Representations and inference from time-varying routine care data

Matthieu Doutreligne
Inria SODA,
Haute Autorité de Santé

Direction of the thesis:
Gaël Varoquaux, INRIA SODA

Co-supervision
Claire Morgand, ARS IDF

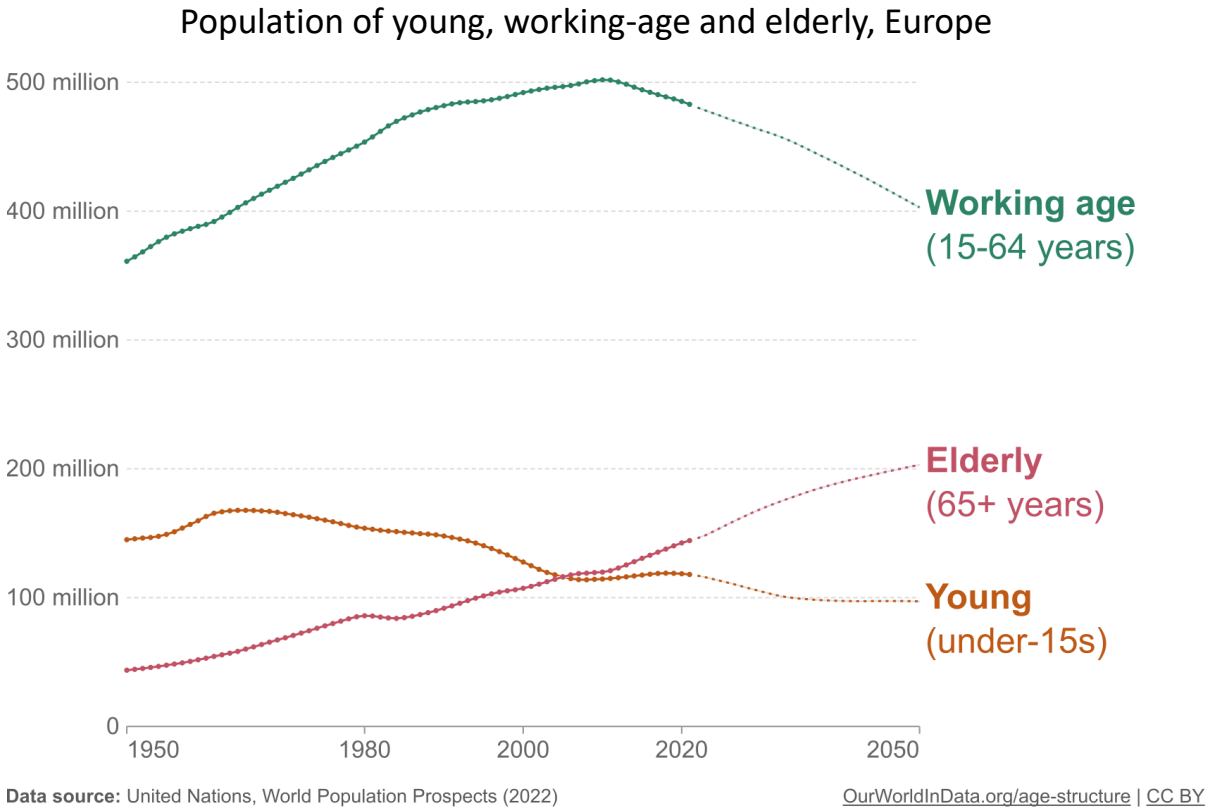
HAS
HAUTE AUTORITÉ DE SANTÉ

Inria

Soda

Important considerations in public health

Fast demographic evolution

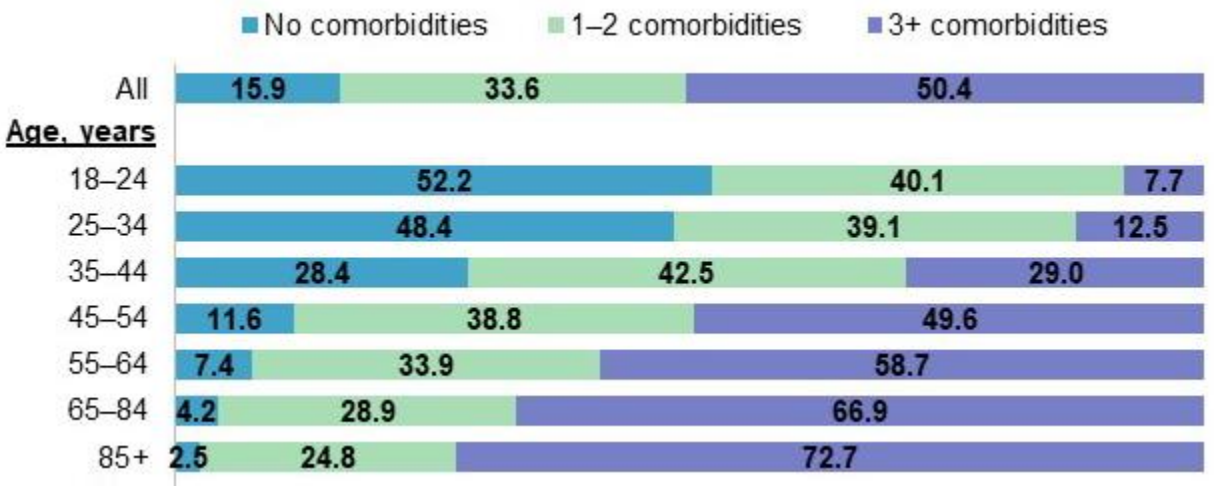


Important considerations in public health

Fast demographic evolution

Consequences:

- Increasing comorbidities



Comorbidities Associated With Adult Inpatient Stays, 2019, Agency for Healthcare Research and Quality

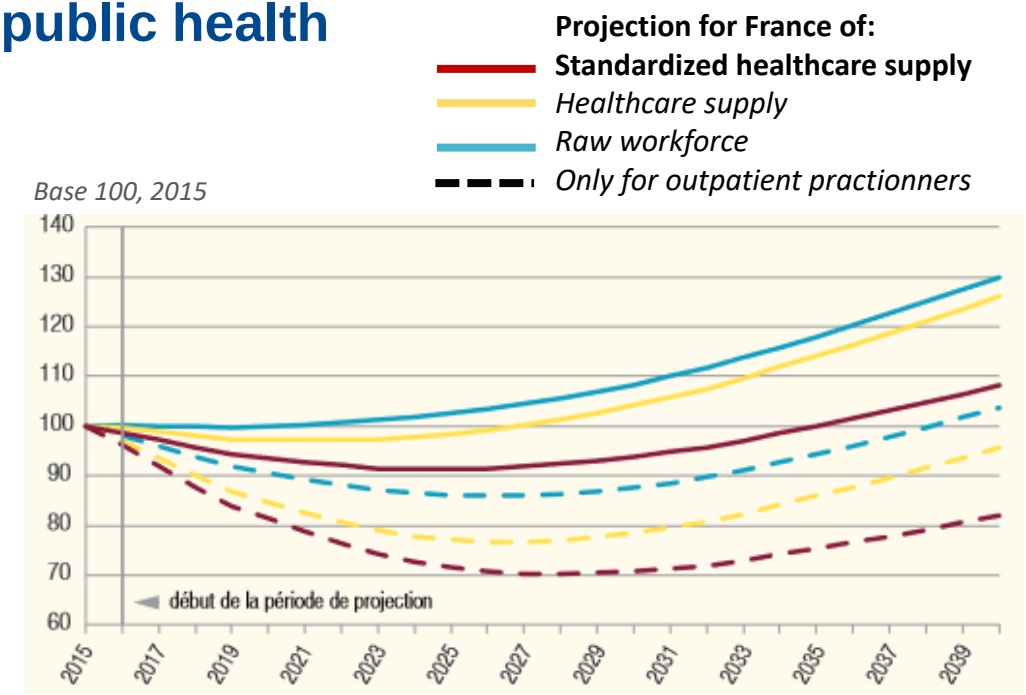
Important considerations in public health

Fast demographic evolution

Consequences:

- Increasing comorbidities
- Scarcity of medical practionners
- Constrained financial resources

