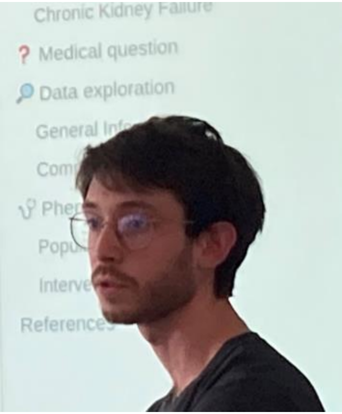
Introduction

Matthieu Doutreligne

Engineer: statistics, computer science, economics, biology

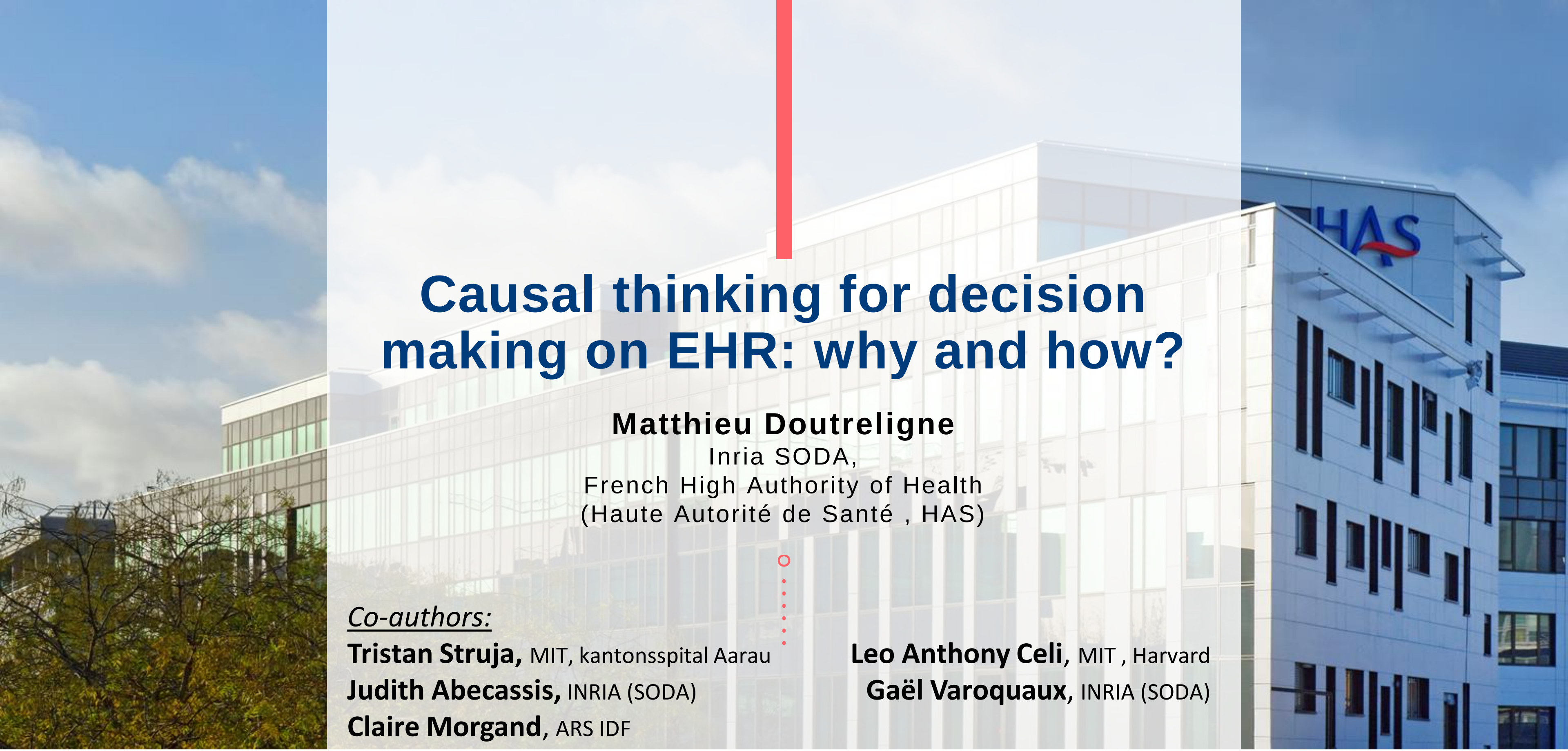
Worked in various health related posts:

- Paris Hospitals (NLP)
- French ministry of health statistical services (claims+Covid)
- Currently :
  - ½ French High Authority of Health (quality of care on EHRs & observational data)
  - ½ 3rd year PhD at Inria in the Social data team: https://team.inria.fr/soda/



"How to do robust and accurate treatment effect estimation from massive routine care data ?"

The PhD in one sentence without any formula



Causal thinking for decision making on EHR: why and how?

Matthieu Doutreligne
Inria SODA,
French High Authority of Health
(Haute Autorité de Santé , HAS)

Co-authors:

Tristan Struja, MIT, kantonsspital Aarau
Judith Abecassis, INRIA (SODA)
Claire Morgand, ARS IDF

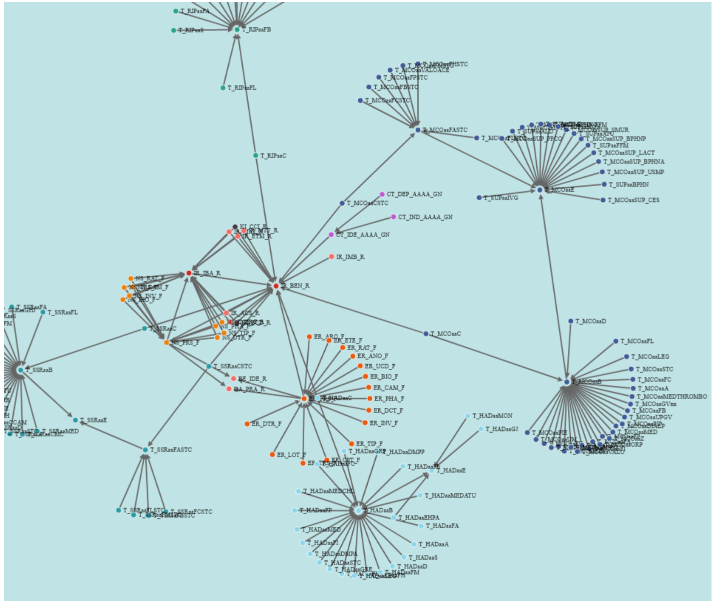
Leo Anthony Celi, MIT , Harvard
Gaël Varoquaux, INRIA (SODA)

I. Motivation

II. Causal framework on EHRs

III. Empirical results

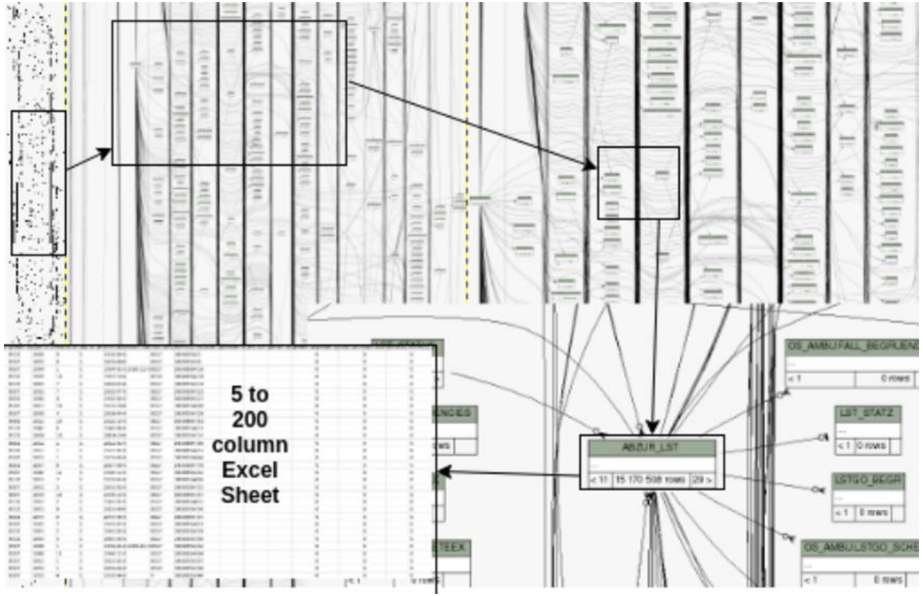
Big healthcare databases with rich data



Claims:

ex. French National Claims, SNDS, 68M patients

Mostly administrative variables eg. billing codes, prescriptions



Electronic Health Records (EHRs):

ex. Paris hospitals, 10M patients

Detailed clinical variables

preprint: https://hal.science/hal-04174834v2/document

👍 Advantages

- **Routine care**
- **Good coverage of the population**
- Cheap data collection

👍 Advantages

- **Routine care**
- **Good coverage of the population**
- Cheap data collection

👎 Difficulties

- **Confounding** (non random interventions)
- **Complexity**
- **Heterogeneous quality**

- **Big data** (statistical and technical difficulties)

Powerfull predictive models

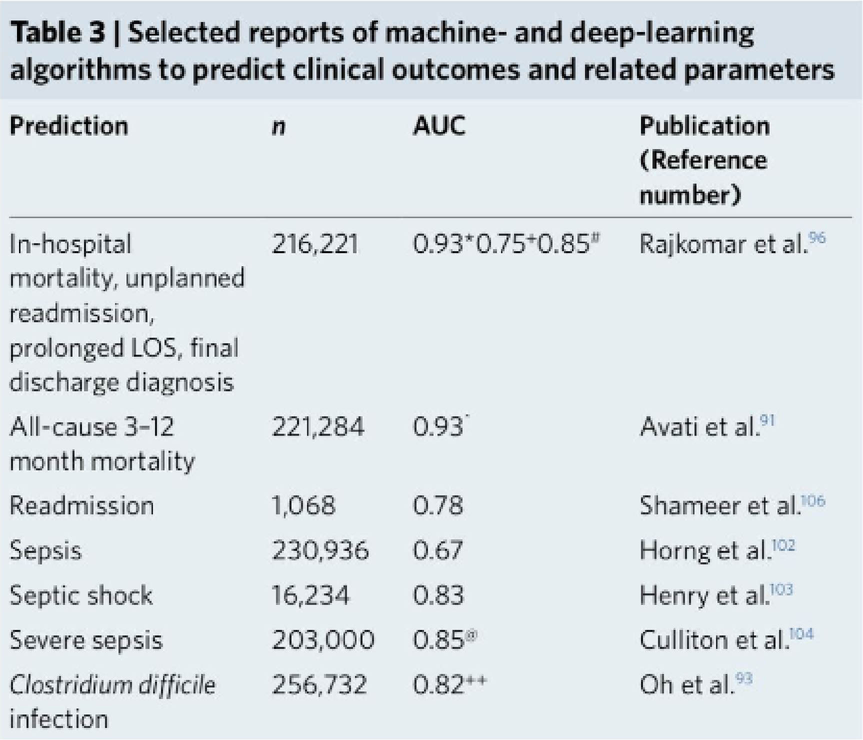
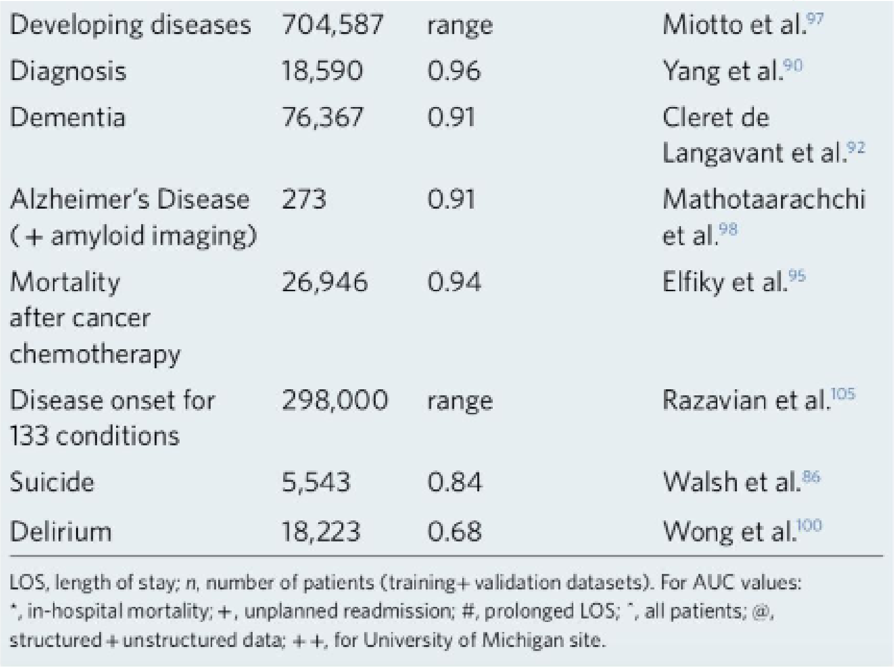



Table 3 | Selected reports of machine- and deep-learning algorithms to predict clinical outcomes and related parameters

Prediction	*n*	AUC	Publication (Reference number)
In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis	216,221	$0.93^{*}0.75^{+}0.85^{\#}$	Rajkomar et al.[96]
All-cause 3–12 month mortality	221,284	$0.93^{\wedge}$	Avati et al.[91]
Readmission	1,068	0.78	Shameer et al.[106]
Sepsis	230,936	0.67	Horng et al.[102]
Septic shock	16,234	0.83	Henry et al.[103]
Severe sepsis	203,000	$0.85^{@}$	Culliton et al.[104]
Clostridium difficile infection	256,732	0.82^{++}	Oh et al.[93]
Developing diseases	704,587	range	Miotto et al.[97]
Diagnosis	18,590	0.96	Yang et al.[90]
Dementia	76,367	0.91	Cleret de Langavant et al.[92]
Alzheimer's Disease (+ amyloid imaging)	273	0.91	Mathotaarachchi et al.[98]
Mortality after cancer chemotherapy	26,946	0.94	Elfiky et al.[95]
Disease onset for 133 conditions	298,000	range	Razavian et al.[105]
Suicide	5,543	0.84	Walsh et al.[86]
Delirium	18,223	0.68	Wong et al.[100]

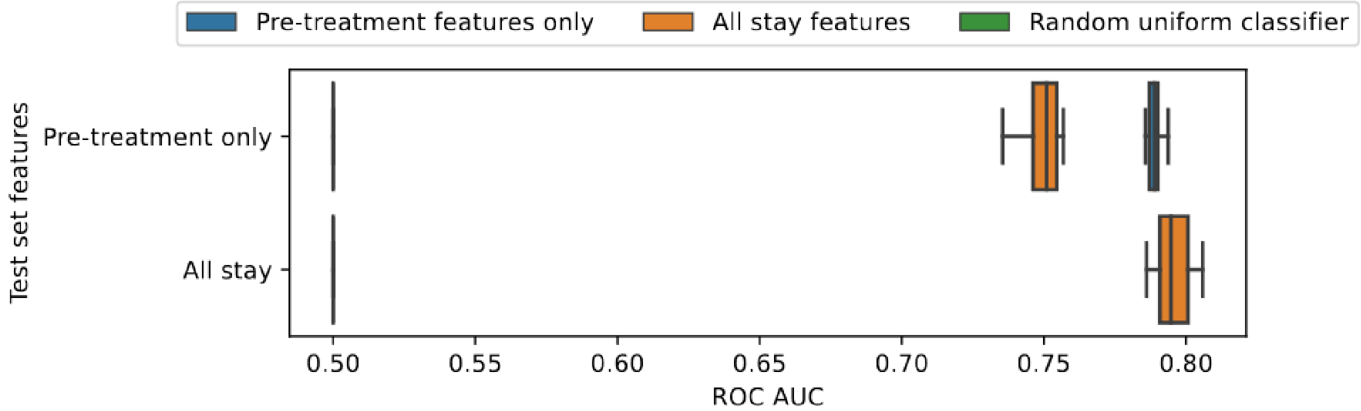
LOS, length of stay; *n*, number of patients (training+ validation datasets). For AUC values: *, in-hospital mortality; +, unplanned readmission; #, prolonged LOS; ^, all patients; @, structured+unstructured data; ++, for University of Michigan site.

