

November, 17th 2021

Towards causal model selection

Matthieu Doutreligne

Claire Morgand, HAS (Service Evaluation et Outils pour la Qualité et la Sécurité des Soins)

Gaël Varoquaux, INRIA (Equipe Social Data)

Towards causal model selections for big observational data

I. Causal inference intro and motivations

II. Upper bound on the PEHE

III. Empirical Study

IV. Ongoing questions

Towards causal model selections for big observationnal data

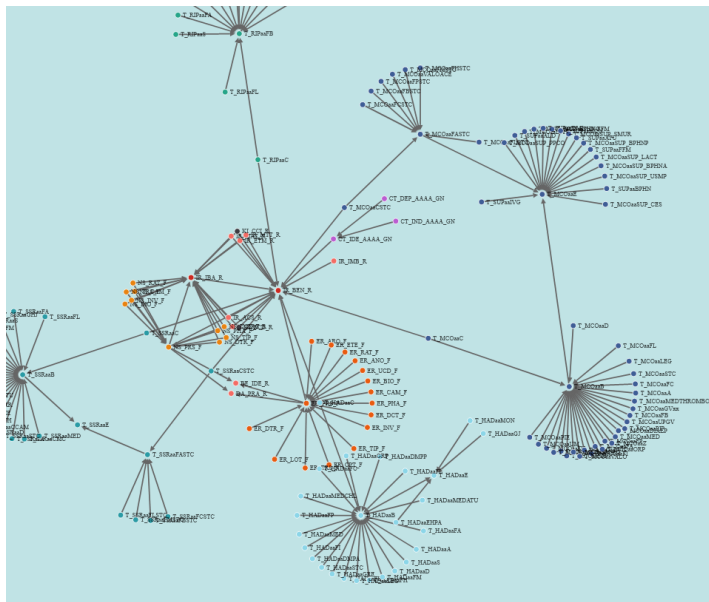
I. Causal inference intro and motivations

II. Upper bound on the PEHE

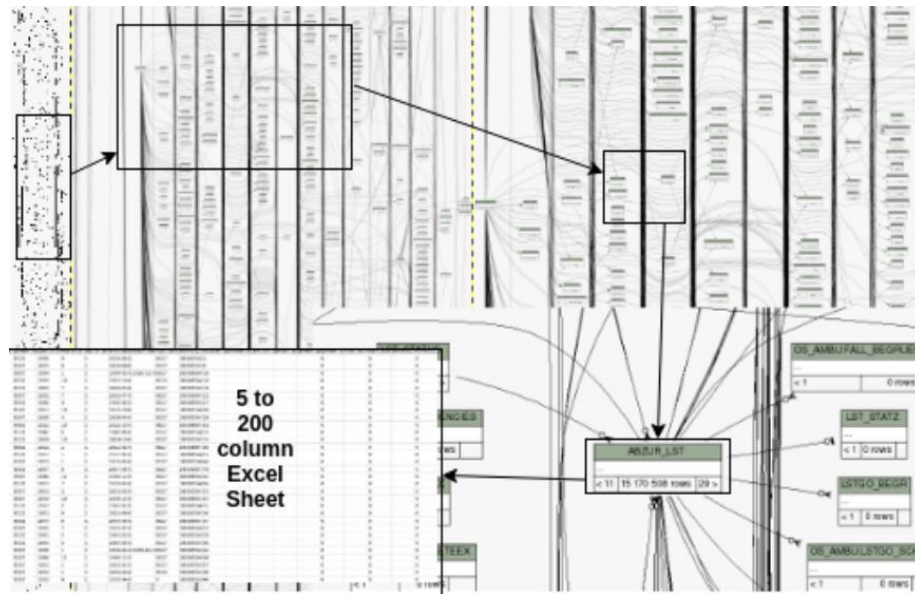
III. Empirical Study

IV. Ongoing questions

Big Healthcare Databases: aka observational data



Medico-administrative data (claims) :
ex. [SNDS](#)
Healthcare consumption, reimbursements



Electronic Health Records:
ex. [APHP](#) datamart, Lille, Bordeaux, ...
Detailed clinicals variables, notes, ...

 **Promesses**

Real world data, almost free data, huge pile of data (stastical power)

 **Difficulties**

Quality, confounders, complexity, heterogeneity, missingness, high-dimensionality, big volumes

 **Promesses**

Real world data, almost free data, huge pile of data (stastical power)

 **Difficulties**

Quality, confounders, complexity, heterogeneity, missingness, high-dimensionality, big volumes

GOAL

**Evaluate health technology and practices
Focus on guideline evaluation**

Guidelines evaluation



Target Patient Population with Features X

Example (stroke initial healthcare)

Patients with stroke related symptoms (TIA, stroke)



For whom, it is recommended to **Intervene with action A**

Perform cerebral scan / MRI as soon as possible



Aiming at improving a pertinent clinical **outcome Y**

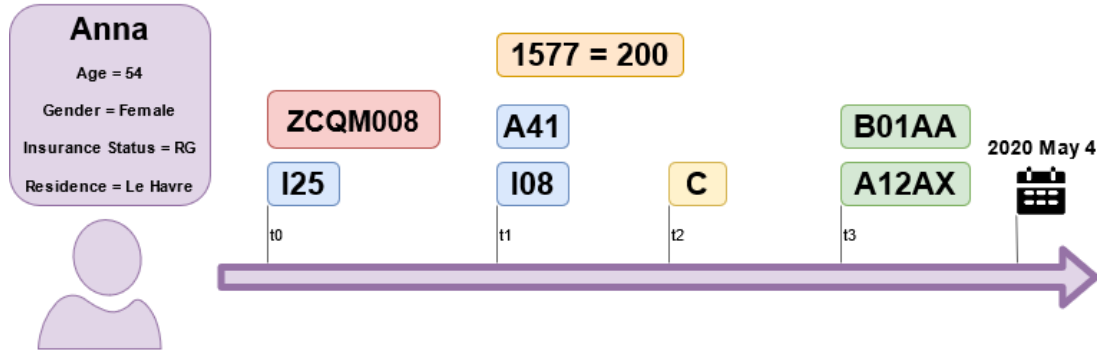
Better survival or reeducation



How to measure the effect of A on Y for the target population with features X ?