Increasing number of available treatments to choose among



Les médecins d'ici à 2040 : une population plus jeune, plus féminisée et plus souvent salariée, Etudes et Résultats, Drees, Ministère de la Santé et de la Prévention, mai 2017

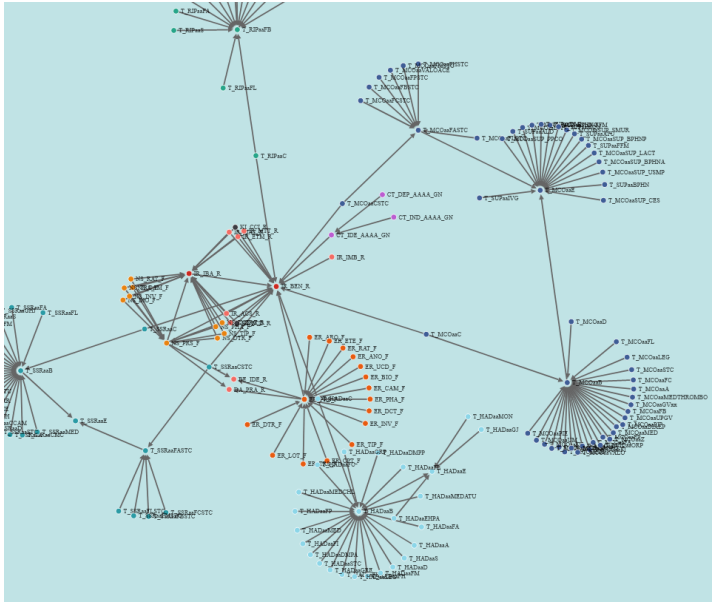
Population ageing puts a lot of constraint on modern healthcare systems

These constraints call for resource optimization

Healthcare data can contribute thanks to better:

- Planning
- Prevention
- Choice of effective interventions adapted to each patient

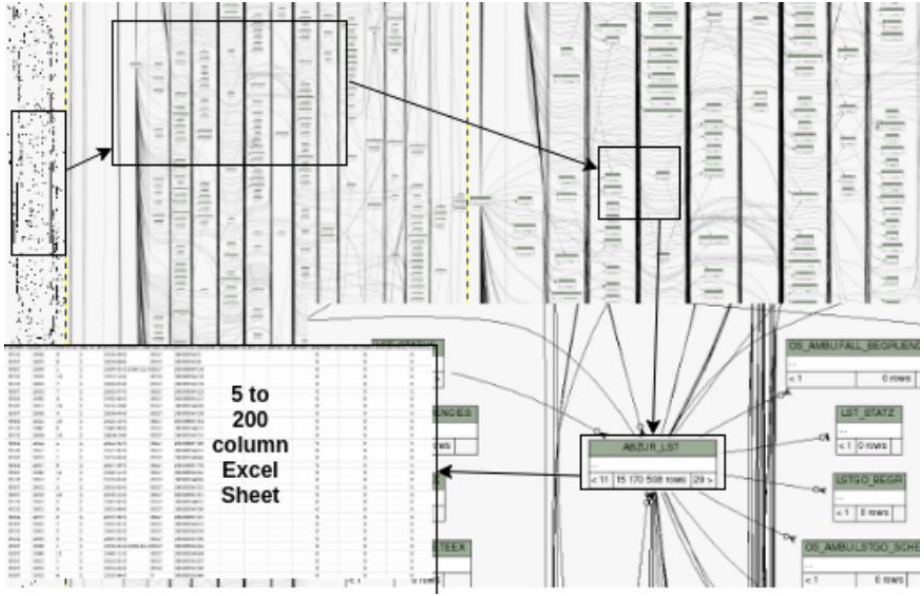
Routine healthcare databases (Real World Data)



Claims:

ex. French National Claims, SNDS, 68M patients

Mostly administrative variables eg. billing codes, prescriptions



Clinical Health Records (CHRs):

ex. Paris hospitals (AP-HP), 10M patients

Detailed clinical variables

Large routine care databases are increasingly available

Characteristics of routine care data

Benefits for public health

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

Characteristics of routine care data

Benefits for public health

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

Drawbacks

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

- **High dimensional data**

The characteristics of routine care data require dedicated methods and questions

How can routine care data contribute to better resource allocation?

Contributions

I. Exploring a complexity gradient in representation and predictive models for EHRs

ongoing work

II. Prediction is not all we need: Causal thinking for decision making on EHRs

submitted to Lancet Digital Health

III. How to select predictive models for causal inference?

Rework in-progress for submission to Jamia

IV. Potential and challenges of Clinical Data Warehouse, a case study in France

published in PLOS Digital Health

Examples of resource optimization

- Prevent acute events by considering risk reduction procedures
- Allocate human resources in priority to potentially long stays
- Avoid early hospital discharges to diminish preventable readmission

Examples of resource optimization

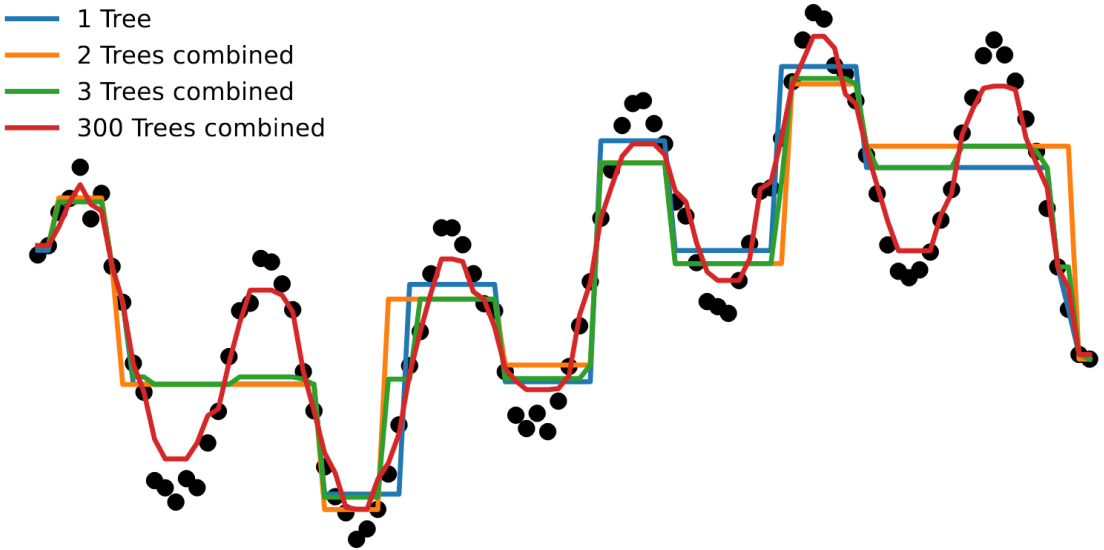
- **Prevent** acute events by considering risk reduction procedures
- Allocate human resources in priority to **potentially** long stays
- Avoid early hospital discharges to diminish **preventable** readmission

We need to better understand the future

Predictive models are a key ingredient for resources optimization

Machine learning, a toolbox for predictive models

- Find an **estimator** $f : x \rightarrow y$ that approximates the true value of y so that $f(x) \approx y$
  Modern algorithms automatically extract patterns **linking similar *x* to similar *y***