Source: **High-performance medicine: the convergence of human and artificial intelligence** Eric Topol, Nature Medicine Jan 2019

So personnalized medicine is solved ? Great !

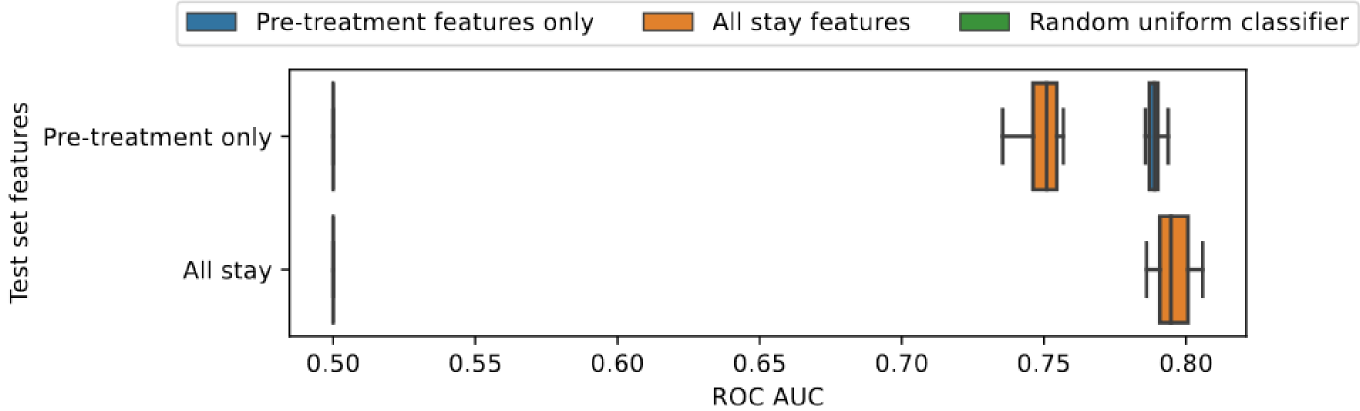
But methodological failure modes: simple example on Mimic

- Predict 28-day mortality, interested in fluid rescusitation treatment
- Train with post-treatment variables
- Evaluate on a clinically useful data set with only pre-treatment variables



But methodological failure modes: simple example on Mimic

- Predict 28-day mortality, interested in fluid rescusitation treatment
- Train with post-treatment variables
- Evaluate on a clinically useful data set with only pre-treatment variables



Who would do that ? 😵‍💫 Answer: A lot of studies !

See: *Yuan, W., Beaulieu-Jones, B. K., Yu, K. H., Lipnick, S. L., Palmer, N., Loscalzo, J., ... & Kohane, I. S. (2021). Temporal bias in case-control design: preventing reliable predictions of the future. Nature communications, 12(1), 1107.*

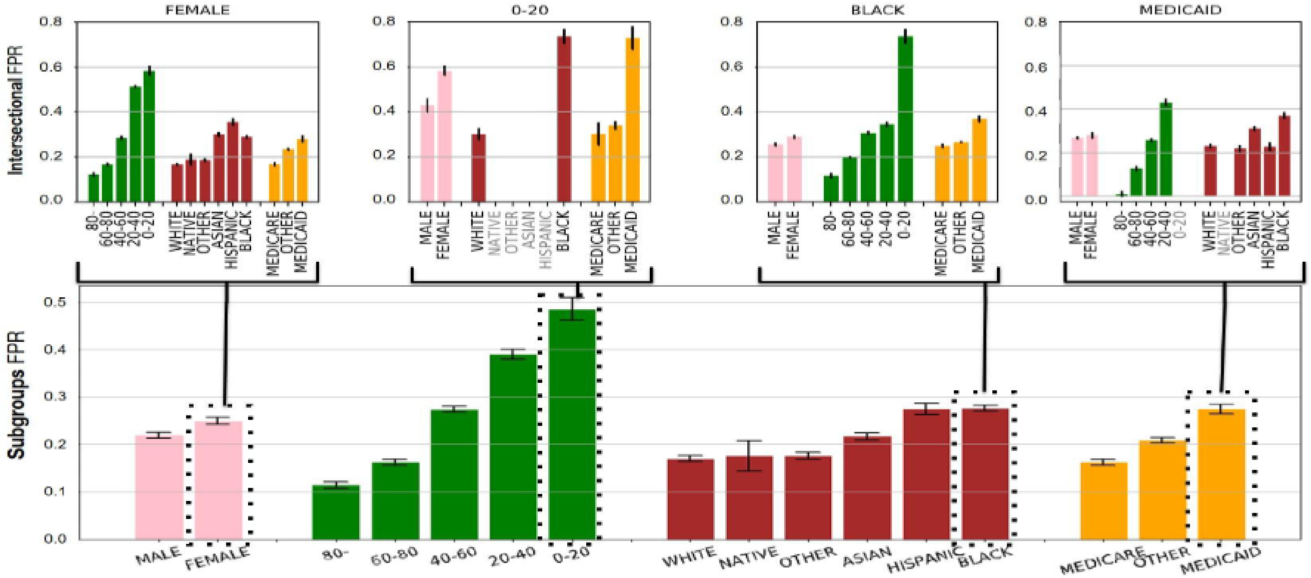
And other failure modes…

eg. Exclusion of under-served populations for chest X-ray diagnosis

Largest underdiagnosis rates in:

- Female
- 0-20
- Black
- Medicaid insurance

Automating Che**Xclusion** With **EHR + ML**



Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi.
“Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations” Nature Medicine 2021.

preprint: https://hal.science/hal-04174834v2/document

These failures occur because of shortcut features

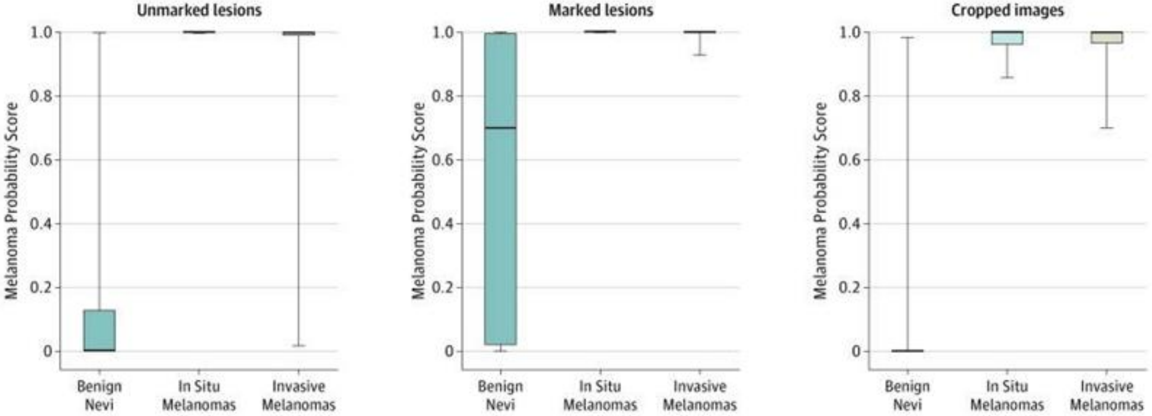
Prediction: malignent melanoma

Intervention: excision of nevi

Shortcut: surgical marks

Begign nevi

CNN predicted score

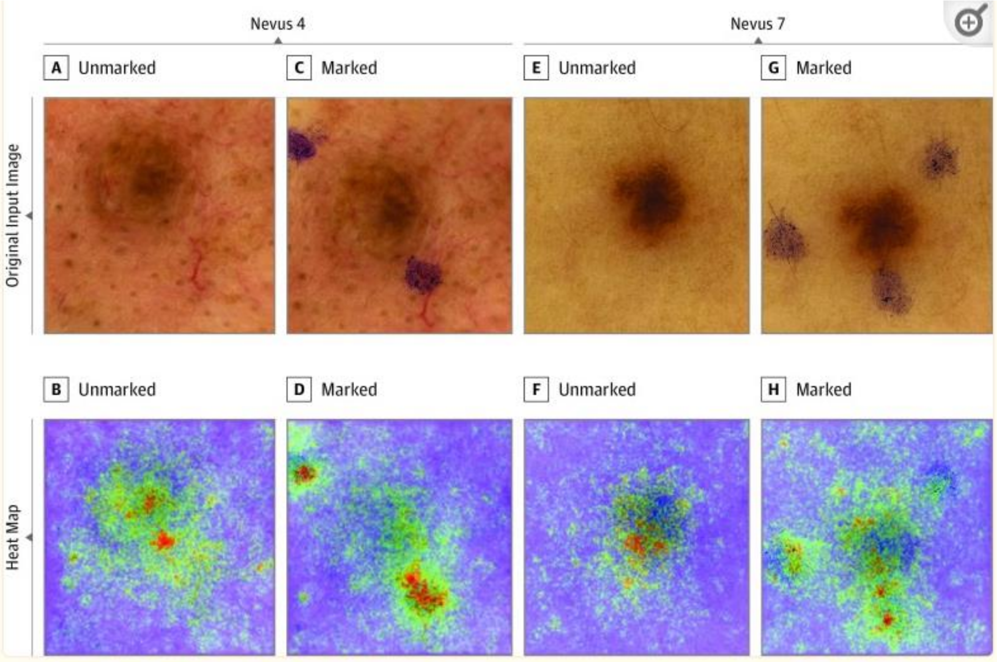


True labels

Winkler, Fink, Toberer, Enk, Deinlein, Hofmann-Wellenhof, Thomas, Lallas, Blum, Stolz, et al. (2019). "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition". In: JAMA dermatology

These failures occur because of shortcut features

Begign nevi



Prediction: malignent melanoma
Intervention: excision of nevi
Shortcut: surgical marks

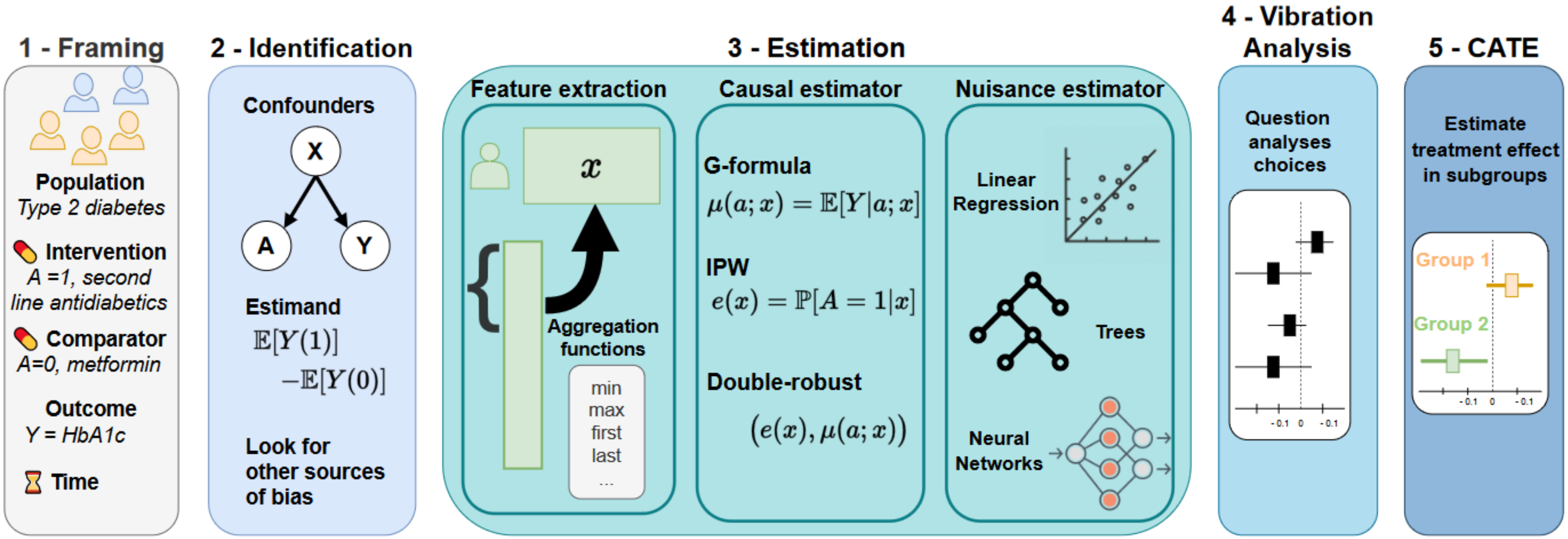
Winkler, Fink, Toberer, Enk, Deinlein, Hofmann-Wellenhof, Thomas, Lallas, Blum, Stolz, et al. (2019). "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition". In: JAMA dermatology

I. Motivation

II. Causal framework on EHRs

III. Empirical results

Causal Framework with EHRs



1 - Framing

Population

Type 2 diabetes

Intervention

A =1, second line antidiabetics

Comparator

A=0, metformin

Outcome

Y = HbA1c

Time

2 - Identification

Confounders

Estimand

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

Look for other sources of bias

3 - Estimation

Feature extraction

Aggregation functions

min
max
first
last
...