Neyman-Rubin Potential Outcome

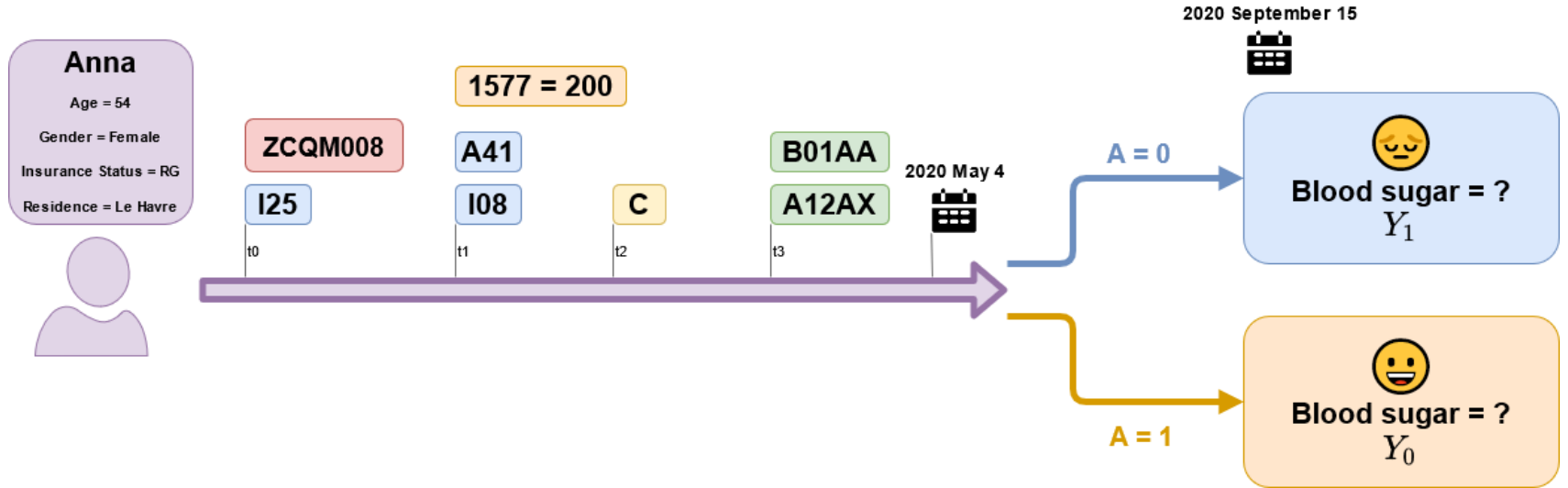


Covariates X

( biology / comorbidities...)



Neyman-Rubin Potential Outcome



Covariates X
 (📌 biology / comorbidities...)

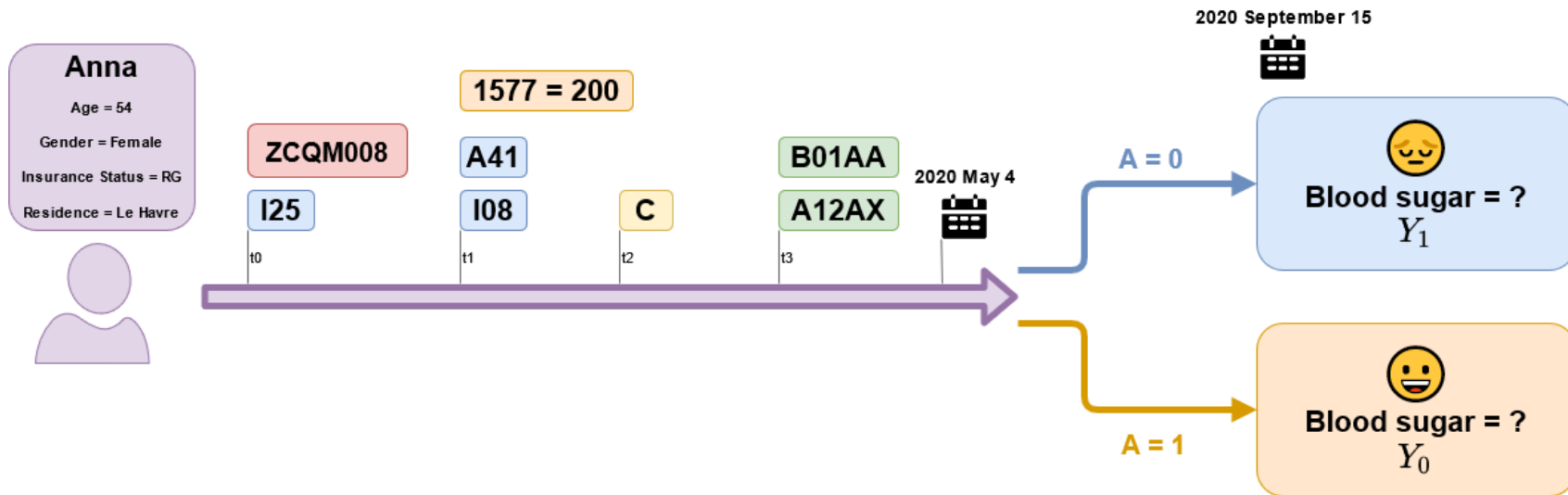
Intervention A
 (💊 drug / act)

Outcome Y



Neyman-Rubin Potential Outcome

$$\Delta = Y(1) - Y(0)$$



Covariates X

(🪄 biology / comorbidities...)

Intervention A
(💊 drug / act)