Boosted trees:
iterative ensemble of trees

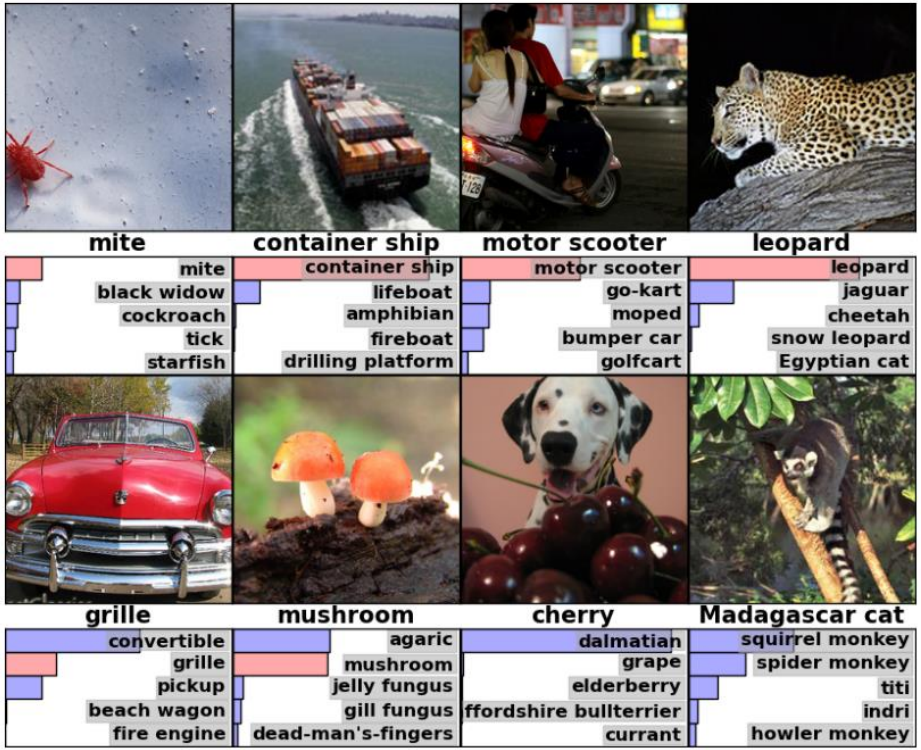
- Models are selected for their **predictive accuracy on out-of-sample data**

Machine learning does not focus on the form of the estimator but on predictive accuracy

Machine learning predicts well for various complex data

Images

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.



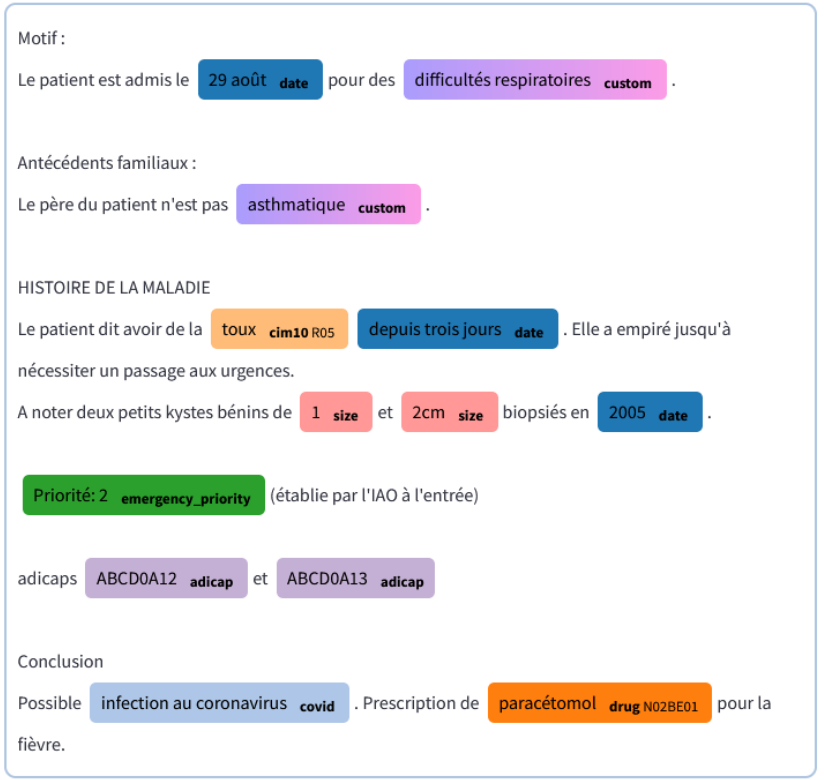
Machine learning predicts well for various complex data

Images

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Text

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.



Machine learning algorithms range from large language models to regularized linear models

Machine learning predicts well for various complex data

Images

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

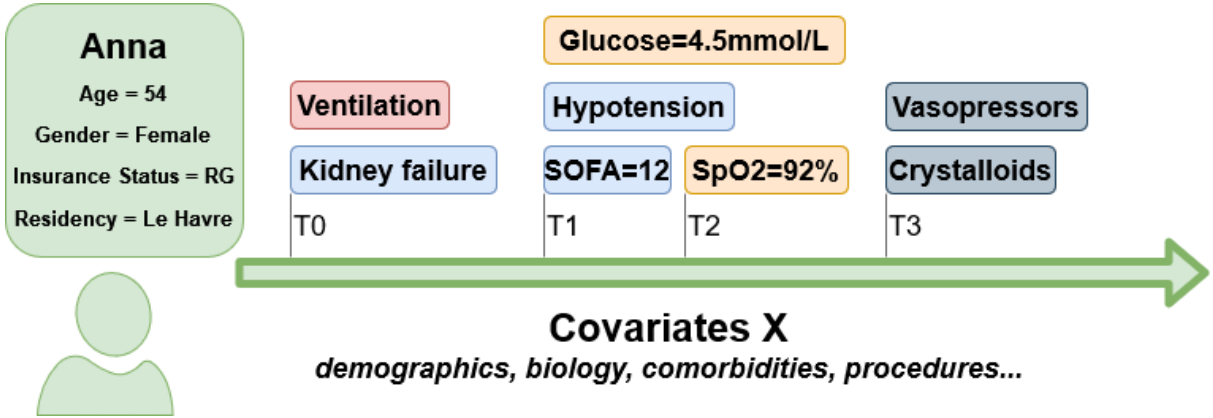
Text

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Routine care data?

Sequential (100+ events)

High-dimensional (10, 000+ modalities)



What is the level of complexity required for time-varying routine care data?

Over-optimistic claims of machine learning for healthcare

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). **Scalable and accurate deep learning with electronic health records**. *NPJ digital medicine*

	Hospital A	Hospital B
Inpatient Mortality, AUROC[1](95% CI)		
Deep learning 24 hours after admission	**0.95**(0.94-0.96)	**0.93**(0.92-0.94)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)
Full feature simple baseline at 24 hours after admission	0.93 (0.91-0.94)	0.90 (0.88-0.92)
Baseline (aEWS[2]) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
30-day Readmission, AUROC (95% CI)		
Deep learning at discharge	**0.77**(0.75-0.78)	**0.76**(0.75-0.77)
Full feature enhanced baseline at discharge	0.75 (0.73-0.76)	0.75 (0.74-0.76)
Full feature simple baseline at discharge	0.74 (0.73-0.76)	0.73 (0.72-0.74)
Baseline (mHOSPITAL[3]) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
Length of Stay at least 7 days AUROC (95% CI)		
Deep learning 24 hours after admission	**0.86**(0.86-0.87)	**0.85**(0.85-0.86)
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)
Baseline (mLiu[4]) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)

Full feature enhance baseline = Linear model on top of measurements grouped by time buckets (1 day, 1 week, 1 month, 1 year, >1year)

🤔 **Deep learning is not significantly better** for two tasks

Elaborate deep learning model does not outperform a simple linear model

How complex a predictive model for routine care data should be?