Causal estimator

G-formula

$\mu(a;x) = \mathbb{E}[Y|a;x]$

IPW

$e(x) = \mathbb{P}[A=1|x]$

Double-robust

$(e(x), \mu(a;x))$

Nuisance estimator

Linear Regression

Trees

Neural Networks

4 - Vibration Analysis

Question analyses choices

5 - CATE

Estimate treatment effect in subgroups

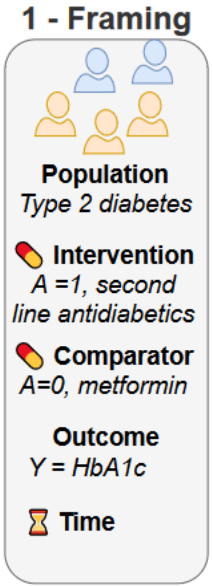
HAS HAUTE AUTORITÉ DE SANTÉ

Inria

Soda

preprint: https://hal.science/hal-04174834v2/document

Causal Framework: Study design



1 - Framing

Population
Type 2 diabetes

Intervention
A =1, second line antidiabetics

Comparator
A=0, metformin

Outcome
Y = HbA1c

Time

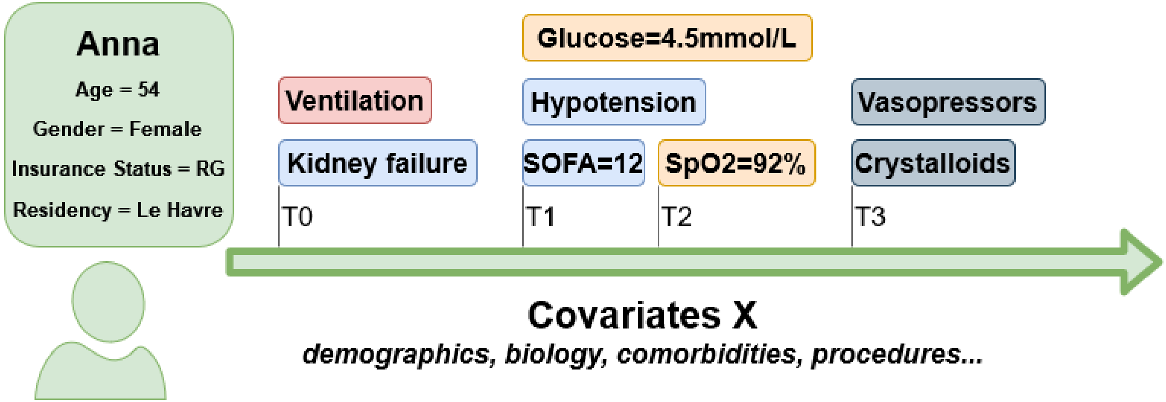
Emulate the **ideal trial** that you would conduct if you could recruit the patients.

Hernan, Miguel A (2021). “Methods of public health research–strengthening causal inference from observational data”. In: New England Journal of Medicine

Study design – Frame the question to avoid biases

Target Population with features X

Eg. Patients with sepsis in the ICU



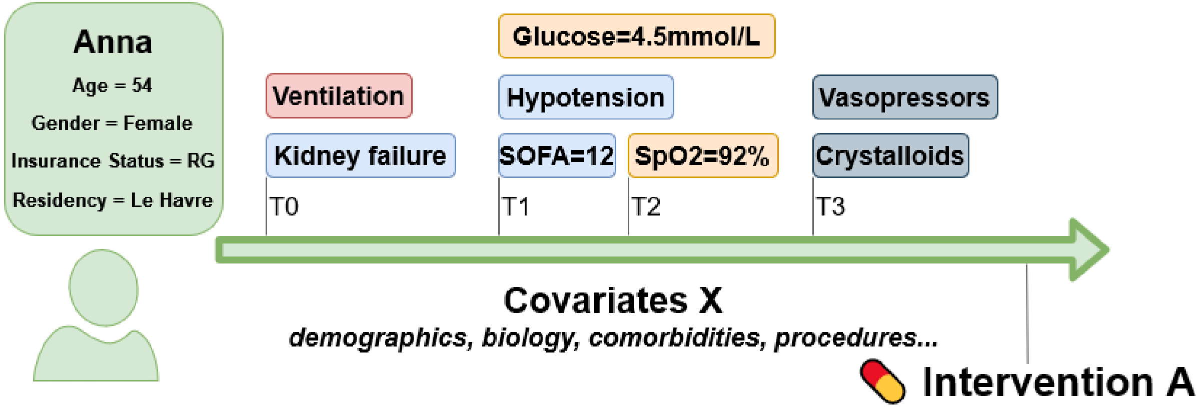
Covariates X

demographics, biology, comorbidities, procedures…

Study design – Frame the question to avoid biases

For whome, we consider giving
treament A=1 or control A=0

Eg. Combination of crystalloids and albumin or Crystalloids only

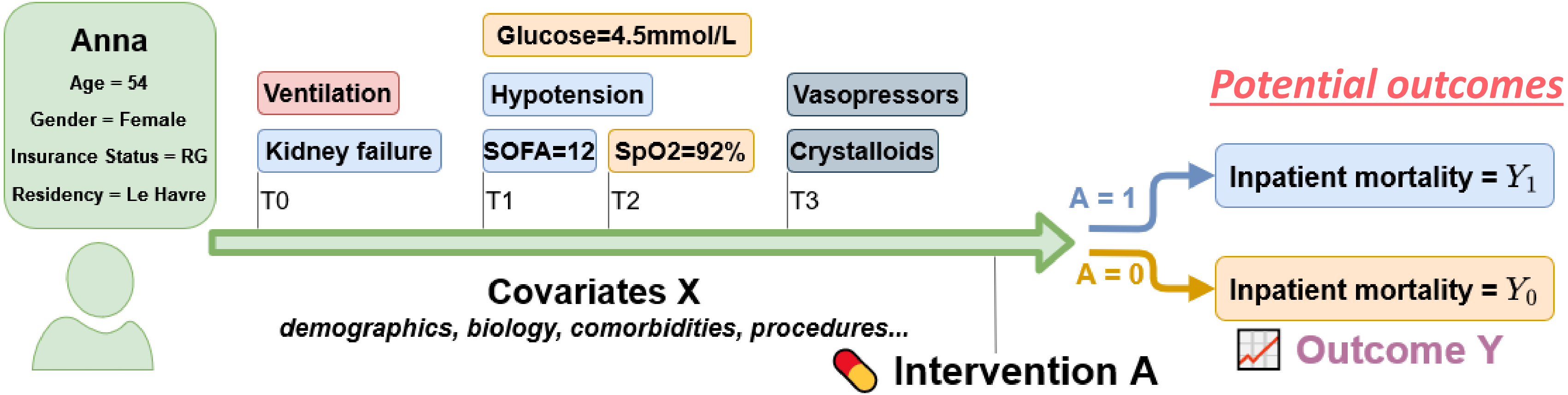


preprint: https://hal.science/hal-04174834v2/document

Study design – Frame the question to avoid biases

To improve a **clinical outcome Y**

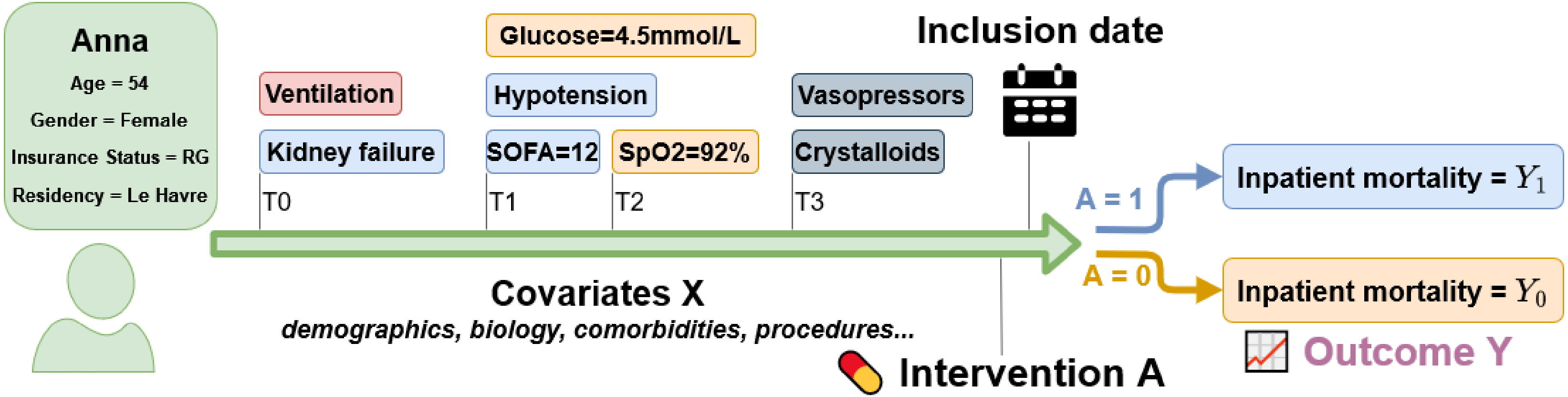
Eg. 28-day survival



Study design – Frame the question to avoid biases

Following patients during a **specific time-period**

Eg. During 24 first hours of hospitalization



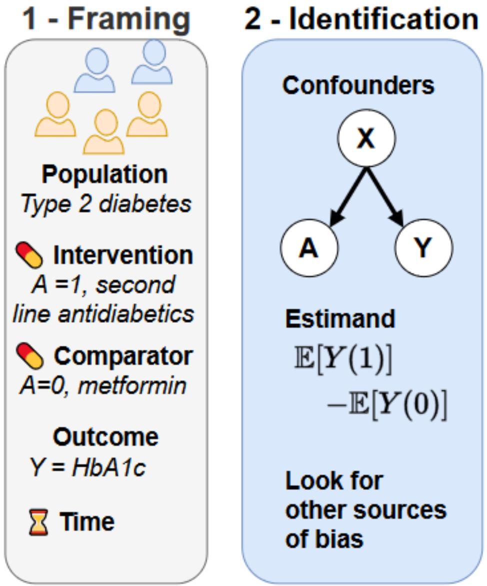
preprint: https://hal.science/hal-04174834v2/document

Study design – Frame the question to avoid biases

	Example (Mimic database usecase)
Target Population with features X	*Patients with sepsis in the ICU*
For whome, we consider giving **the treament A=1 or the control A=0**	*Combination of crystalloids and albumin or Crystalloids only*
To improve a **clinical outcome Y**	*28-day survival*
Following patients during a **specific time-period**	*During 24 first hours of hospitalization*

? **Contrast** **the intervention** against **the control** on **the outcome in the target population**

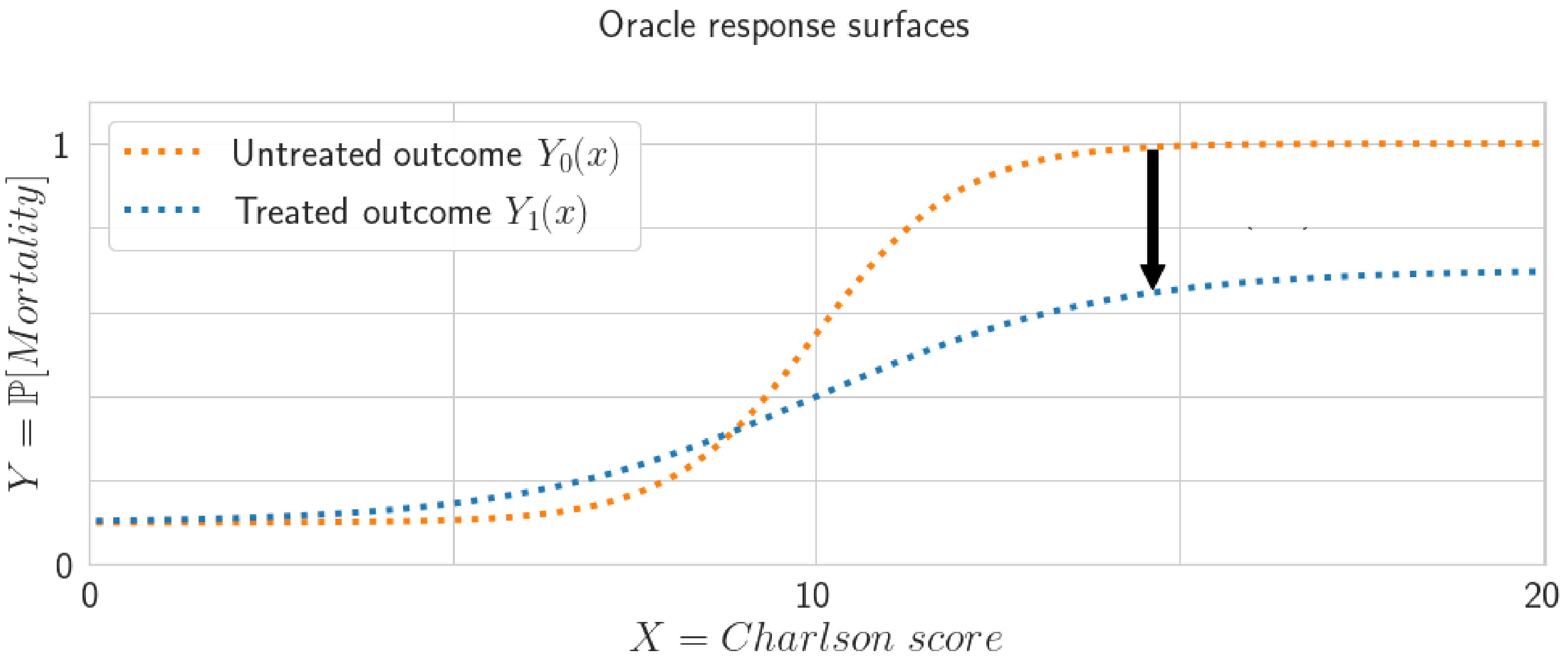
Causal Framework in real life : Identification



List necessary information to answer the causal question

VanderWeele, Tyler J (2019). "Principles of confounder selection". In: European journal of epidemiology

1D example



preprint: https://hal.science/hal-04174834v2/document

1D example

Estimate one of:

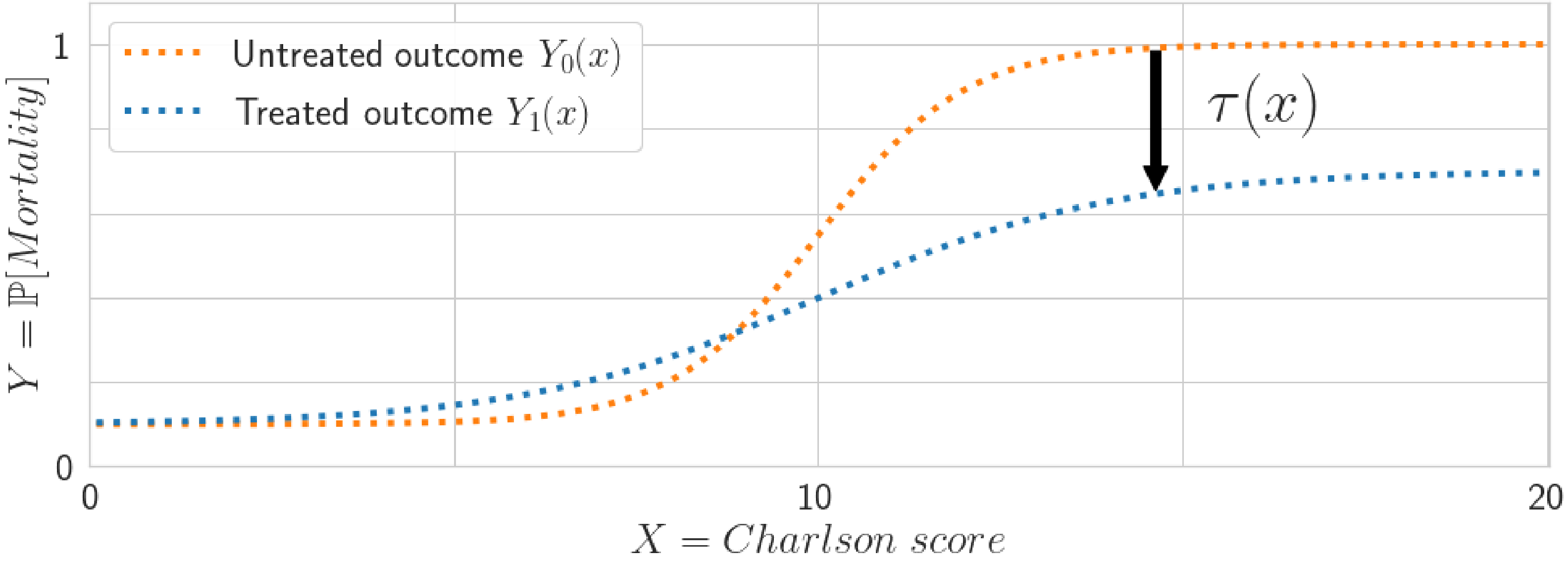
- Average Treatment Effect (ATE)