☑️ **Outcome Y**

Target Estimands

Complete (unobserved) distribution

$$(Y(1), Y(0), X, A) \sim \mathcal{D}^*$$

Factual (observed) distribution

$$(Y(A), X, A) \sim \mathcal{D}$$

Target Estimands

Complete (unobserved) distribution

$$(Y(1), Y(0), X, A) \sim \mathcal{D}^*$$

Factual (observed) distribution

$$(Y(A), X, A) \sim \mathcal{D}$$



Individual Treatment Effect

$$\Delta = Y(1) - Y(0)$$

Target Estimands

Complete (unobserved) distribution

$$(Y(1), Y(0), X, A) \sim \mathcal{D}^*$$

Factual (observed) distribution

$$(Y(A), X, A) \sim \mathcal{D}$$

 Individual Treatment Effect

$$\Delta = Y(1) - Y(0)$$

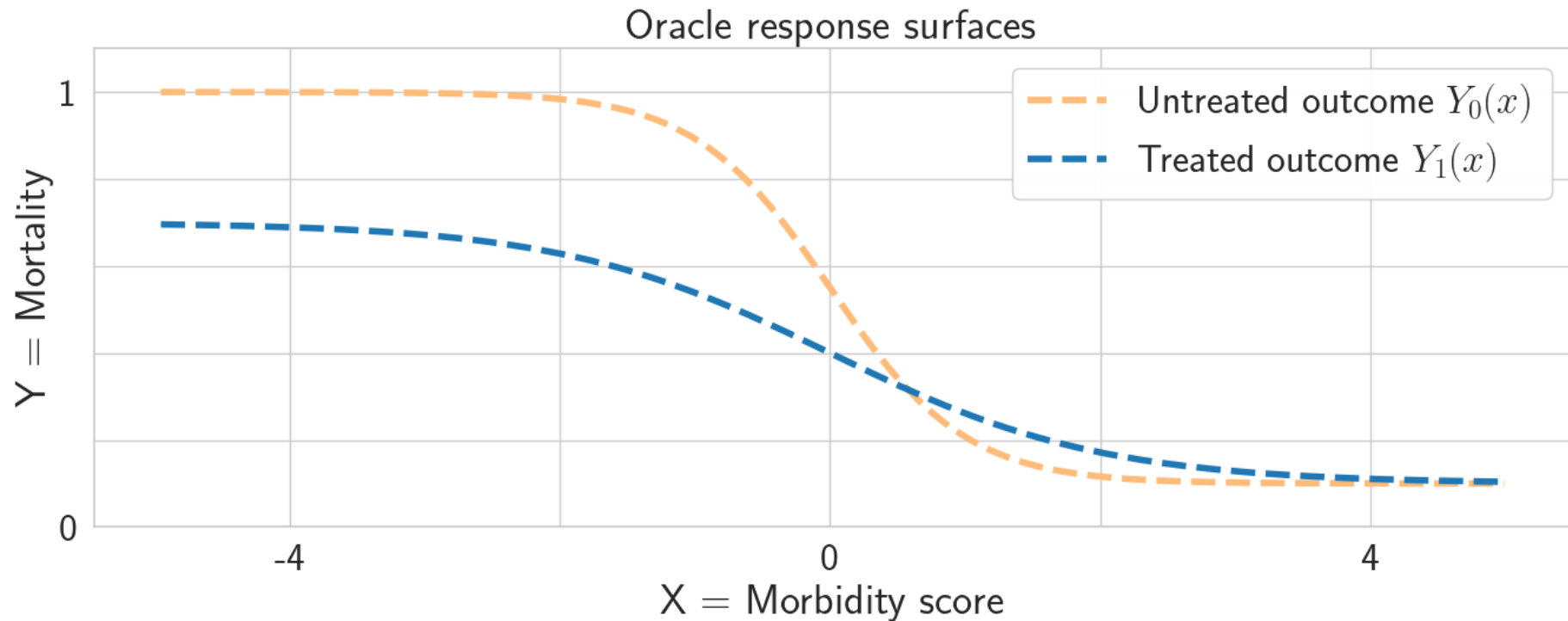
Average Treatment Effect (ATE) :

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

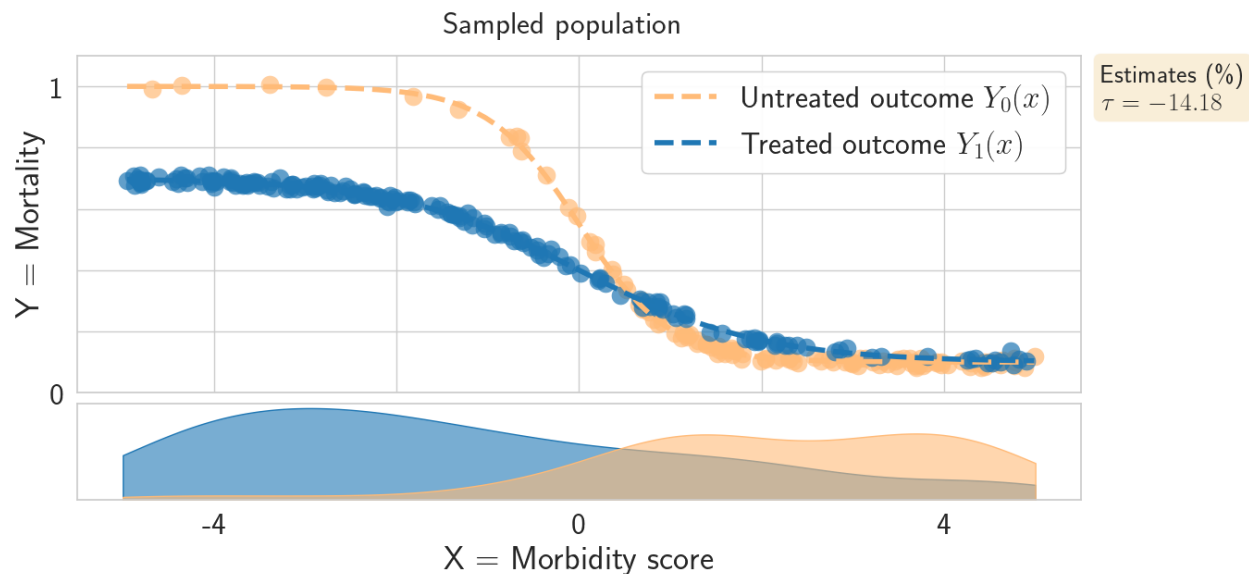
Conditional Treatment Effect (CATE) :