Empirical study: planning and prevention with AP-HP data

Raw cohort from AP-HP (Paris hospitals) of 200,000 patients, two tasks:

- Long length of stay
- **Prognosis**

	Prognosis
Description	Next stay prognosis: ICD10 chapter classification
Task	Multi-Label binary classification (20 classes)
Cohort Size	10,786
Prevalence	From 1.3 to 55.9%
Number of cases	From 139 to 6,029

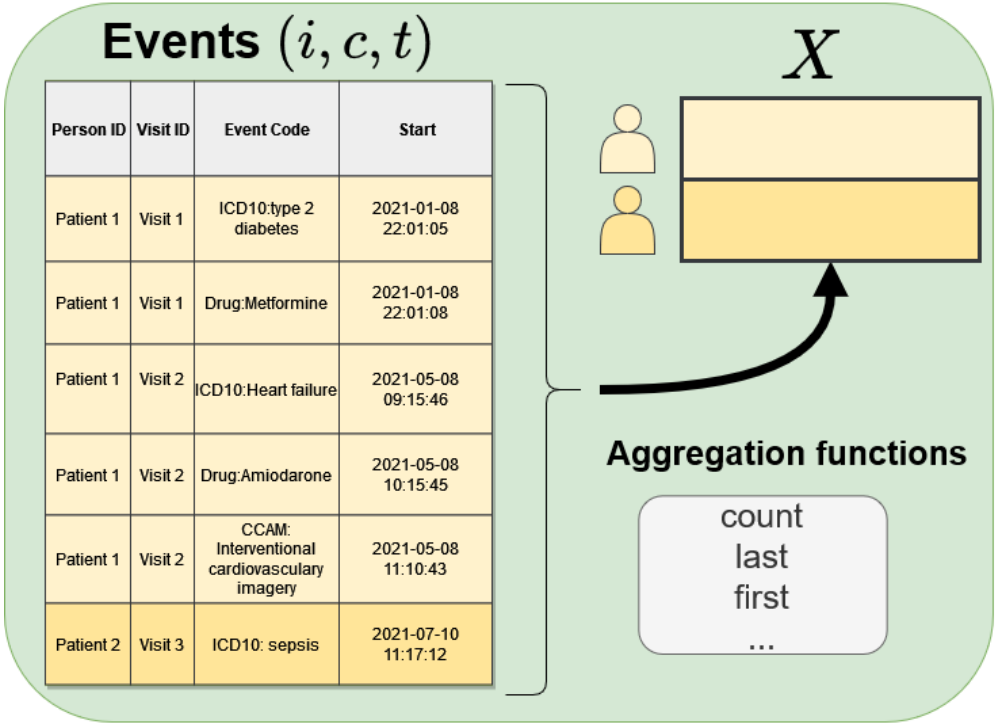
What model better predict the next stay diagnosis from the data of the previous stay?

Focus on ICD10 code prediction: chain aggregation and estimation

Timeline aggregation:

- Demographics (only static variables)
- Decayed counting
- Local embeddings
- SNDS embeddings
- Transformer embeddings

Increasing complexity



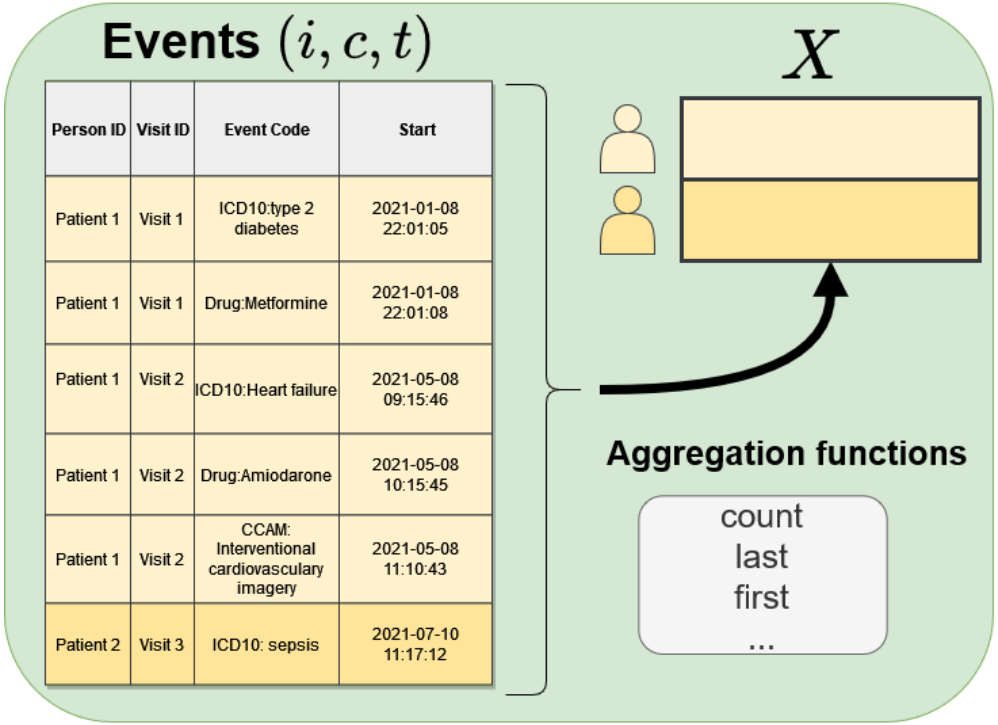
Focus on ICD10 code prediction: chain aggregation and estimation

Timeline aggregation:

- Demographics (only static variables)
- Decayed counting
- Local embeddings
- SNDS embeddings
- Transformer embeddings

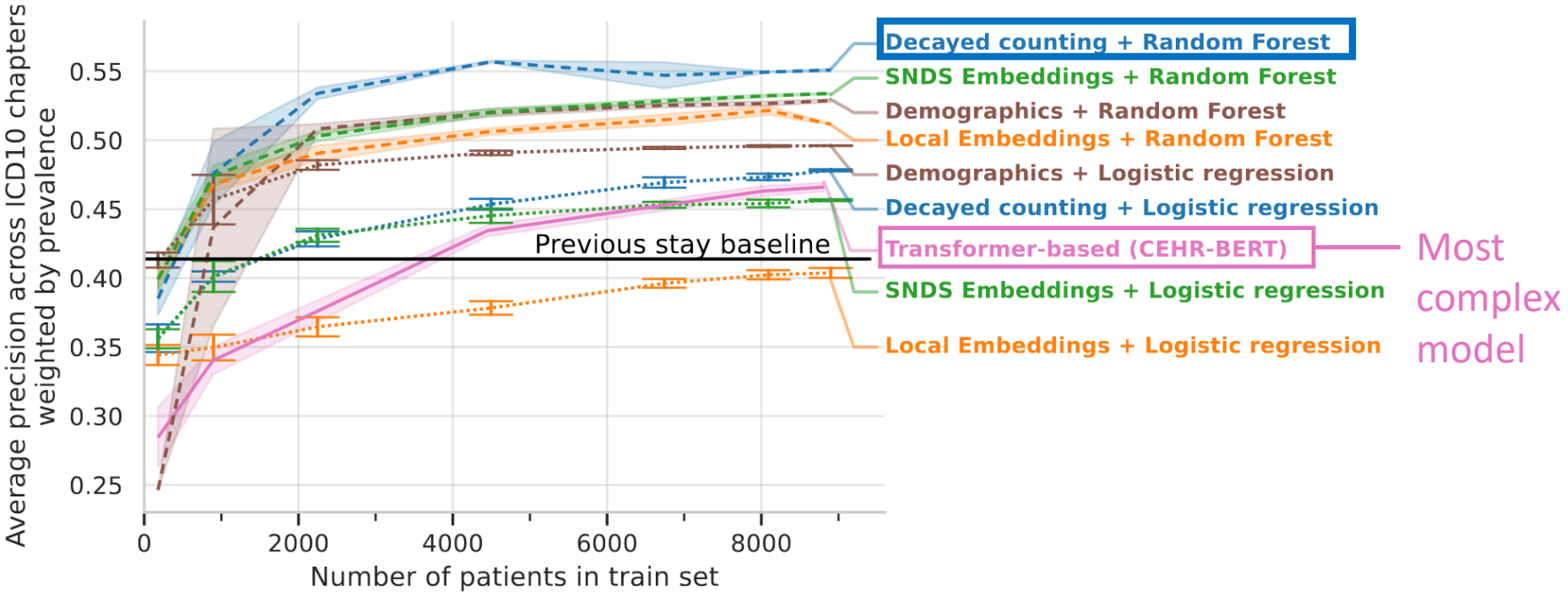
Estimator:

- Linear model
- Random Forest
- Transformer prediction head



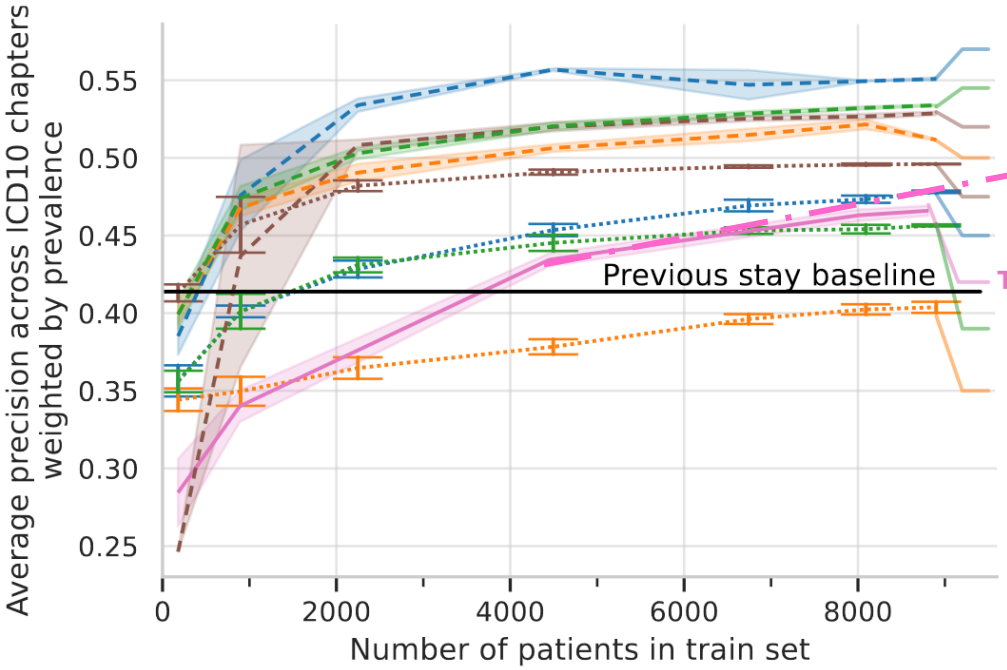
We benchmark a gradient of models: from simple to complex.

Results: ICD10 code prediction



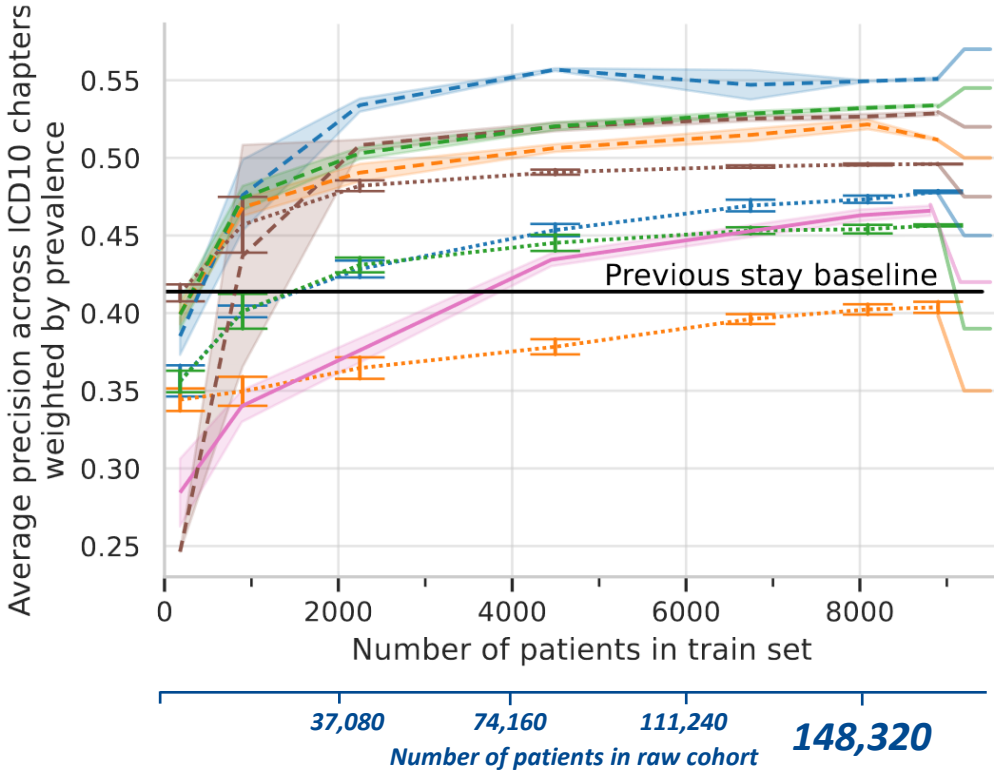
1. Complexity does not win

Results: ICD10 code prediction

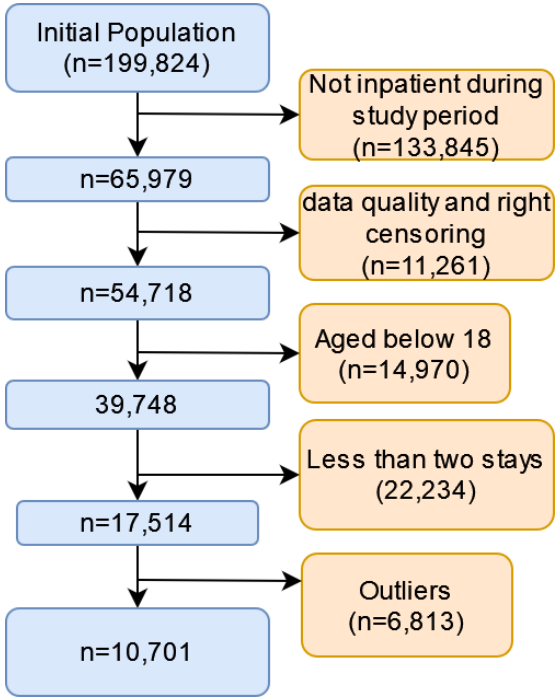


2. With more data, a complex transformer architecture could be the best model

Results: ICD10 code prediction



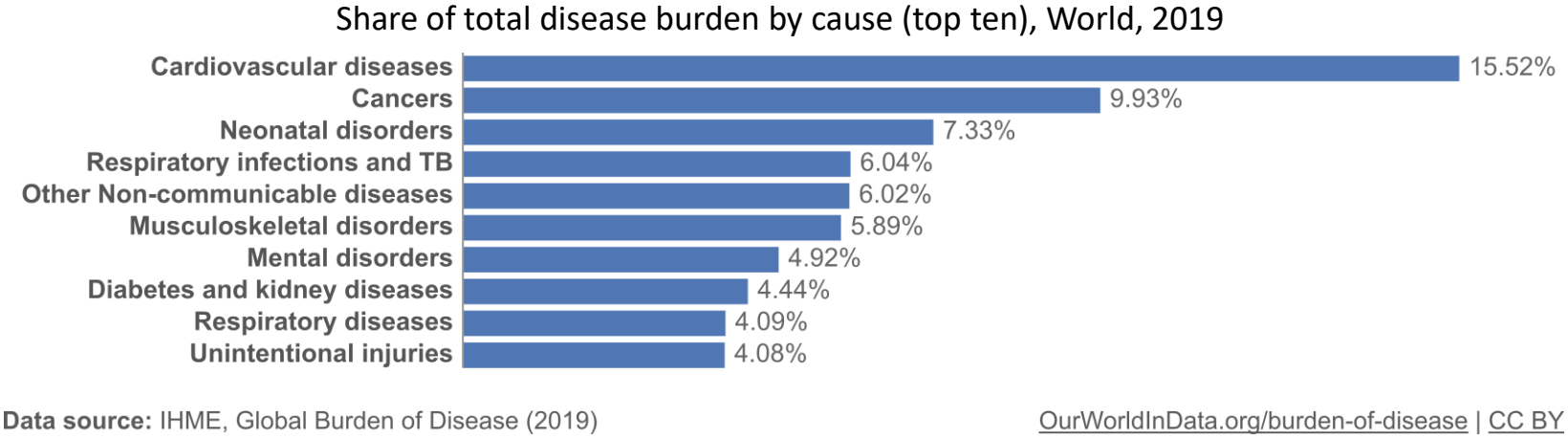
Selection flowchart (train + test)



3. We already use a lot of data: big healthcare data is not so big

A prevention task requiring big data: Major Adverse Cardiovascular Events

Cardiovascular diseases are the leading cause of death worldwide (>15%)



A prevention task requiring big data: Major Adverse Cardiovascular Events

Cardiovascular diseases are the leading cause of death worldwide (>15%).