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

- Conditional Average Treatment Effect (CATE)

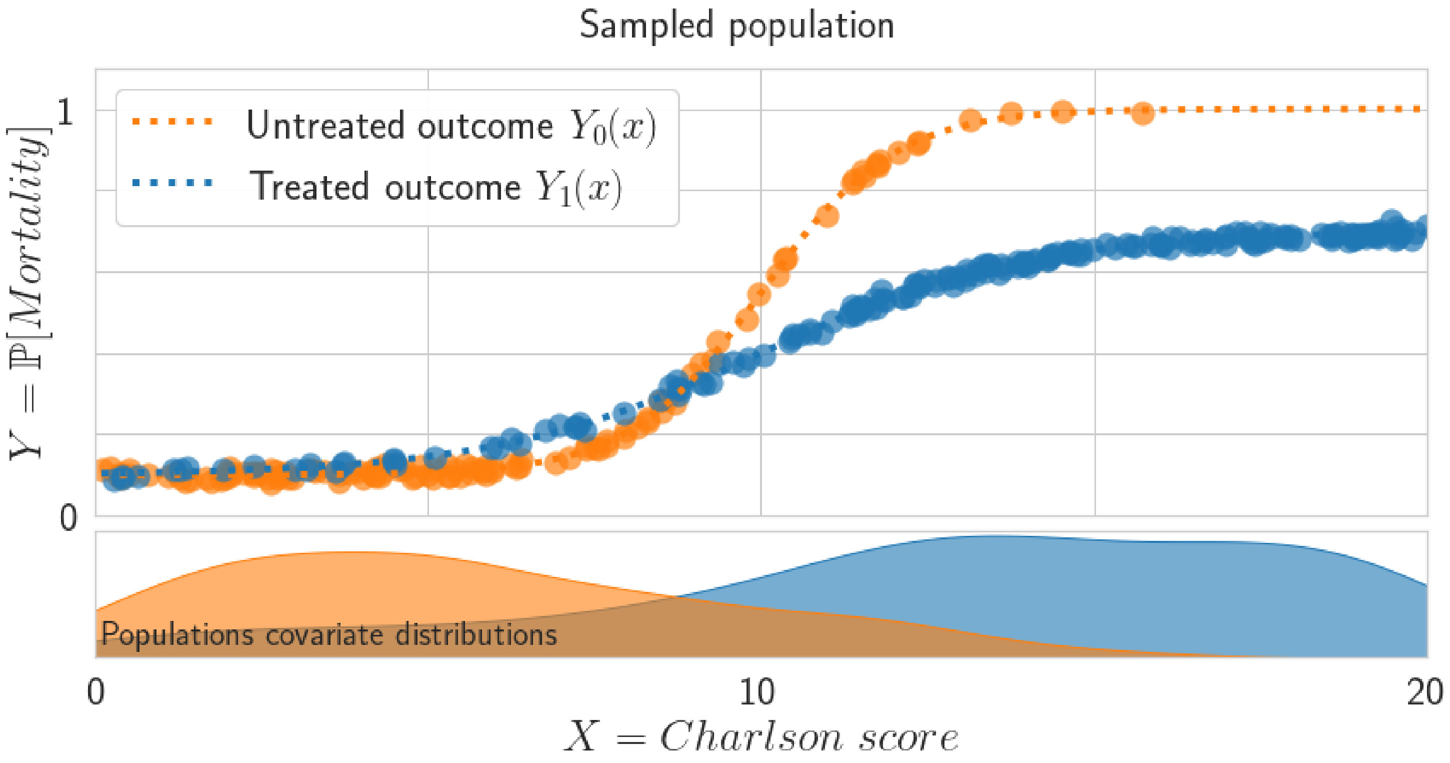
$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Oracle response surfaces



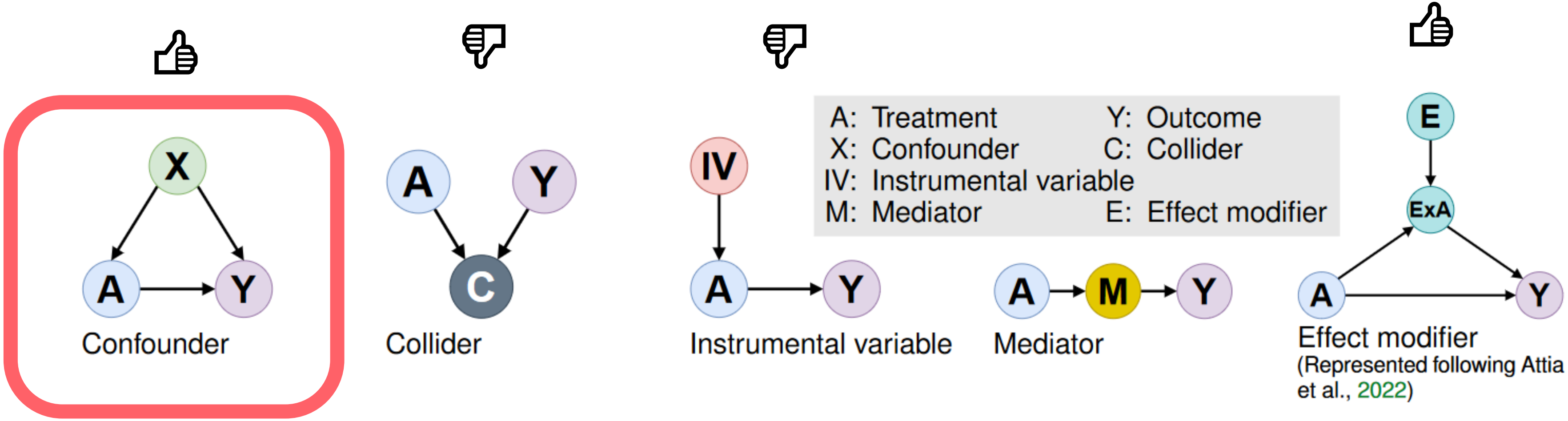
preprint: https://hal.science/hal-04174834v2/document

1D example


Sampled population
Untreated outcome $Y_0(x)$
Treated outcome $Y_1(x)$
$Y = \mathbb{P}[Mortality]$
1
0
Populations covariate distributions
0
10
20
$X = Charlson\ score$

Identification - List necessary information to answer the causal question

Categorize variables in the data base

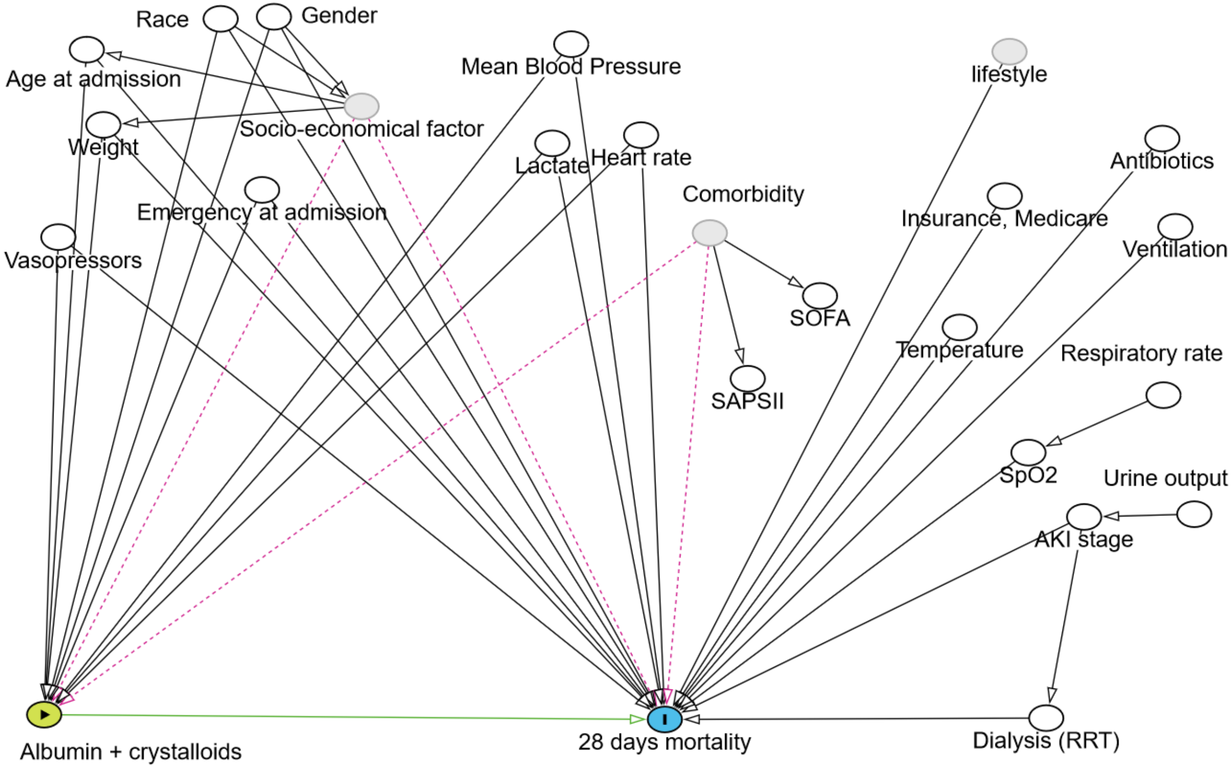


Focus on confounding

Identification - List necessary information to answer the causal question

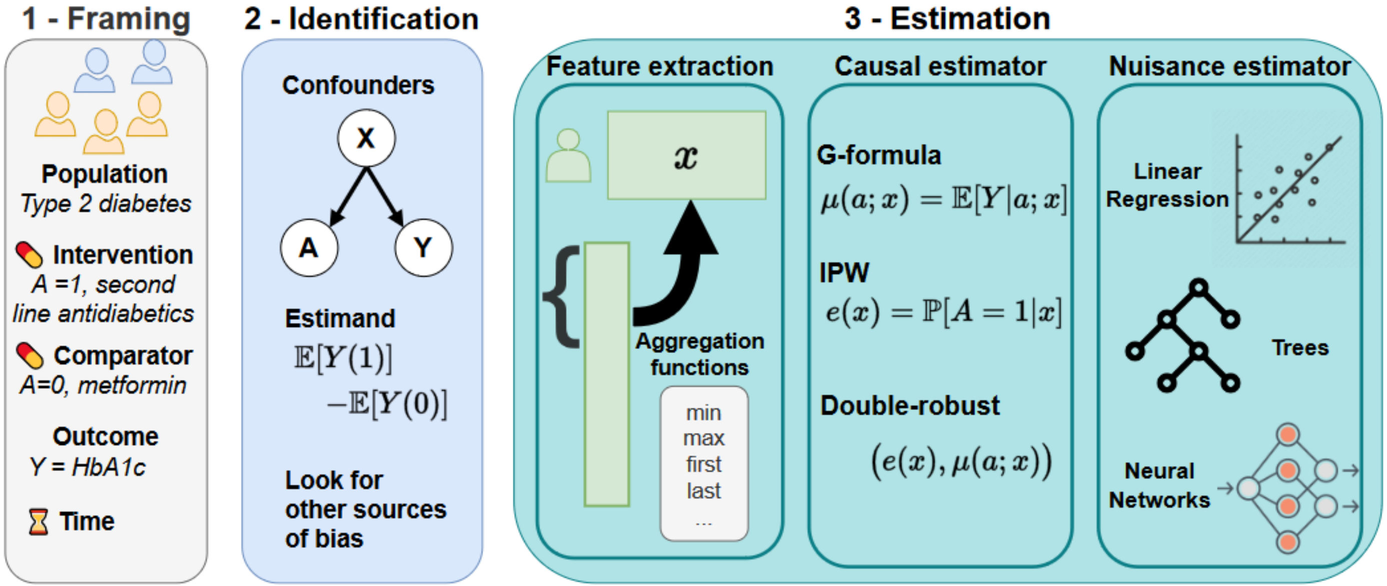
Causal graph to list confounders
(we used daggity)

Red arrows point to **missing confounders** that we hope to **control with proxies**



preprint: https://hal.science/hal-04174834v2/document

Causal Framework: Estimation



1 - Framing

Population
Type 2 diabetes

Intervention
A =1, second line antidiabetics

Comparator
A=0, metformin

Outcome
Y = HbA1c

Time

2 - Identification

Confounders

Estimand

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

Look for other sources of bias

3 - Estimation

Feature extraction

x

Aggregation functions

min
max
first
last
...