$$\tau(x) = \mathbb{E} [Y(1) - Y(0) | X = x]$$

Simulated example



Simulated example



$P(X, A)$

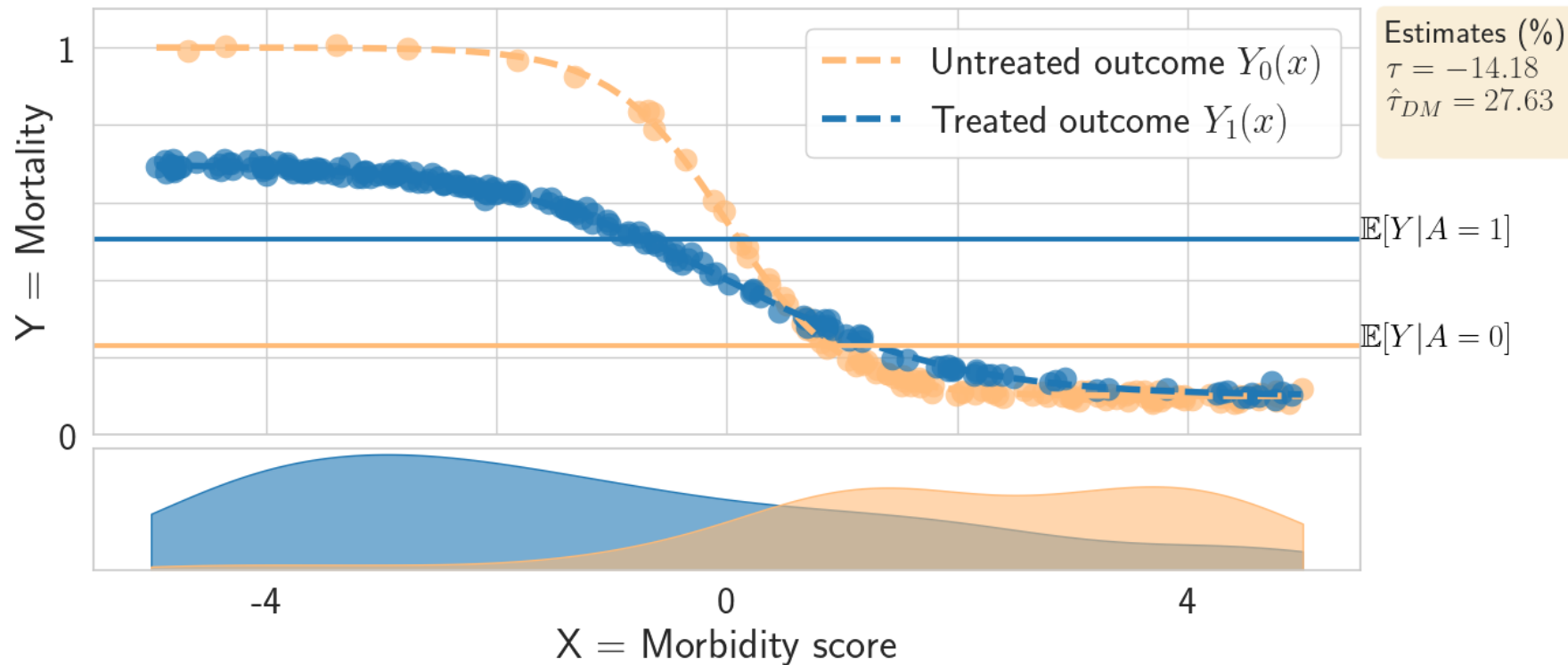
A Naive solution: The Difference in Mean

$$Z_{DM} = \frac{1}{\sum_i A_i} \sum_{i: A_i=1} Y_i - \frac{1}{\sum_i 1-A_i} \sum_{i: A_i=0} Y_i$$

Simulated example

Difference in Mean

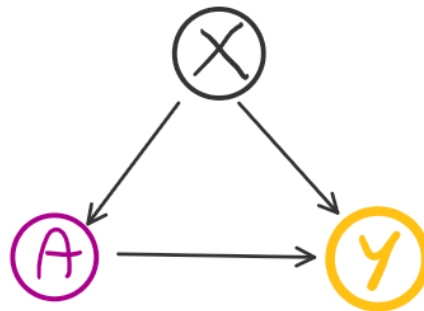
Sampled population



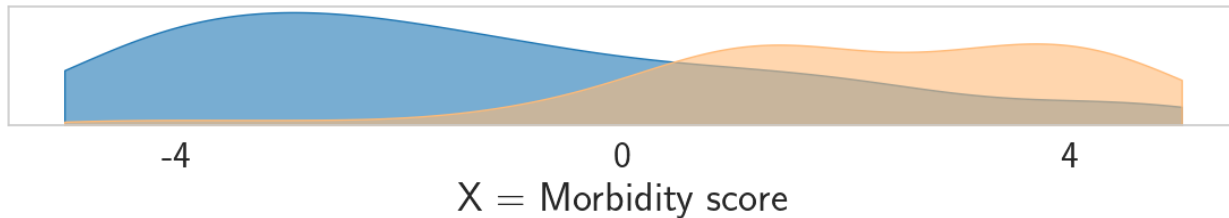
A Naive solution: The Difference in Mean

$$Z_{DM} = \frac{1}{\sum_i A_i} \sum_{i: A_i=1} Y_i - \frac{1}{\sum_i 1-A_i} \sum_{i: A_i=0} Y_i$$

☹️ **Coufounders X** (Treatment bias)



Treated and **non-treated**
are not the same





Causal assumptions: 1 – Ignorability (conditionnal exchangeability)

Enough information available to capture differences between **treated** and **control** populations

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$$

Causal assumptions: 1 – Ignorability (conditionnal exchangeability)

Enough information available to capture differences between **treated** and **control** populations

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$$


⚠ Not verifiable with data only -> call to domain expert 

Causal assumptions: 1 – Ignorability (conditionnal exchangeability)

Enough information available to capture differences between **treated** and **control** populations

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$$

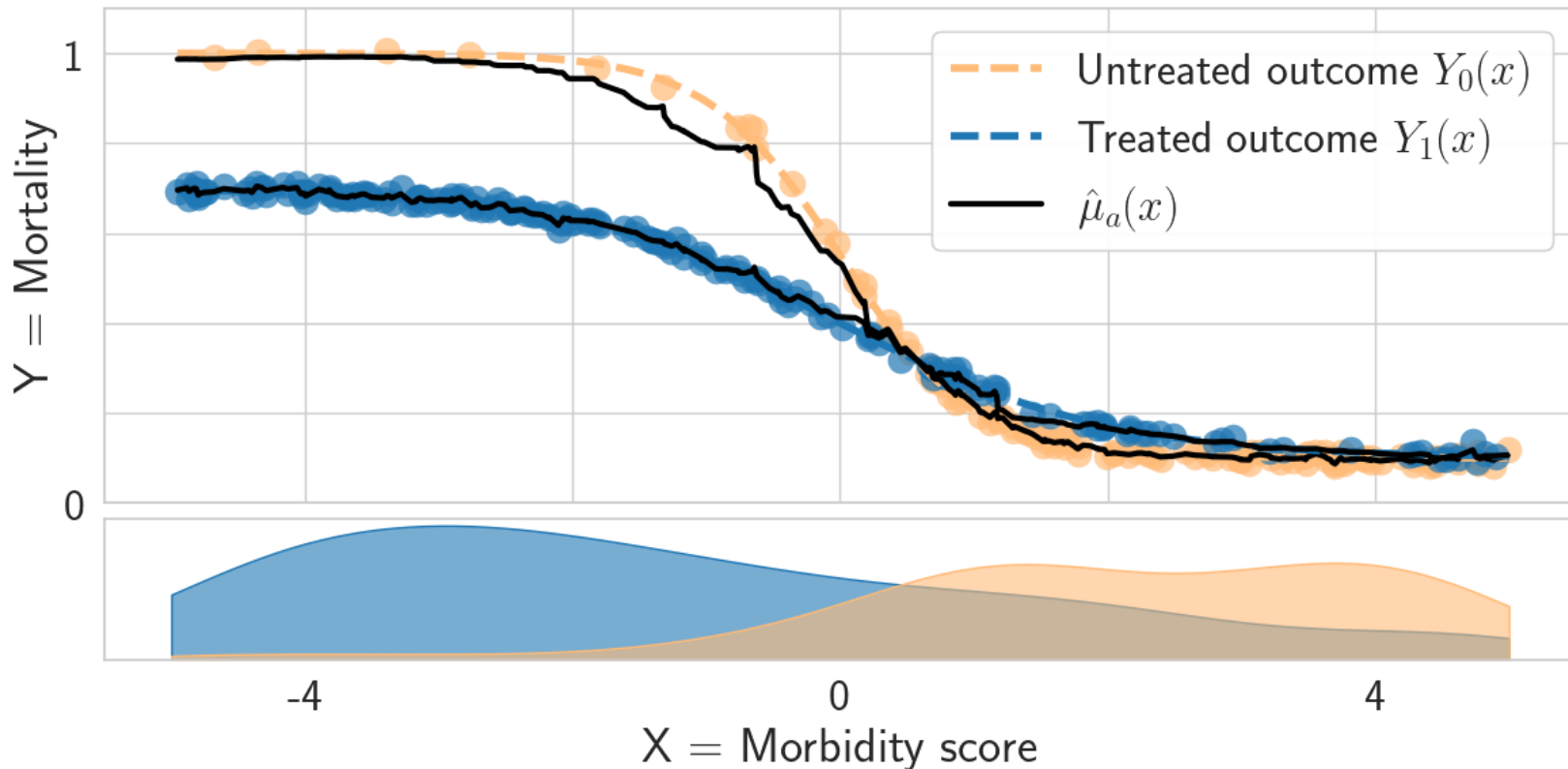
⚠ Not verifiable with data only -> call to domain expert 