But in AP-HP data, the number of cases is small from a statistical learning perspective.

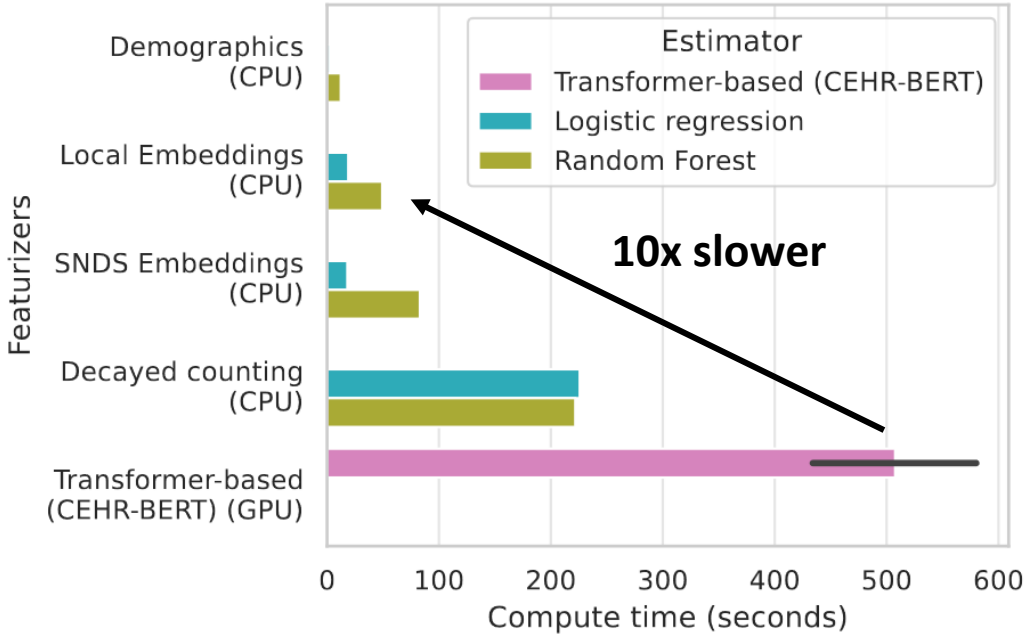
	MACE
Task	Binary classification
Description	MACE prognosis at one year
Cohort Size	165,948
Prevalence	2.6 %
Number of cases	4,315

A prevention task requiring big data: Major Adverse Cardiovascular Events

Cardiovascular diseases are the leading cause of death worldwide (>15%).

But in AP-HP data, the number of cases is still rare from a statistical learning perspective.

Computing resources are lacking for complex models.



Computational resources are needed for prevention but hard to collocate within hospital

How complex a prective model should be?

- Simple representations and estimators predict well for medium sized datasets.

  Random forest outperforms a transformer with 5 ROC-AUC points

- Data is not so big due to inclusion criteria and low prevalence.

  From 2,000,000 patients to 4,316 cases

- Benchmarking predictive models requires more computing power that what is actually available for routine care data.

  Less than a good laptop for each project

Prediction might not be all we need

- UK deployed the cardiovascular disease (CVD) Qrisk score, it is **accurate and well calibrated.**
- It is used to recommend statins, even for moderate CVD risks and primary prevention.

UK, N.C.G.C. (2014). Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease.

Prediction might not be all we need

- UK deployed the cardiovascular disease (CVD) Qrisk score, it is **accurate and well calibrated.**
- It is used to recommend statins, even at moderate CVD risks and primary prevention.
- However, it **did not reduce the burden of cardiovascular diseases.**

Eriksen, C. U., Rotar, O., Toft, U. & Jørgensen, T. What is the effectiveness of systematic population-level screening programmes for reducing the burden of cardiovascular diseases? (World Health Organization. Regional Office for Europe, 2021).

Prediction might not be all we need

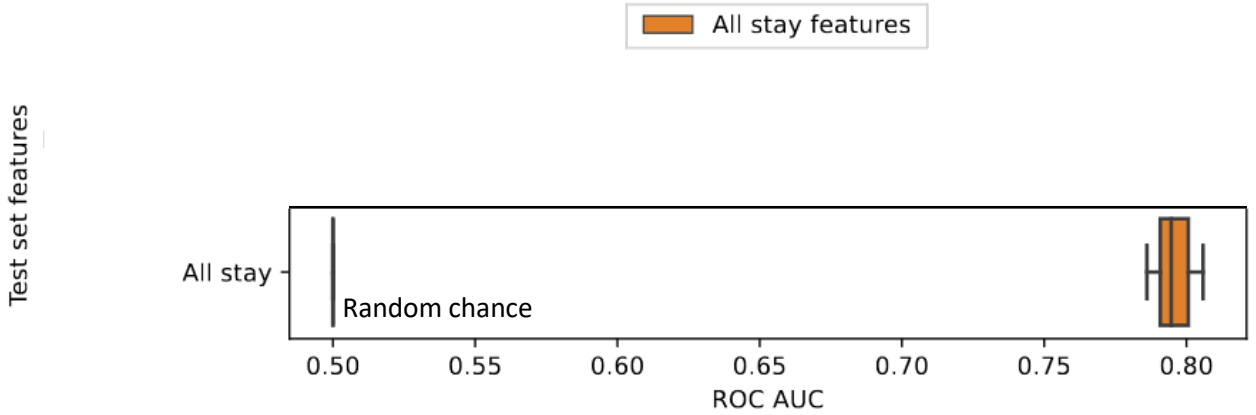
- UK deployed the cardiovascular disease (CVD) Qrisk score, it is **accurate and well calibrated.**
- It is used to recommend statins, even at moderate CVD risks and primary prevention.
- However, it **did not reduce the burden of cardiovascular diseases.**
- Probably because it **did not target the responders and compliers.**

Krska, J., du Plessis, R., & Chellaswamy, H. (2016). Implementation of NHS Health Checks in general practice: variation in delivery between practices and practitioners. *Primary health care research & development*

Prediction fails when not associated to an appropriate and realistic intervention

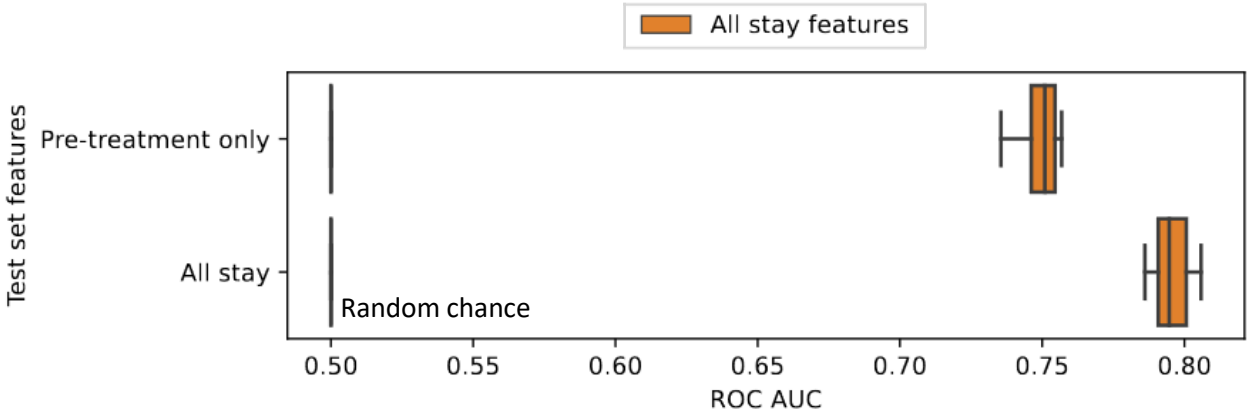
28-day mortality prediction informing the administration of fluids for sespis

- Train with post-treatment variables
- Evaluate on out-of-sample with the same variables (all stay)