Causal estimator

G-formula

$$\mu(a;x) = \mathbb{E}[Y|a;x]$$

IPW

$$e(x) = \mathbb{P}[A=1|x]$$

Double-robust

$$(e(x), \mu(a;x))$$

Nuisance estimator

Linear Regression

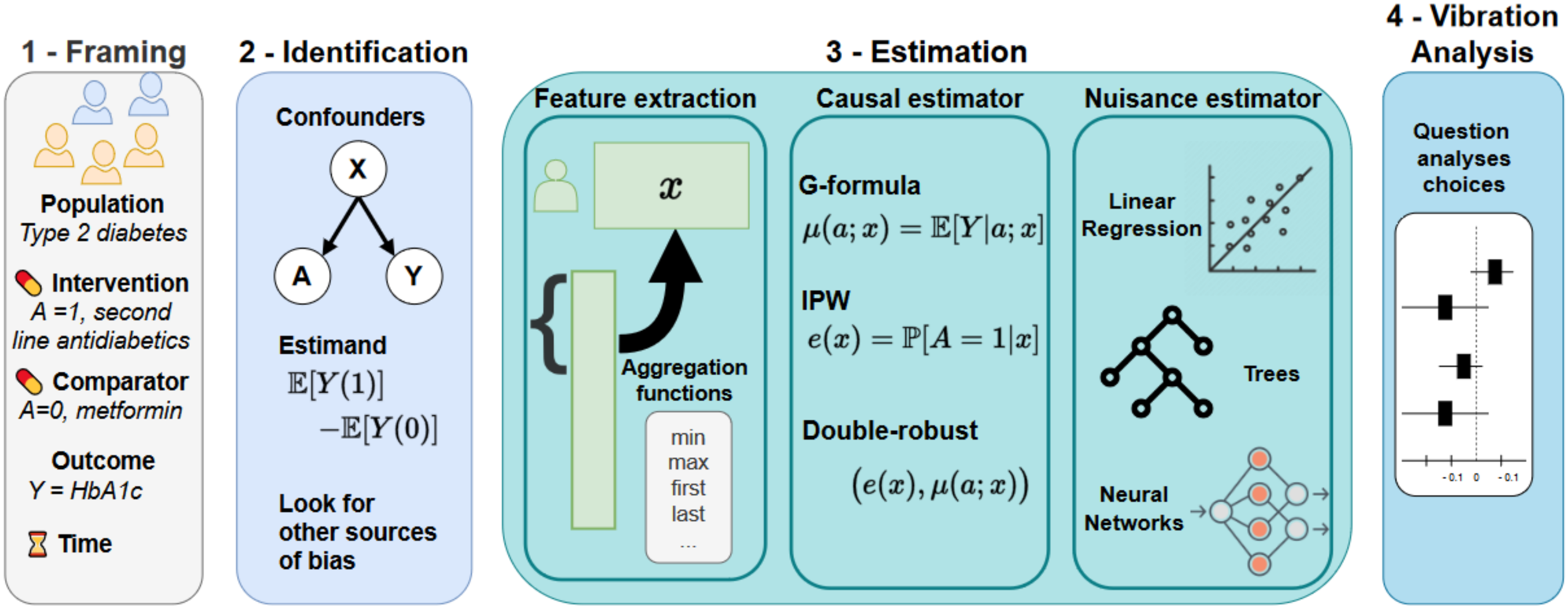
Trees

Neural Networks

Select appropriate estimators

Wager, Stefan (2020). Stats 361: Causal inference.

Causal Framework: Vibration analysis



1 - Framing

Population
Type 2 diabetes

Intervention
A =1, second line antidiabetics

Comparator
A=0, metformin

Outcome
Y = HbA1c

Time

2 - Identification

Confounders

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

Look for other sources of bias

3 - Estimation

Feature extraction

Aggregation functions

min
max
first
last
...

Causal estimator

G-formula
$\mu(a;x) = \mathbb{E}[Y|a;x]$

IPW
$e(x) = \mathbb{P}[A=1|x]$

Double-robust
$(e(x), \mu(a;x))$

Nuisance estimator

Linear Regression

Trees

Neural Networks

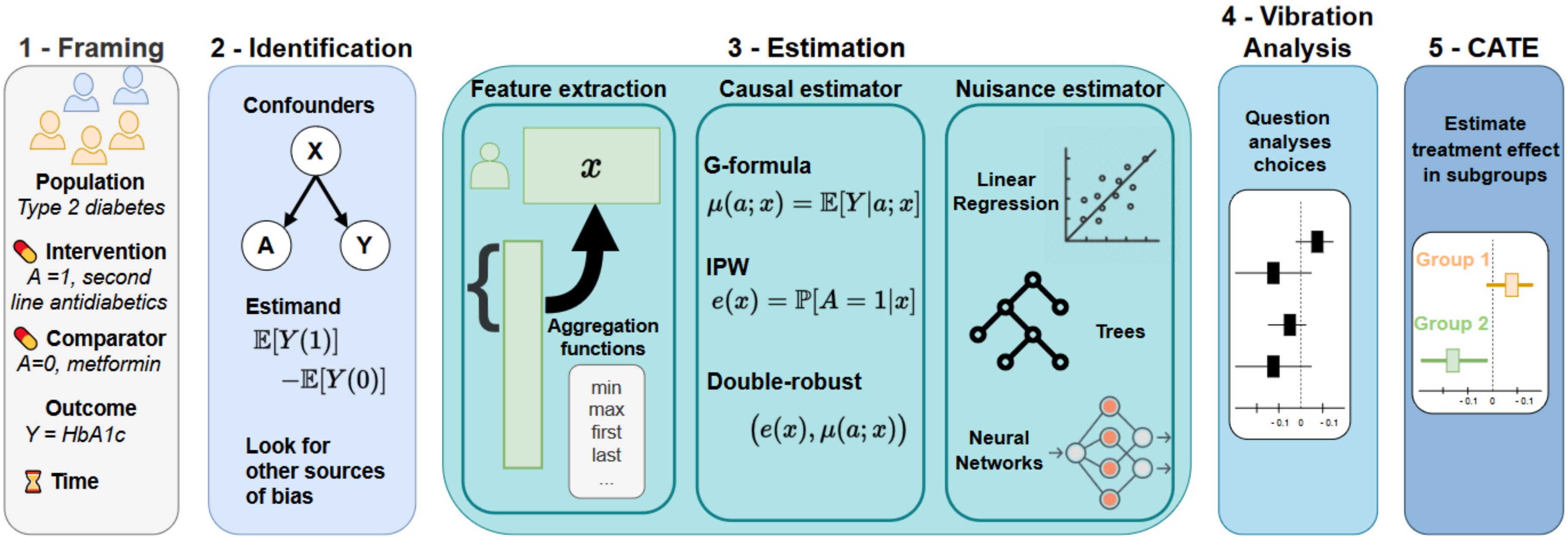
4 - Vibration Analysis

Question analyses choices

Assess the robustness of the hypotheses

Patel, Burford, and Ioannidis (2015). “Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations”. In: Journal of clinical epidemiology

Causal Framework: Treatment heterogeneity



1 - Framing

Population
Type 2 diabetes

Intervention
A =1, second line antidiabetics

Comparator
A=0, metformin

Outcome
Y = HbA1c

Time

2 - Identification

Confounders

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

Look for other sources of bias

3 - Estimation

Feature extraction

Aggregation functions

Causal estimator

G-formula
$\mu(a; x) = \mathbb{E}[Y|a; x]$

IPW
$e(x) = \mathbb{P}[A = 1|x]$

Double-robust
$(e(x), \mu(a; x))$

Nuisance estimator

4 - Vibration Analysis

Question analyses choices

5 - CATE

Estimate treatment effect in subgroups

Compute treatment effects on subpopulations

Robertson, Sarah E, Andrew Leith, Christopher H Schmid, and Issa J Dahabreh (2021). "Assessing heterogeneity of treatment effects in observational studies". In: American Journal of Epidemiology

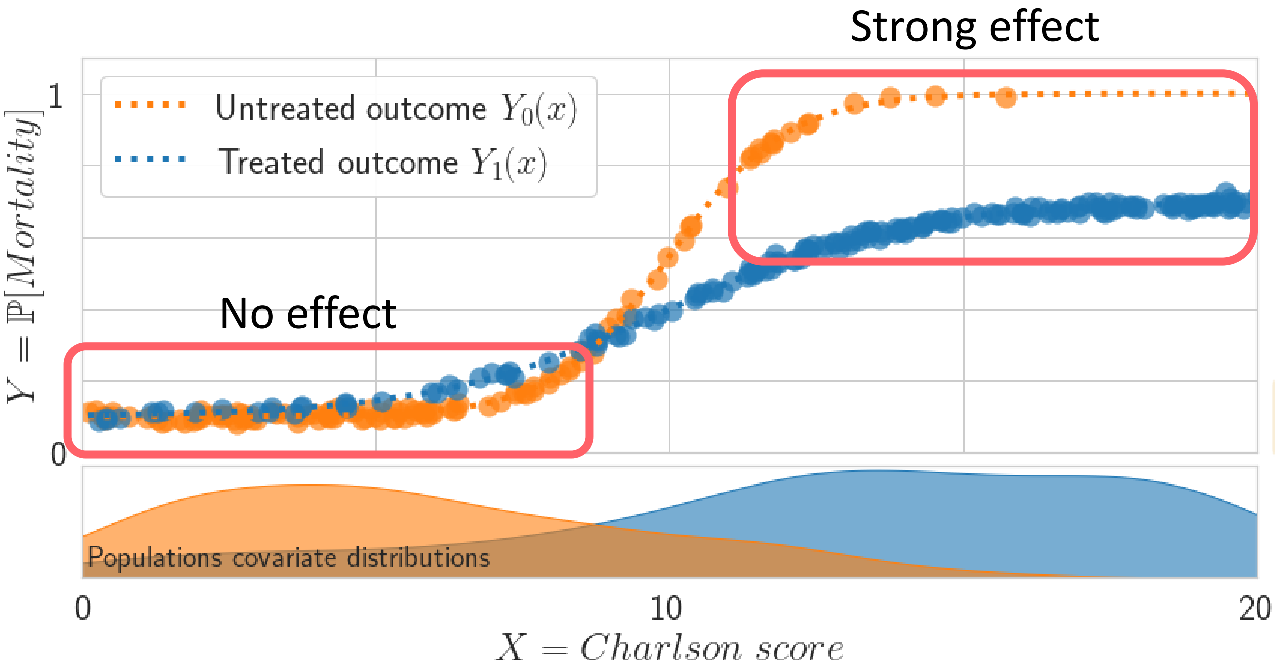
Treatment heterogeneity – Compute treatment effects on subpopulations

Does the effect varies in different subpopulations?

If yes, there is room for personalized treatment !

How to do that ?

- Take the most reliable estimate from previous steps.
- Regress the individual estimations against targeted sources heterogeneity.



Causal thinking for decision making on EHR: why and how?

I. Motivation

II. Causal framework on EHRs

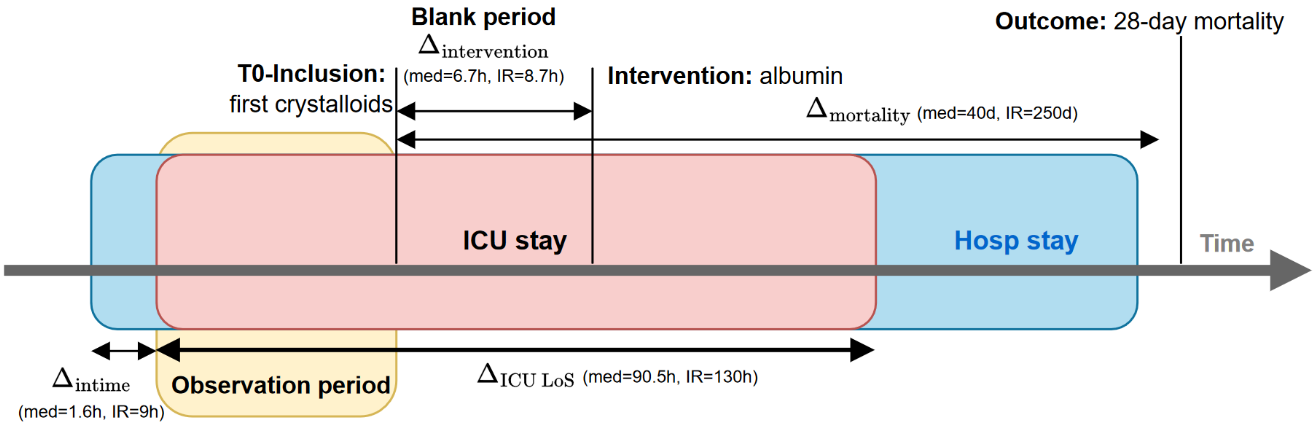
III. Empirical results

Let’s run some inference (Mimic-IV)

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

Medical question: What is the effect of albumin in combination with crystalloids compared to crystalloids alone on 28-day mortality in patients with sepsis?

Cohort: 3,559 treated and 14,862 controls.


Blank period
$\Delta_{\text{intervention}}$
(med=6.7h, IR=8.7h)
Outcome: 28-day mortality
T0-Inclusion: first crystalloids
Intervention: albumin
$\Delta_{\text{mortality}}$ (med=40d, IR=250d)
ICU stay
Hosp stay
Time
Δ_{intime}
(med=1.6h, IR=9h)
Observation period
$\Delta_{\text{ICU LoS}}$ (med=90.5h, IR=130h)

preprint: https://hal.science/hal-04174834v2/document

Let’s run some inference (Mimic-IV)

😵 **Estimation choices:**

🛠 **Feature aggregations:**

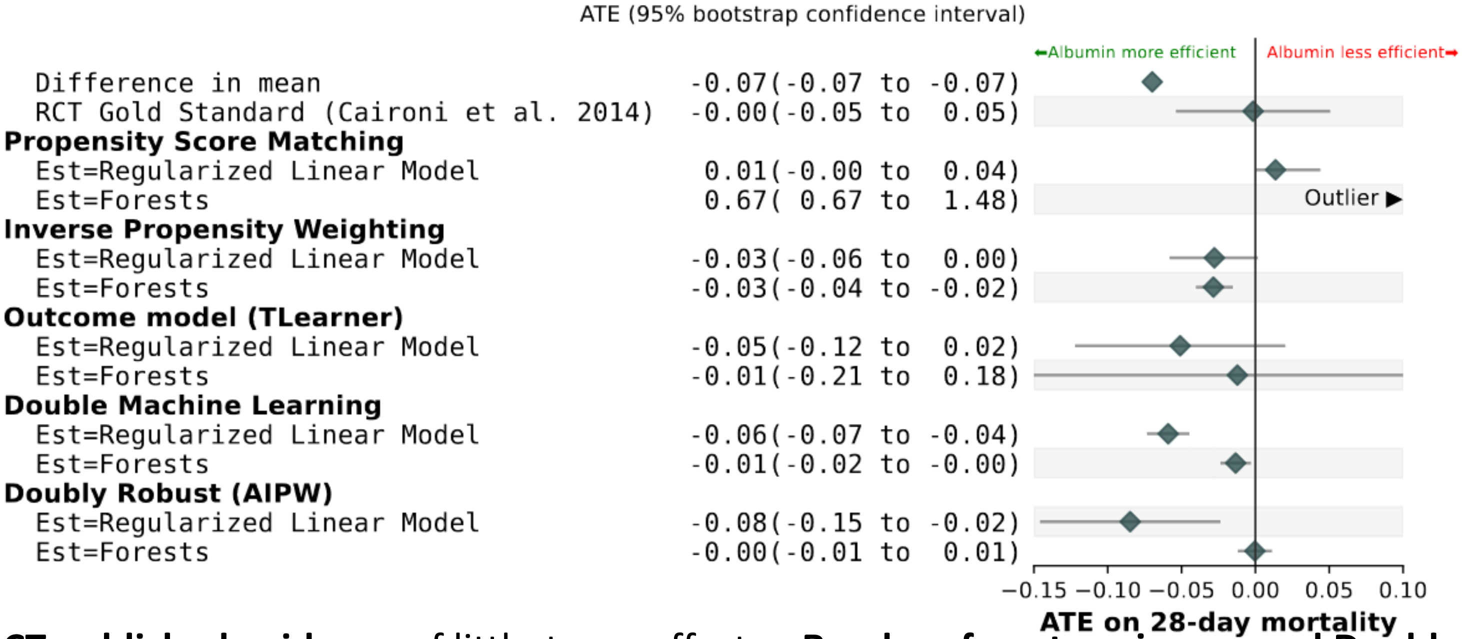
- Last value before the start of the follow-up period,
- First observed value,
- Both the first and last values as concatenated features.