 Legally, a practitioner has to log into the medical records all the information on which he/she based his/her decision !

Outcome model

💡 outcome Modelization

g-formula, regression, response surface fitting



Estimates (%)

$\tau = -14.18$

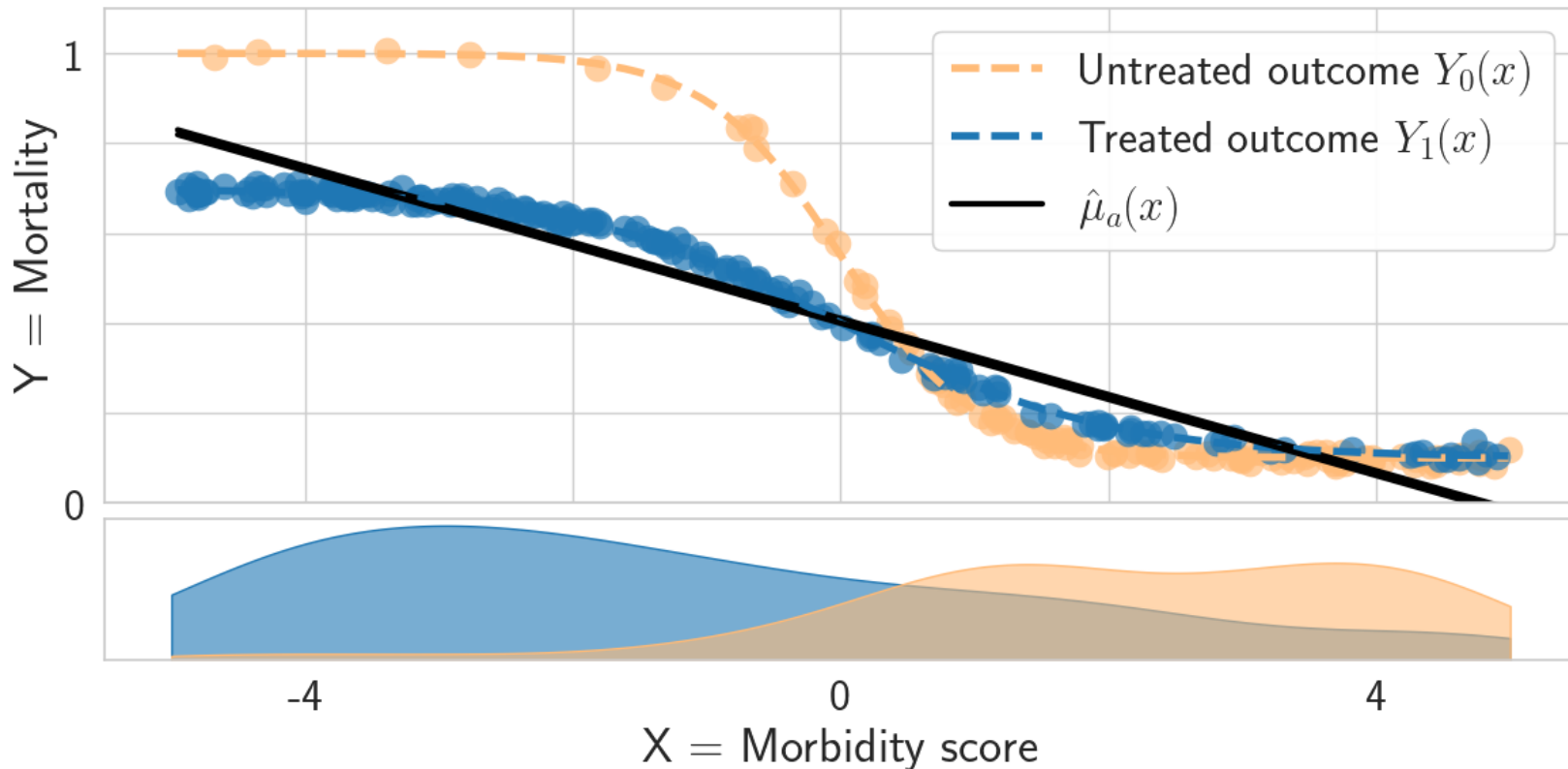
$\hat{\tau}_{DM} = 27.63$

$\hat{\tau}_G(\hat{\mu}) = -12.68$

$\tau\text{-risk}(\hat{\mu}) = 0.15$

Outcome model

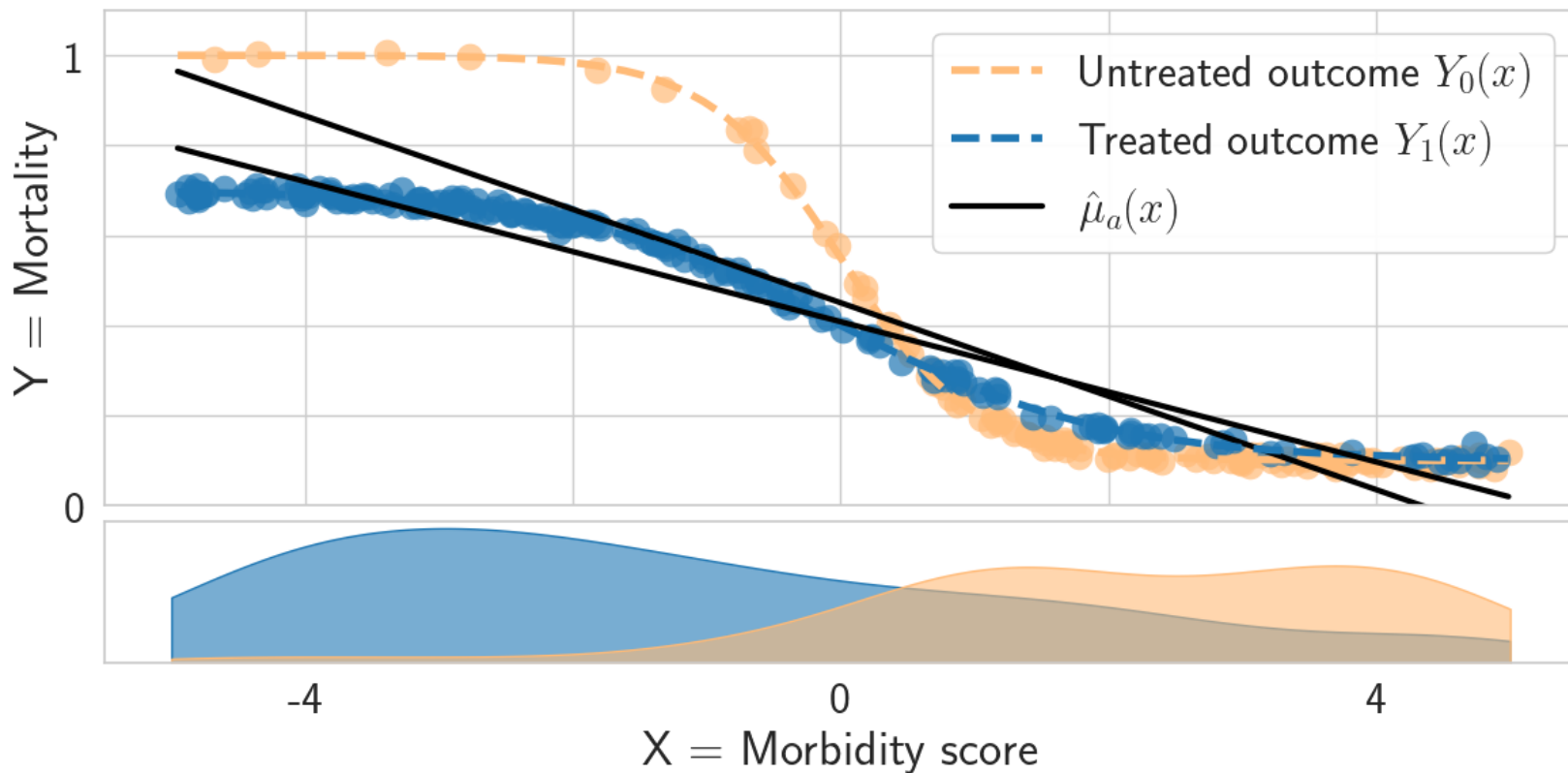
⚠ specification bias



Estimates (%)
 $\tau = -14.18$
 $\hat{\tau}_{DM} = 27.63$
 $\hat{\tau}_G(\hat{\mu}) = -1.07$
 $\tau\text{-risk}(\hat{\mu}) = 4.84$