28-day mortality prediction informing fluid administration for sespis

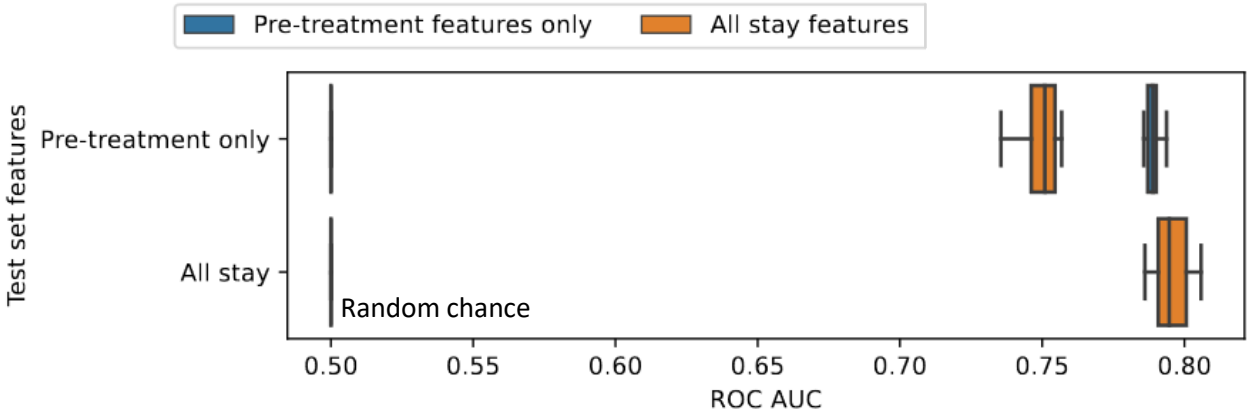
- Train with post-treatment variables
- Evaluate on a **actionnable dataset with only pre-treatment variables**



Relying on post-treatment variables (*shortcut variables*) hurts the performances

28-day mortality prediction informing fluid administration for sespis

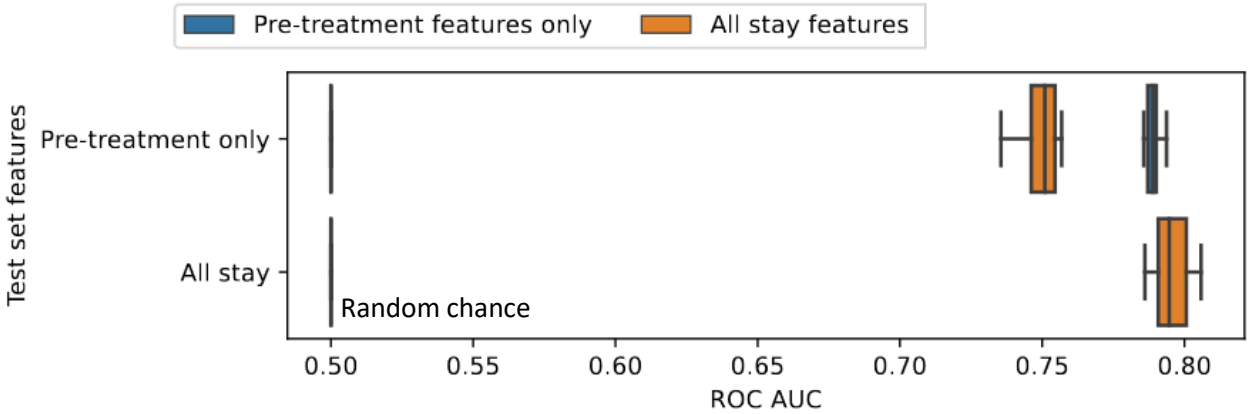
- Train with post-treatment variables
- Evaluate on a **actionnable dataset with only pre-treatment variables**



Taking into account the actionable intervention is needed to build useful algorithms

28-day mortality prediction informing fluid administration for sepsis

- Train with post-treatment variables
- Evaluate on a **actionnable dataset with only pre-treatment variables**



😳 **Who would do that?**
Answer: A lot of studies!

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.

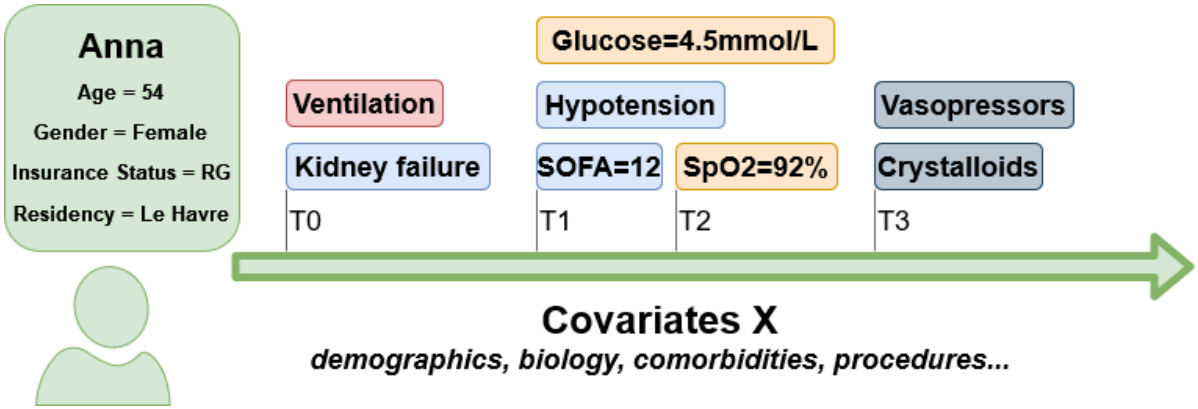
Failing to consider appropriate data damages predictive algorithms

Frame the problem

Richardson, W Scott, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. (1995). "The well-built clinical question: a key to evidence-based decisions". In: Acp j club

P I C O

Population



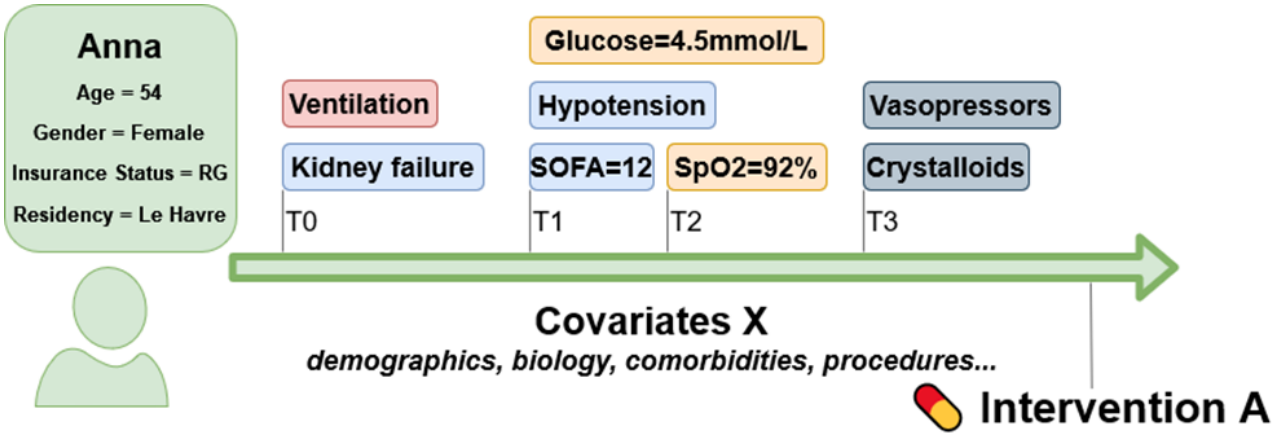
Example: Patients with sepsis in the ICU