📈 **Causal estimators:** Inverse Propensity Weighting (IPW), outcome modeling (G-formula) with T-Learner, Augmented Inverse Propensity Weighting (AIPW) and Double Machine Learning (DML).

⚙ **Outcome and treatment estimators:** regularized logistic regression and random forest

Let’s run some inference (Mimic-IV)

Aggregation: first and last pre-treatment measures



	ATE (95% bootstrap confidence interval)
Difference in mean	-0.07(-0.07 to -0.07)
RCT Gold Standard (Caironi et al. 2014)	-0.00(-0.05 to 0.05)
Propensity Score Matching	
Est=Regularized Linear Model	0.01(-0.00 to 0.04)
Est=Forests	0.67(0.67 to 1.48)
Inverse Propensity Weighting	
Est=Regularized Linear Model	-0.03(-0.06 to 0.00)
Est=Forests	-0.03(-0.04 to -0.02)
Outcome model (TLearner)	
Est=Regularized Linear Model	-0.05(-0.12 to 0.02)
Est=Forests	-0.01(-0.21 to 0.18)
Double Machine Learning	
Est=Regularized Linear Model	-0.06(-0.07 to -0.04)
Est=Forests	-0.01(-0.02 to -0.00)
Doubly Robust (AIPW)	
Est=Regularized Linear Model	-0.08(-0.15 to -0.02)
Est=Forests	-0.00(-0.01 to 0.01)

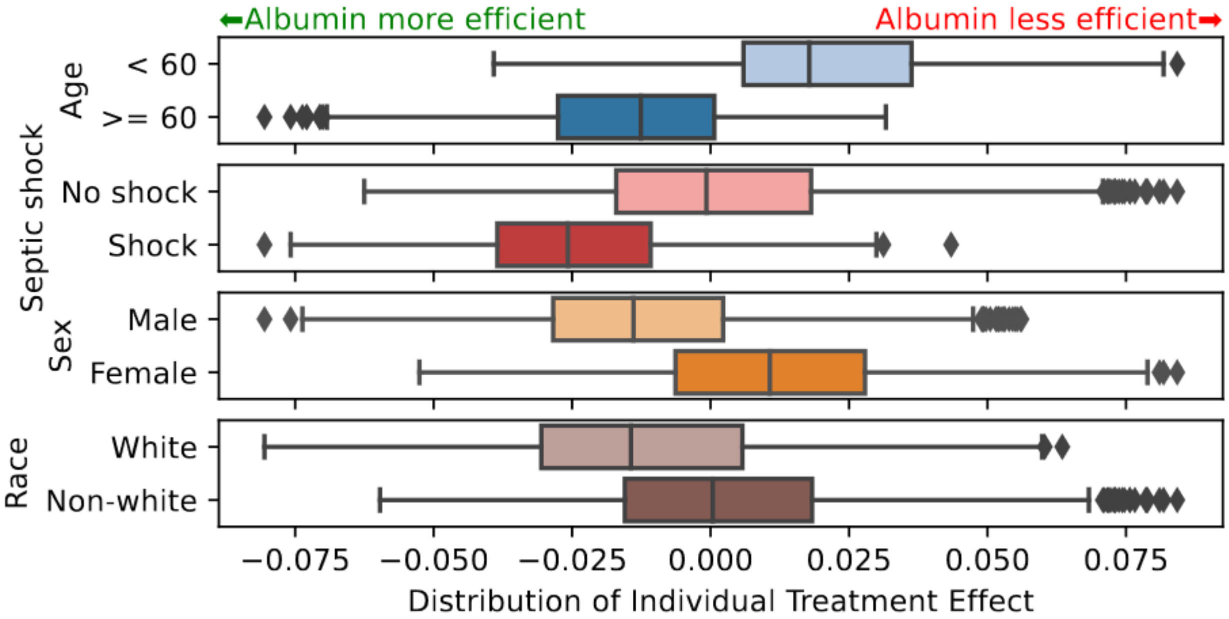
Recover RCT published evidence of little-to-no effect -> **Random forests nuisance and Double ML or AIPW**

Li, Binghu, Hongliang Zhao, Jie Zhang, Qingguang Yan, Tao Li, and Liangming Liu (2020). “Resuscitation fluids inseptic shock: a network meta-analysis of randomized controlled trials”. In: Shock

HAUTE AUTORITÉ DE SANTÉ

preprint: https://hal.science/hal-04174834v2/document

Heterogeneity of Treatment Effect

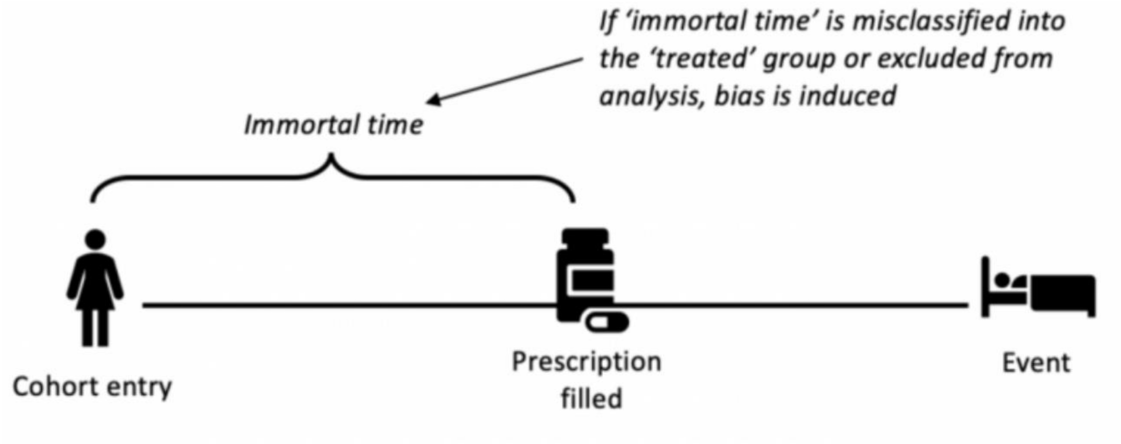


Recover RCT post-hoc subgroup analysis: increasing treatment effect (relative risk) for patients with septic shock: RR=**0.87**; 95% CI, 0.77 to 0.99 vs **1.13**;95% CI, 0.92 to 1.39

Caironi, Pietro, Gianni Tognoni, Serge Masson, Roberto Fumagalli, Antonio Pesenti, Marilena Romero, CaterinaFanizza, Luisa Caspani, Stefano Faenza, Giacomo Grasselli, et al. (2014). "Albumin replacement in patients withsevere sepsis or septic shock". In: New England Journal of Medicine

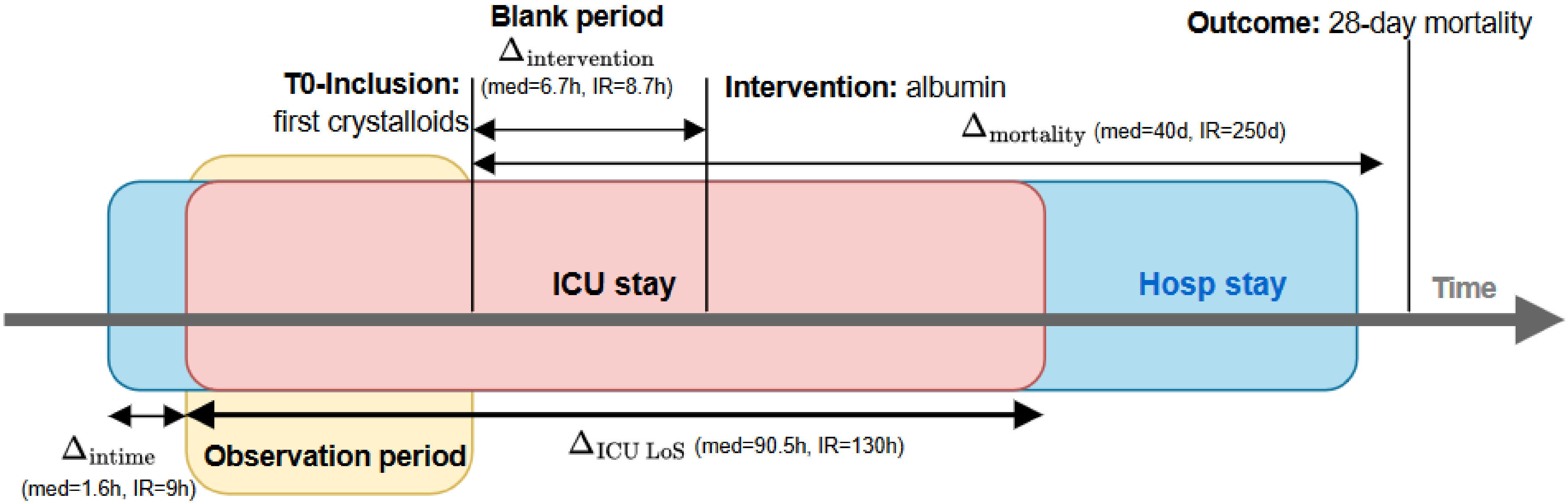
Back-up Slides

Immortal time bias introduced with different inclusion times


If 'immortal time' is misclassified into the 'treated' group or excluded from analysis, bias is induced
Immortal time
Cohort entry
Prescription filled
Event

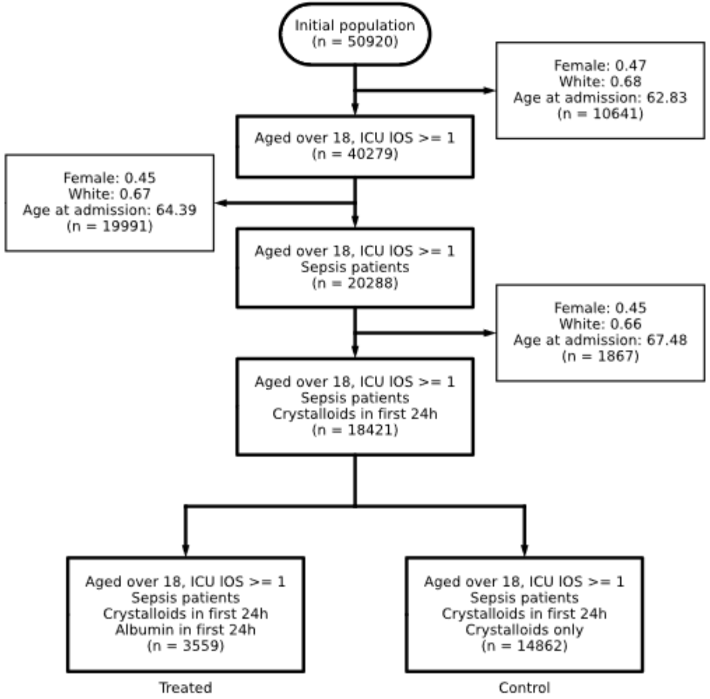
Lee, H. and D. Nunan (2020). Immortal time bias, Catalogue of Bias Collaboration. https://catalogofbias.org/biases/immortaltimebias/

Immortal time bias introduced with different inclusion times



preprint: https://hal.science/hal-04174834v2/document

Selection flowchart



Initial population
(n = 50920)

Female: 0.47
White: 0.68
Age at admission: 62.83
(n = 10641)

Aged over 18, ICU lOS >= 1
(n = 40279)

Female: 0.45
White: 0.67
Age at admission: 64.39
(n = 19991)

Aged over 18, ICU lOS >= 1
Sepsis patients
(n = 20288)

Female: 0.45
White: 0.66
Age at admission: 67.48
(n = 1867)

Aged over 18, ICU lOS >= 1
Sepsis patients
Crystalloids in first 24h
(n = 18421)

Aged over 18, ICU lOS >= 1
Sepsis patients
Crystalloids in first 24h
Albumin in first 24h
(n = 3559)

Treated

Aged over 18, ICU lOS >= 1
Sepsis patients
Crystalloids in first 24h
Crystalloids only
(n = 14862)

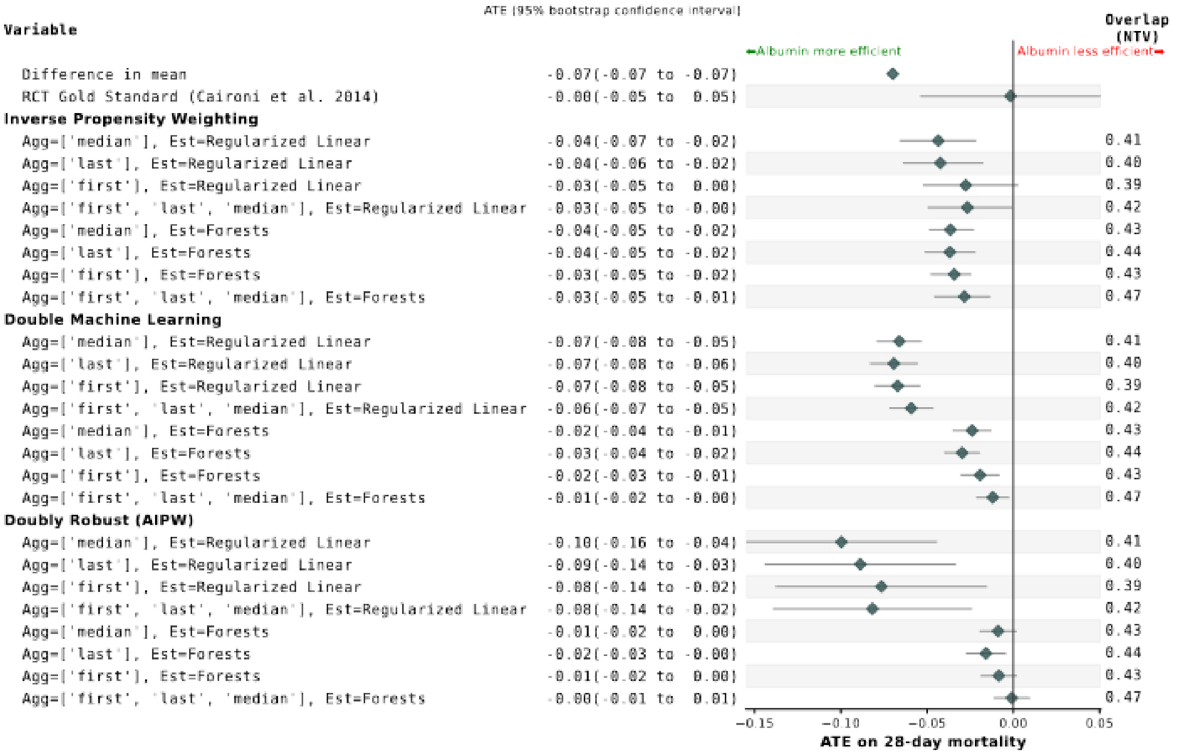
Control

Figure 12: *Selection flowchart on MIMIC-IV for the emulated trial.*

preprint: https://hal.science/hal-04174834v2/document

Does aggregation matters?