Outcome model

⚠ Specification + Extrapolation biases



Estimates (%)

$\tau = -14.18$

$\hat{\tau}_{DM} = 27.63$

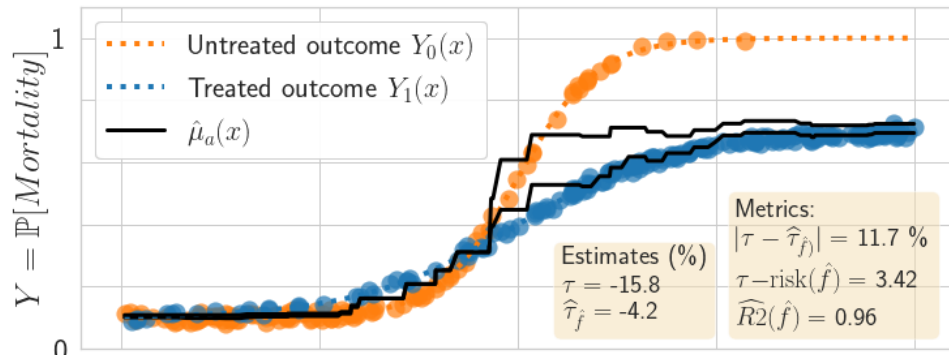
$\hat{\tau}_G(\hat{\mu}) = -3.92$

$\tau\text{-risk}(\hat{\mu}) = 2.53$

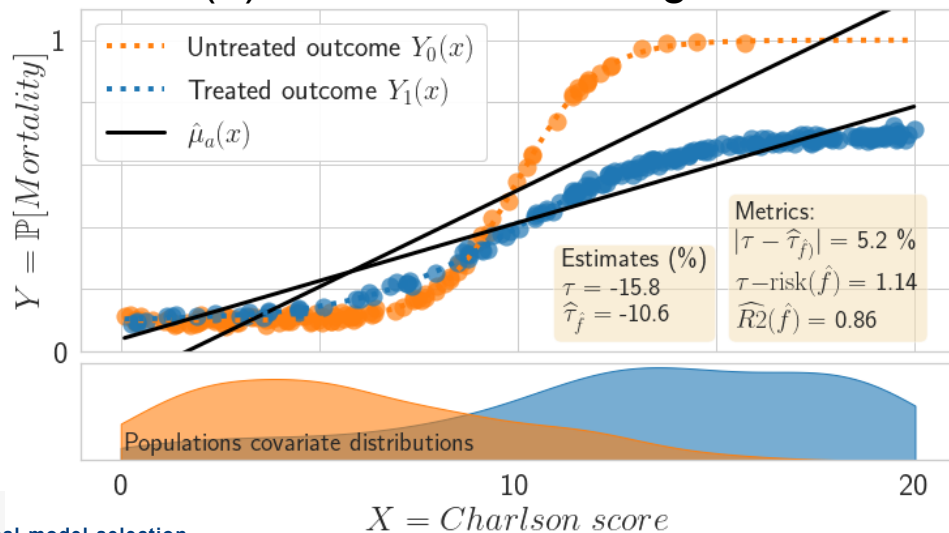
Toy example:

- (a) Random-forest estimator with **high regression performance** (high R2) yielding **poor ATE inference** (large error between true effect tau and predicted tau_f),
- (b) Linear estimator with **smaller regression performance** leading to **better ATE and CATE inference**.

