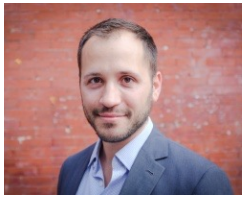
# Acknowledgments



George Hripcsak


David Blei


Noemie Elhadad


Nicholas Tatonetti
(Cedars-Sinai)


Yixin Wang
(Univ. of Michigan)


Anna Ostropolets


Lauren Richter


Martijn Schuemie
(Janssen R&D)


Wenyu Song
(Mass General Brigham)

COLUMBIA | COLUMBIA UNIVERSITY DEPARTMENT OF BIOMEDICAL INFORMATICS

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS



linyingz@wustl.edu

@Z_Linying

Washington University School of Medicine in St. Louis

# Questions?

linyingz@wustl.edu
https://linyingzhang.com