It seems not



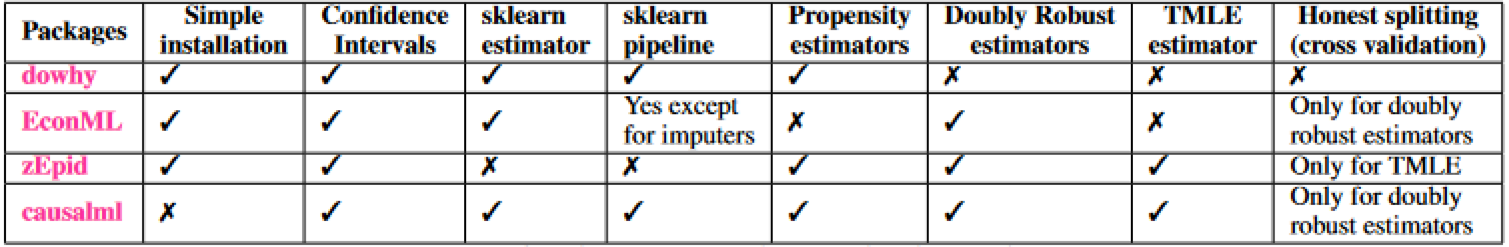
Variable	ATE (95% bootstrap confidence interval)	Overlap (NTV)
Difference in mean	-0.07(-0.07 to -0.07)	
RCT Gold Standard (Caironi et al. 2014)	-0.00(-0.05 to 0.05)	
Inverse Propensity Weighting		
Agg=['median'], Est=Regularized Linear	-0.04(-0.07 to -0.02)	0.41
Agg=['last'], Est=Regularized Linear	-0.04(-0.06 to -0.02)	0.40
Agg=['first'], Est=Regularized Linear	-0.03(-0.05 to 0.00)	0.39
Agg=['first', 'last', 'median'], Est=Regularized Linear	-0.03(-0.05 to -0.00)	0.42
Agg=['median'], Est=Forests	-0.04(-0.05 to -0.02)	0.43
Agg=['last'], Est=Forests	-0.04(-0.05 to -0.02)	0.44
Agg=['first'], Est=Forests	-0.03(-0.05 to -0.02)	0.43
Agg=['first', 'last', 'median'], Est=Forests	-0.03(-0.05 to -0.01)	0.47
Double Machine Learning		
Agg=['median'], Est=Regularized Linear	-0.07(-0.08 to -0.05)	0.41
Agg=['last'], Est=Regularized Linear	-0.07(-0.08 to -0.06)	0.40
Agg=['first'], Est=Regularized Linear	-0.07(-0.08 to -0.05)	0.39
Agg=['first', 'last', 'median'], Est=Regularized Linear	-0.06(-0.07 to -0.05)	0.42
Agg=['median'], Est=Forests	-0.02(-0.04 to -0.01)	0.43
Agg=['last'], Est=Forests	-0.03(-0.04 to -0.02)	0.44
Agg=['first'], Est=Forests	-0.02(-0.03 to -0.01)	0.43
Agg=['first', 'last', 'median'], Est=Forests	-0.01(-0.02 to -0.00)	0.47
Doubly Robust (AIPW)		
Agg=['median'], Est=Regularized Linear	-0.10(-0.16 to -0.04)	0.41
Agg=['last'], Est=Regularized Linear	-0.09(-0.14 to -0.03)	0.40
Agg=['first'], Est=Regularized Linear	-0.08(-0.14 to -0.02)	0.39
Agg=['first', 'last', 'median'], Est=Regularized Linear	-0.08(-0.14 to -0.02)	0.42
Agg=['median'], Est=Forests	-0.01(-0.02 to 0.00)	0.43
Agg=['last'], Est=Forests	-0.02(-0.03 to -0.00)	0.44
Agg=['first'], Est=Forests	-0.01(-0.02 to 0.00)	0.43
Agg=['first', 'last', 'median'], Est=Forests	-0.00(-0.01 to 0.01)	0.47

Figure 16: *Vibration analysis dedicated to the aggregation choices. The choices of aggregation only marginally modify the results. When assessed with Normalized Total Variation, the overlap assumption is respected for all our choices of aggregation. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 50 bootstrap repetitions.*

HAUTE AUTORITÉ DE SANTÉ

preprint: https://hal.science/hal-04174834v2/document

Practical implementations issues



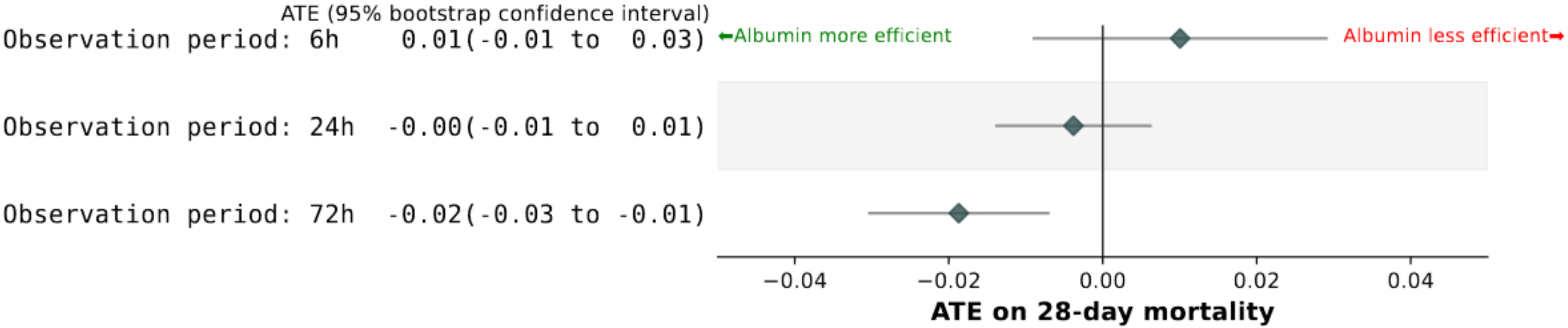
Packages	Simple installation	Confidence Intervals	sklearn estimator	sklearn pipeline	Propensity estimators	Doubly Robust estimators	TMLE estimator	Honest splitting (cross validation)
dowhy	✓	✓	✓	✓	✓	✗	✗	✗
EconML	✓	✓	✓	Yes except for imputers	✗	✓	✗	Only for doubly robust estimators
zEpid	✓	✓	✗	✗	✓	✓	✓	Only for TMLE
causalml	✗	✓	✓	✓	✓	✓	✓	Only for doubly robust estimators

Table 6: *Selection criteria for causal python packages*

Foundings:

- Counterfactual prediction lacks off-the-shelf cross-fitting estimators
- Good practices for imputation not implemented in EconML
- Bootstrap may not yield the more efficient confidence intervals and parametric confidence intervals are rarely implemented

Immortal time bias introduced with different inclusion times



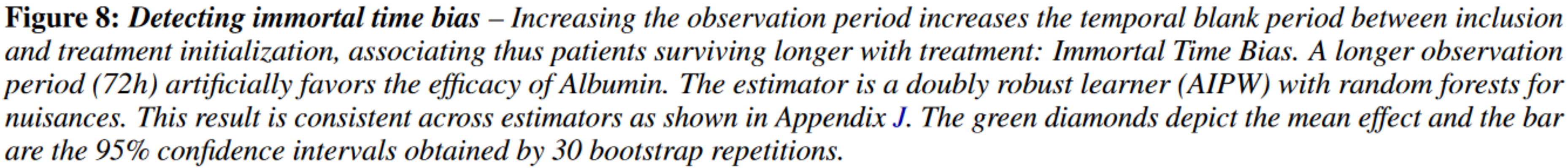


Figure 8: ***Detecting immortal time bias*** – *Increasing the observation period increases the temporal blank period between inclusion and treatment initialization, associating thus patients surviving longer with treatment: Immortal Time Bias. A longer observation period (72h) artificially favors the efficacy of Albumin. The estimator is a doubly robust learner (AIPW) with random forests for nuisances. This result is consistent across estimators as shown in Appendix J. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 30 bootstrap repetitions.*

Another study in nephrology where ITB was harder to control for: https://soda.gitlabpages.inria.fr/deepacau/#intervention-comparator

preprint: https://hal.science/hal-04174834v2/document

Causal estimators

- IPW :

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{a_i y_i}{\hat{e}(x_i)} + \frac{(1-a_i)y_i}{1-\hat{e}(x_i)}$$

- G-formula :

$$\hat{\tau}_G(f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)$$

- Augmented Inverse Propensity Weighting :

$$\widehat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}\left(X_i\right) - \hat{\mu}_{(0)}\left(X_i\right) + \frac{A_i - \hat{e}\left(X_i\right)}{\left(1-\hat{e}\left(X_i\right)\right)\hat{e}\left(X_i\right)}\left(Y_i - \hat{\mu}_{\left(A_i\right)}\left(X_i\right)\right)\right)$$

preprint: https://hal.science/hal-04174834v2/document

Heterogeneous Treatment Effect

- Double ML, built-in:

$$\hat{\tau}(\cdot) = \operatorname{argmin}_{\tau} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left((y_i - m(x_i)) - (a_i - e(x_i))\, \tau\left(x_i^{cate}\right) \right)^2 \right\}$$

- Double Robust, final regression:

$$\arg\min_{\theta} \mathbb{E}_n \left[(\tilde{Y} - \theta(X_{CATE}) \cdot \tilde{A})^2 \right]$$

Where $\tilde{Y} = Y - \hat{\mu}(X, A)$ and $\tilde{A} = A - \hat{e}(X)$

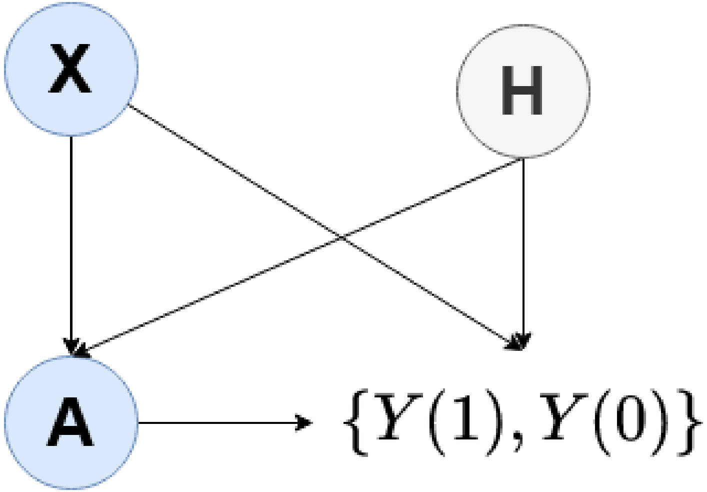
Causal inference: Assumption

1 – Ignorability ie. Unconfoundedness

We have enough information to capture all difference between **treated** and **controls** before intervention ie. Intervention is random conditionnaly on X.

Counter Exemple of missing confounder:

- Patients with head trauma
- X = age
- H = Trauma gravity (*ex. assessed w. Glasgow*)
- A = Neurological evaluation in 2 hours
- Y = Mortality at one week



preprint: https://hal.science/hal-04174834v2/document

Causal inference: Assumption

1 – Ignorability ie. Unconfoundedness

We have enough information to capture all difference between **treated** and **controls** before intervention ie. Intervention is random conditionnaly on X.

Mathematically:

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

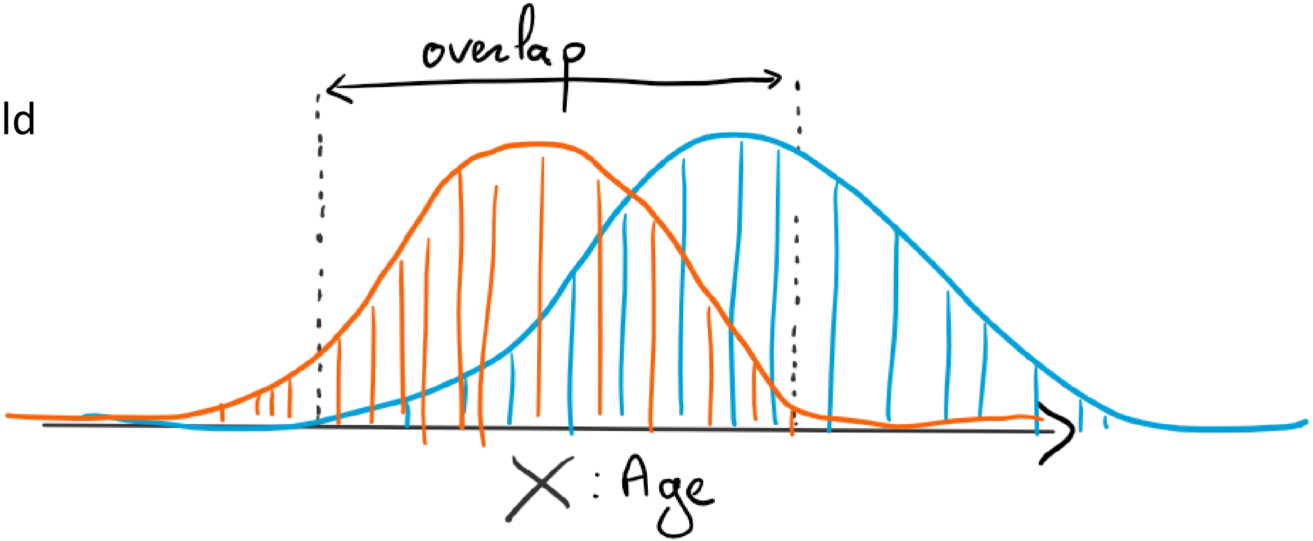
⚠ **Not verifiable with data only:**

To understand why read https://probml.github.io/pml-book/book2.html introduction on causality

Assumptions

2 – Positivity (overlap)

Treated and **controls** should be **close enough**



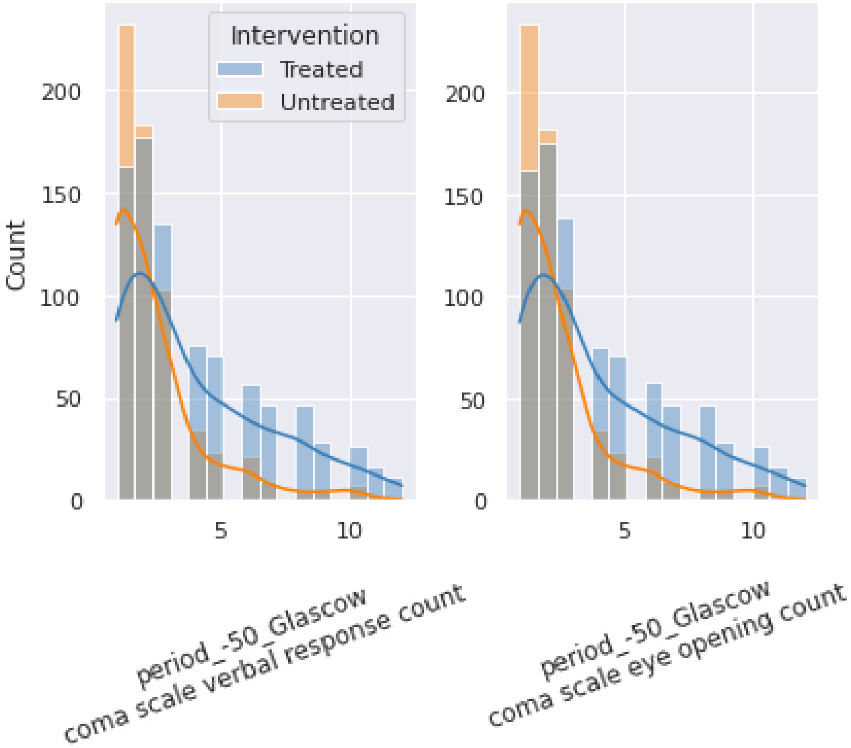
$$\exists \eta > 0, st, \eta < e(x) < 1 - \eta \quad \forall x \in \mathcal{X}$$

preprint: https://hal.science/hal-04174834v2/document

Assumptions

2 – Positivity (overlap)

Treated and **controls** should be **close enough**



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

Assumptions

3 - Consistance

For a patient, **the outcome** corresponds to the **potential outcome** of its treatment.