(a) RF with bad ATE inference

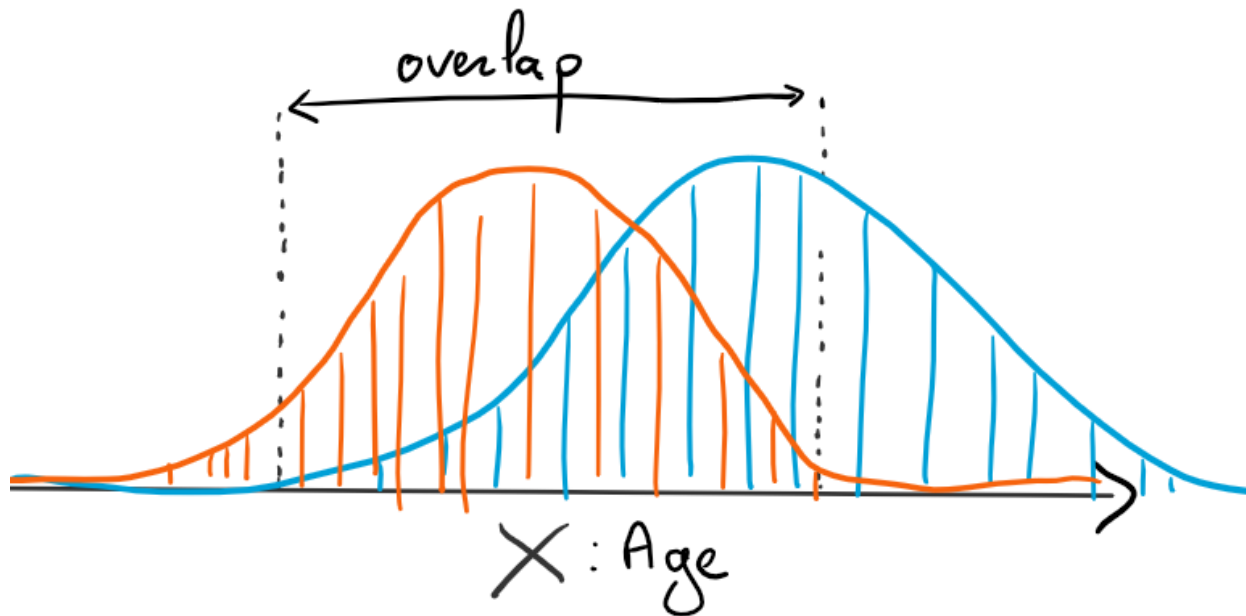


(b) Linear model with good ATE



Causal assumptions: 2 – Positivity (overlap)

Treated and controls should be sufficiently comparable

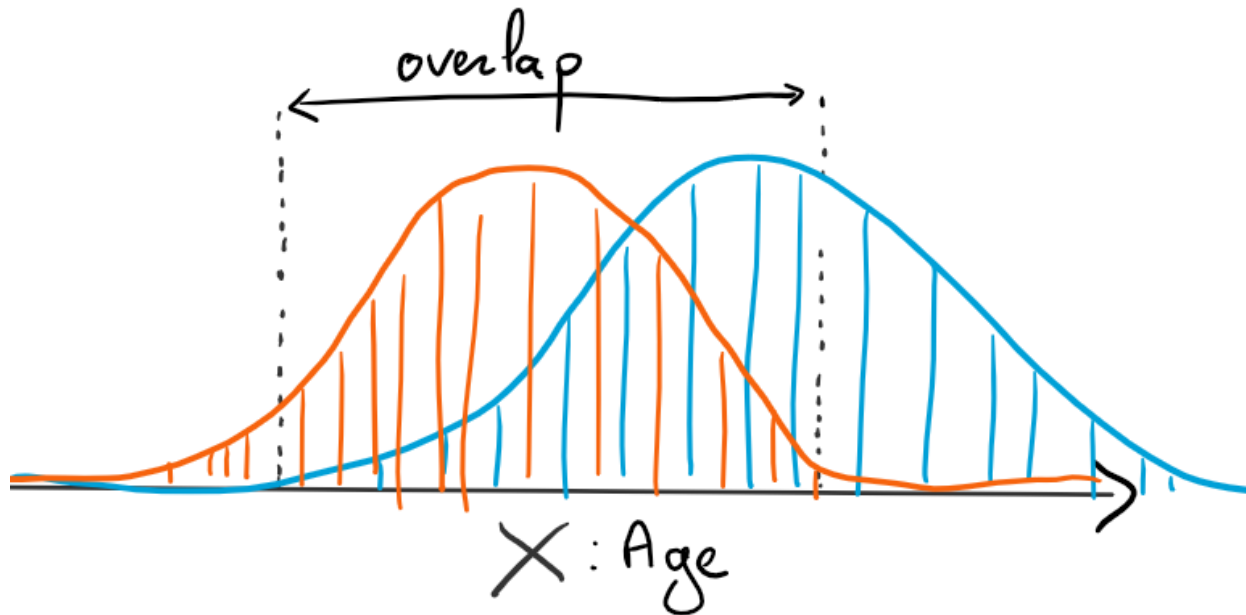




Causal assumptions: 2 – Positivity (overlap)