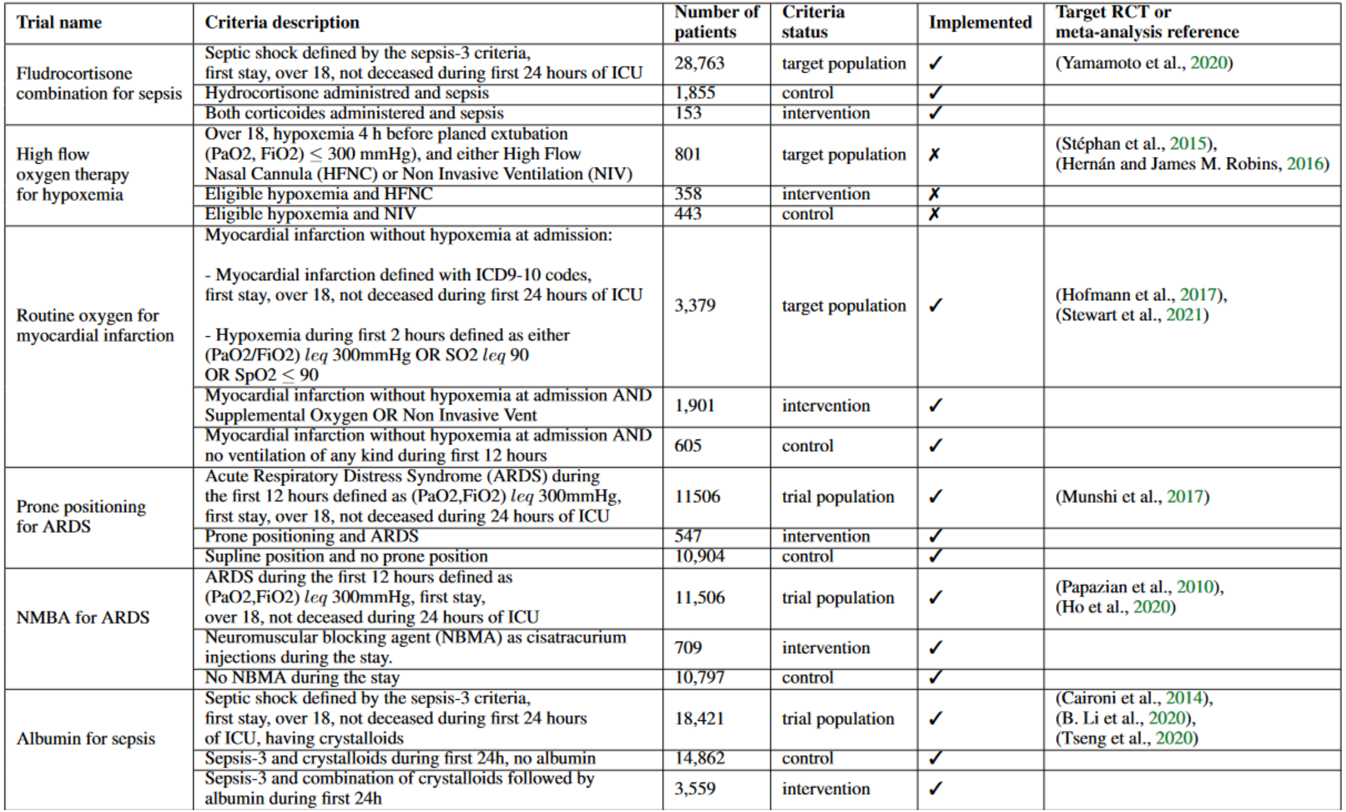
$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

All intervention are identical between individual and there is no interactions.

4 - Observations identiquement et indépendamment distribuées

- Full data (with potential outcomes) are iid.

Other emulated trials which could be studied in Mimic



Trial name	Criteria description	Number of patients	Criteria status	Implemented	Target RCT or meta-analysis reference
Fludrocortisone combination for sepsis	Septic shock defined by the sepsis-3 criteria, first stay, over 18, not deceased during first 24 hours of ICU	28,763	target population	✓	(Yamamoto et al., 2020)
	Hydrocortisone administred and sepsis	1,855	control	✓	
	Both corticoides administered and sepsis	153	intervention	✓	
High flow oxygen therapy for hypoxemia	Over 18, hypoxemia 4 h before planed extubation (PaO2, FiO2) $\leq$ 300 mmHg), and either High Flow Nasal Cannula (HFNC) or Non Invasive Ventilation (NIV)	801	target population	✗	(Stéphan et al., 2015), (Hernán and James M. Robins, 2016)
	Eligible hypoxemia and HFNC	358	intervention	✗	
	Eligible hypoxemia and NIV	443	control	✗	
Routine oxygen for myocardial infarction	Myocardial infarction without hypoxemia at admission: - Myocardial infarction defined with ICD9-10 codes, first stay, over 18, not deceased during first 24 hours of ICU - Hypoxemia during first 2 hours defined as either (PaO2/FiO2) leq 300mmHg OR SO2 leq 90 OR SpO2 $\leq$ 90	3,379	target population	✓	(Hofmann et al., 2017), (Stewart et al., 2021)
	Myocardial infarction without hypoxemia at admission AND Supplemental Oxygen OR Non Invasive Vent	1,901	intervention	✓	
	Myocardial infarction without hypoxemia at admission AND no ventilation of any kind during first 12 hours	605	control	✓	
Prone positioning for ARDS	Acute Respiratory Distress Syndrome (ARDS) during the first 12 hours defined as (PaO2,FiO2) leq 300mmHg, first stay, over 18, not deceased during 24 hours of ICU	11506	trial population	✓	(Munshi et al., 2017)
	Prone positioning and ARDS	547	intervention	✓	
	Supline position and no prone position	10,904	control	✓	
NMBA for ARDS	ARDS during the first 12 hours defined as (PaO2,FiO2) leq 300mmHg, first stay, over 18, not deceased during 24 hours of ICU	11,506	trial population	✓	(Papazian et al., 2010), (Ho et al., 2020)
	Neuromuscular blocking agent (NBMA) as cisatracurium injections during the stay.	709	intervention	✓	
	No NBMA during the stay	10,797	control	✓	
Albumin for sepsis	Septic shock defined by the sepsis-3 criteria, first stay, over 18, not deceased during first 24 hours of ICU, having crystalloids	18,421	trial population	✓	(Caironi et al., 2014), (B. Li et al., 2020), (Tseng et al., 2020)
	Sepsis-3 and crystalloids during first 24h, no albumin	14,862	control	✓	
	Sepsis-3 and combination of crystalloids followed by albumin during first 24h	3,559	intervention	✓	

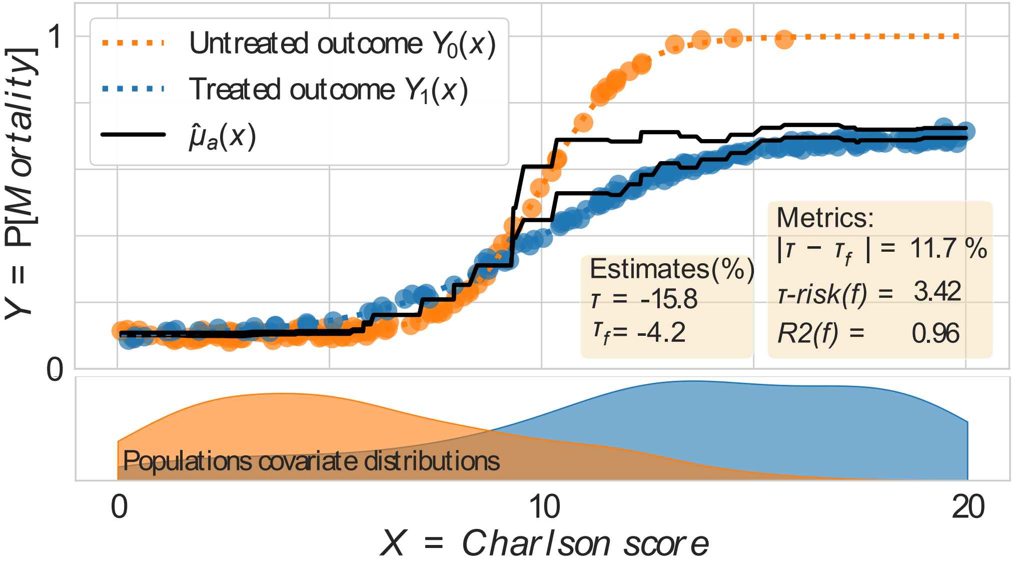
preprint: https://hal.science/hal-04174834v2/document

Select a model: ML 101

🧑‍🏫 **Select model with small MSE(y, f) on Out-Of-Samples**

A) Random Forest

🥇 Perfect R2, 😭 Poor inference



preprint: https://hal.science/hal-04174834v2/document

Select a model: ML 101

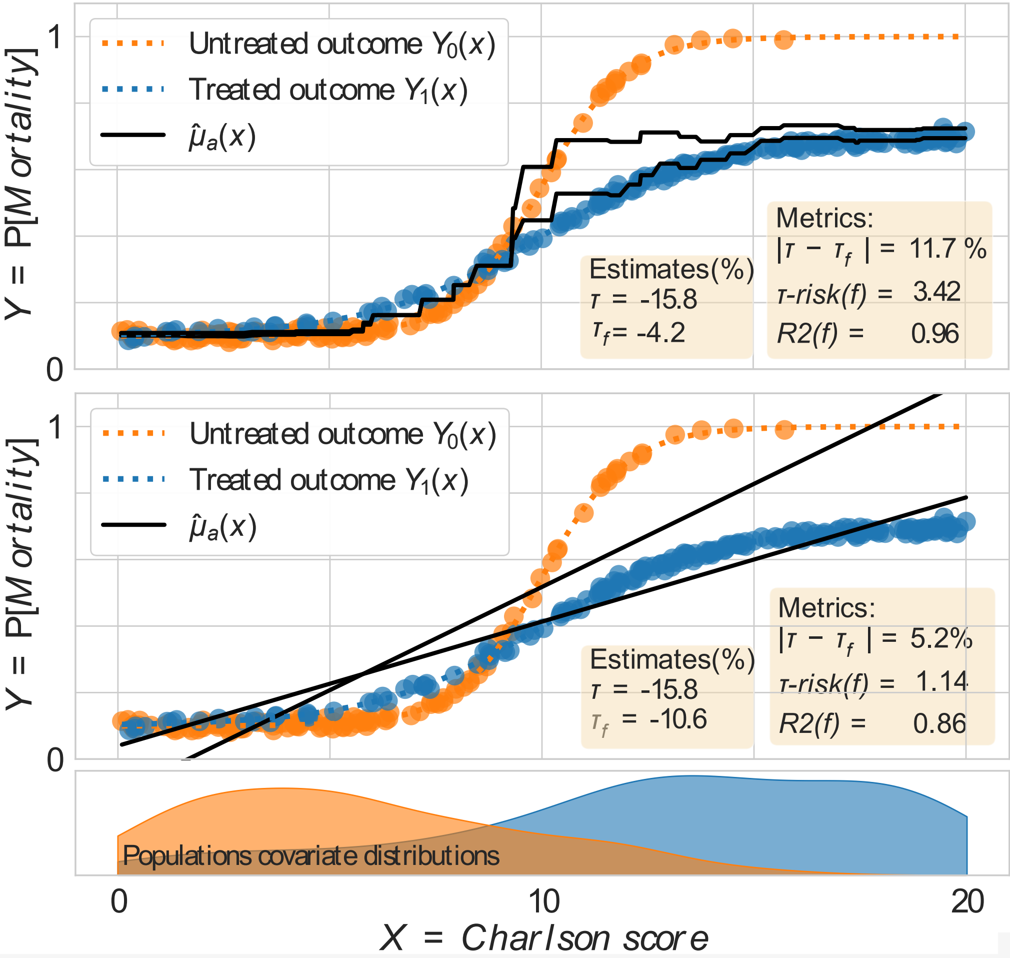
🧑‍🏫 **Select model with smallest MSE(y, f) on Out-Of-Samples**

A) Random Forest

🥇 **Perfect R2,** 😭 **Poor inference**

B) Linear model

😭 **Bad R2,** 🥇 **good inference**



preprint: https://hal.science/hal-04174834v2/document

References, model selection

- V. Dorie, J. Hill, U. Shalit, M. Scott, et D. Cervone, « Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition », *arXiv:1707.02641 [stat]*, juill. 2018, Consulté le: janv. 08, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1707.02641
- Nie, Xinkun, et Stefan Wager. « Quasi-Oracle Estimation of Heterogeneous Treatment Effects ». *arXiv:1712.04912 [econ, math, stat]*, 13 décembre 2017. http://arxiv.org/abs/1712.04912.
- Schuler, Alejandro, Michael Baiocchi, Robert Tibshirani, et Nigam Shah. « A Comparison of Methods for Model Selection When Estimating Individual Treatment Effects ». *ArXiv:1804.05146 [Cs, Stat]*, 13 juin 2018. http://arxiv.org/abs/1804.05146
- Johansson, Fredrik D., Nathan Kallus, Uri Shalit, et David Sontag. « Learning Weighted Representations for Generalization Across Designs ». *arXiv:1802.08598 [stat]*, 26 février 2018. http://arxiv.org/abs/1802.08598.
- Shalit, Uri, Fredrik D. Johansson, et David Sontag. « Estimating individual treatment effect: generalization bounds and algorithms ». *arXiv:1606.03976 [cs, stat]*, 16 mai 2017. http://arxiv.org/abs/1606.03976.
- Alaa, Ahmed, et Mihaela Van Der Schaar. « Validating Causal Inference Models via Influence Functions ». In *International Conference on Machine Learning*, 191-201. PMLR, 2019. http://proceedings.mlr.press/v97/alaa19a.html.