Given the **Propensity score**, $e(x) = \mathbb{P}_{\mathcal{D}}[A = 1|X = x]$

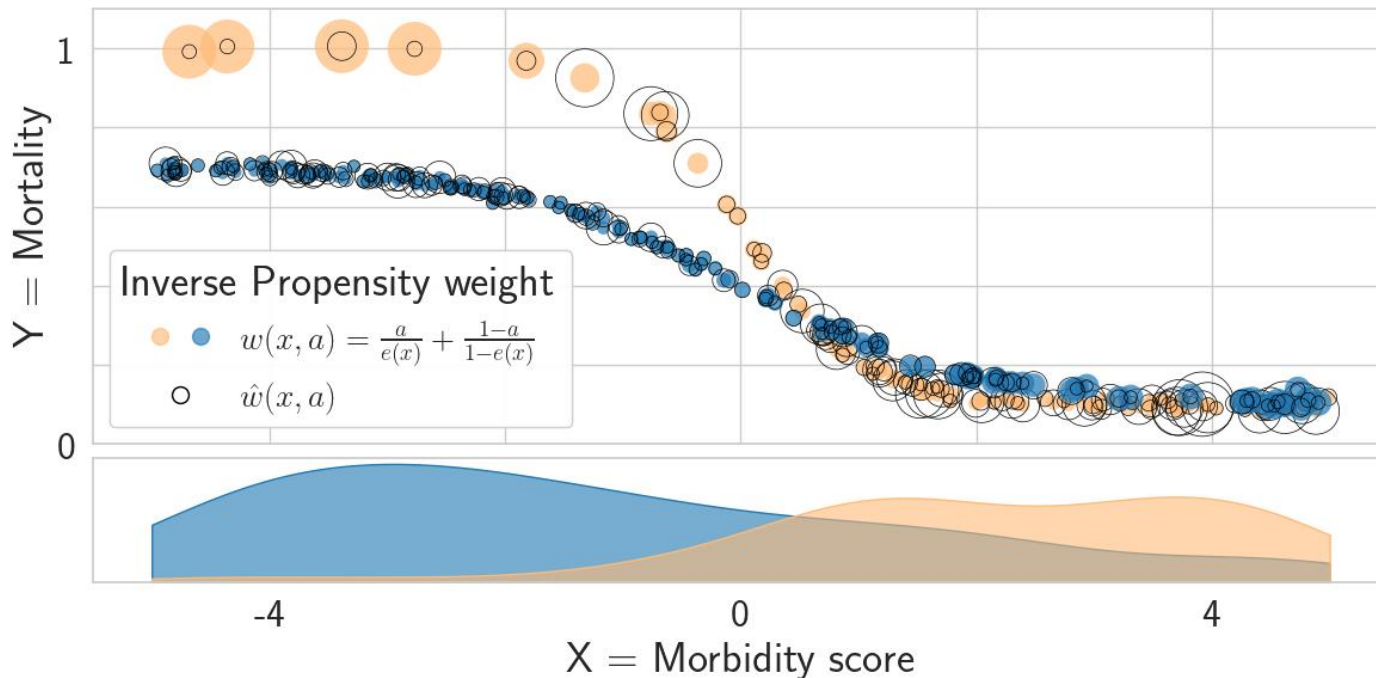
We assume : $\exists \eta > 0, \text{ st, } \eta < e(x) < 1 - \eta \quad \forall x \in \mathcal{X}$





Intervention model

propensity score, reponderation (close, but different from matching)



Estimates (%)


$\tau = -14.18$

$\hat{\tau}_{DM} = 27.63$

$\hat{\tau}_{IPW}^*(e) = -4.22$

$\hat{\tau}_{IPW}(\hat{e}) = -1.33$

Temptative with real data

 **Database:** MIMIC-III (opensource), 67 000 Intense Care Unit hospital stays



Medical question:

What is the effect of **cerebral imagery (A)** on **intra-hospital mortality (Y)** for patients with **stroke related billing diagnoses** ?



Methodological question:

How to choose between two **Average Treatment Effect estimates** ?

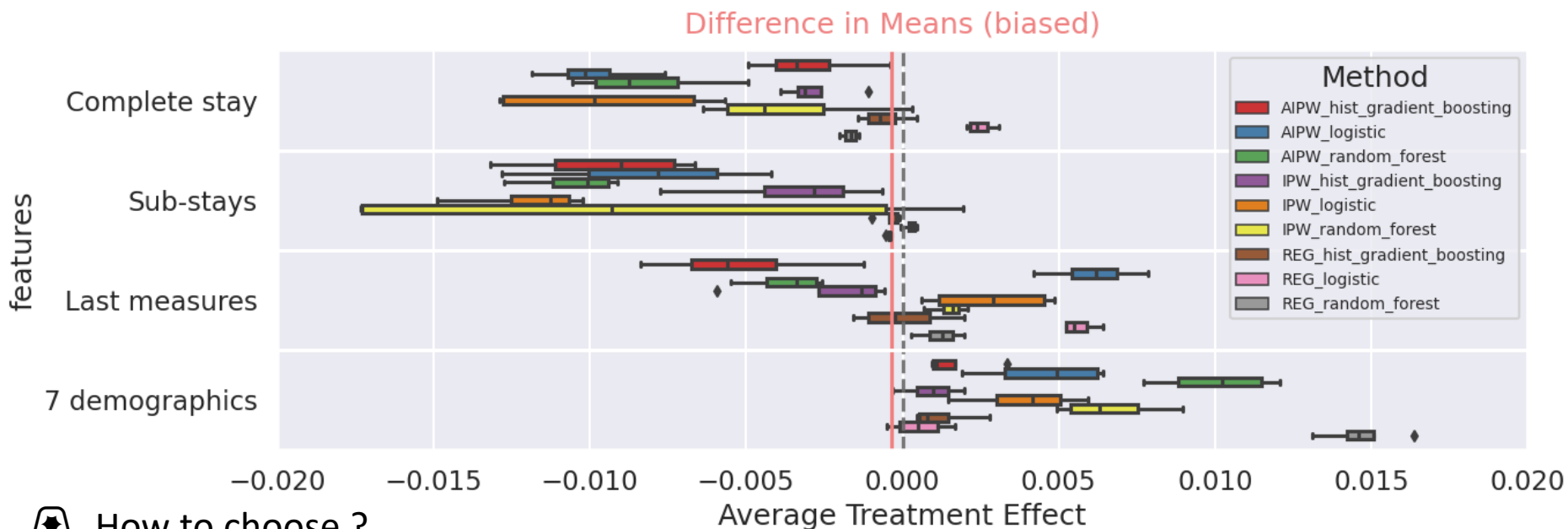
Multiple choices

- **Raw input variables** : baseline, expert selection or 50 most measured
- ✂ **Features representation** : how to flatten the patient covariates ?
- ➡ **Causal estimator** : outcome modeling (g-formula), intervention modeling (reweighting), both (double robust) ?
- ☑ **ML model for outcome and intervention** : logistic, random forest, gradient boosting

Sensitivity Analyse

📁 13 baseline measurements

Average Treatment Effect (ATE) of in-ICU brain imaging on in-hospital mortality for various features representation (n_repetitions=4)



📁 How to choose ?