Define the target population with features X

Frame the problem

PICO

Intervention & Control



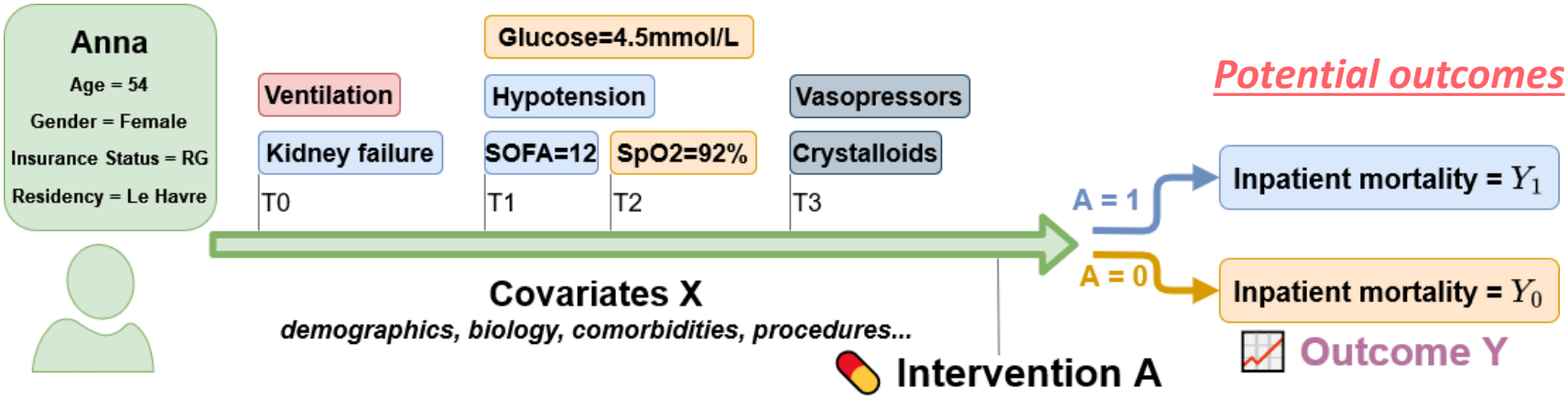
Example: Combination of crystalloids and albumin or Crystalloids only

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

Frame the problem

PICO

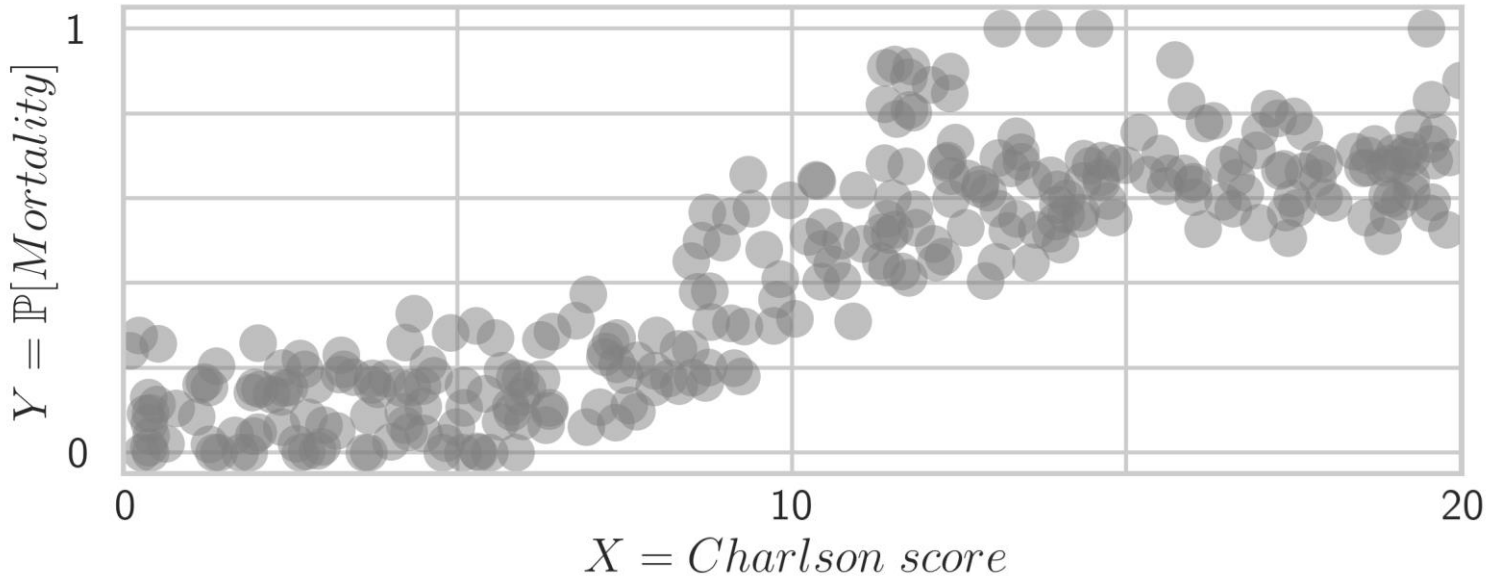
Outcome



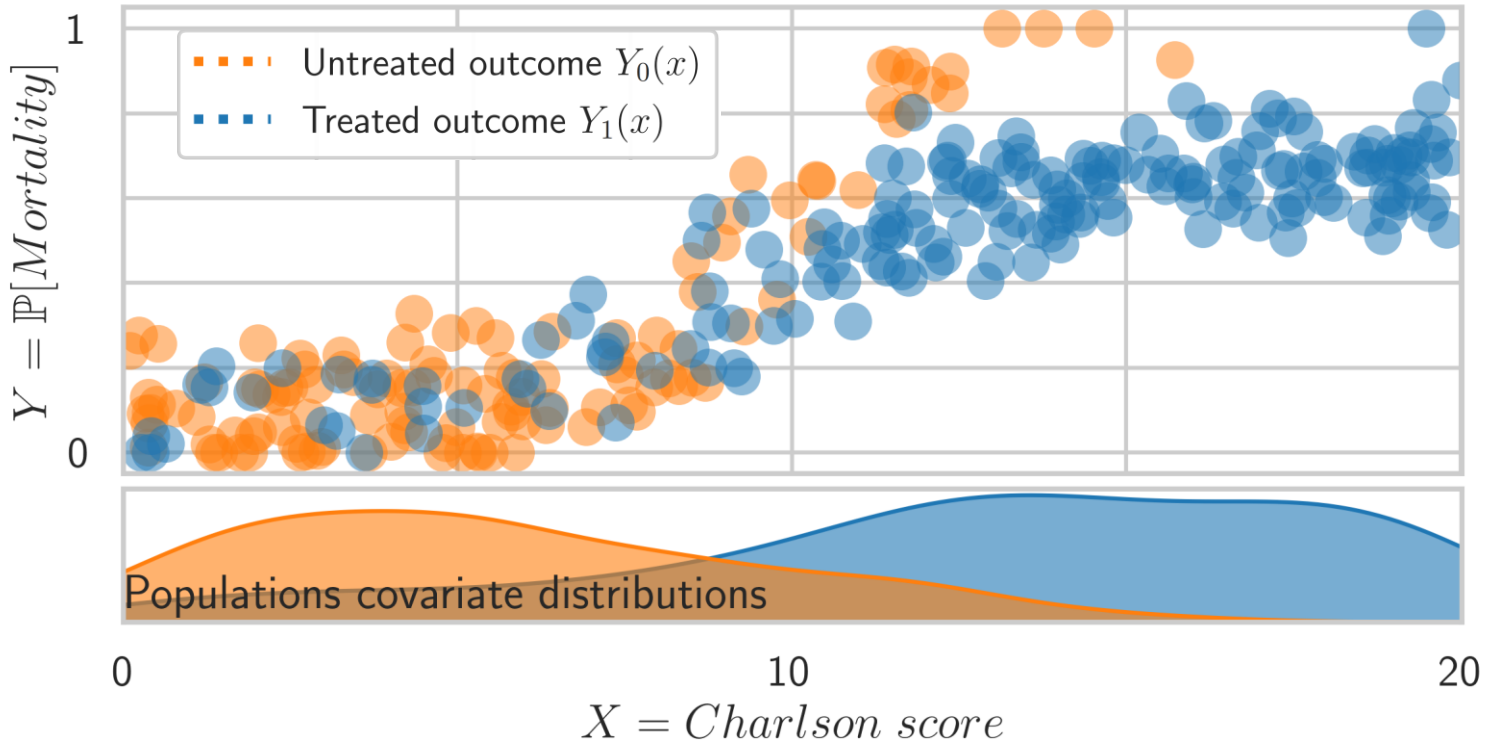
Example: 28-day survival

To improve a **clinical outcome Y**

Application with sampled data: outcomes and features

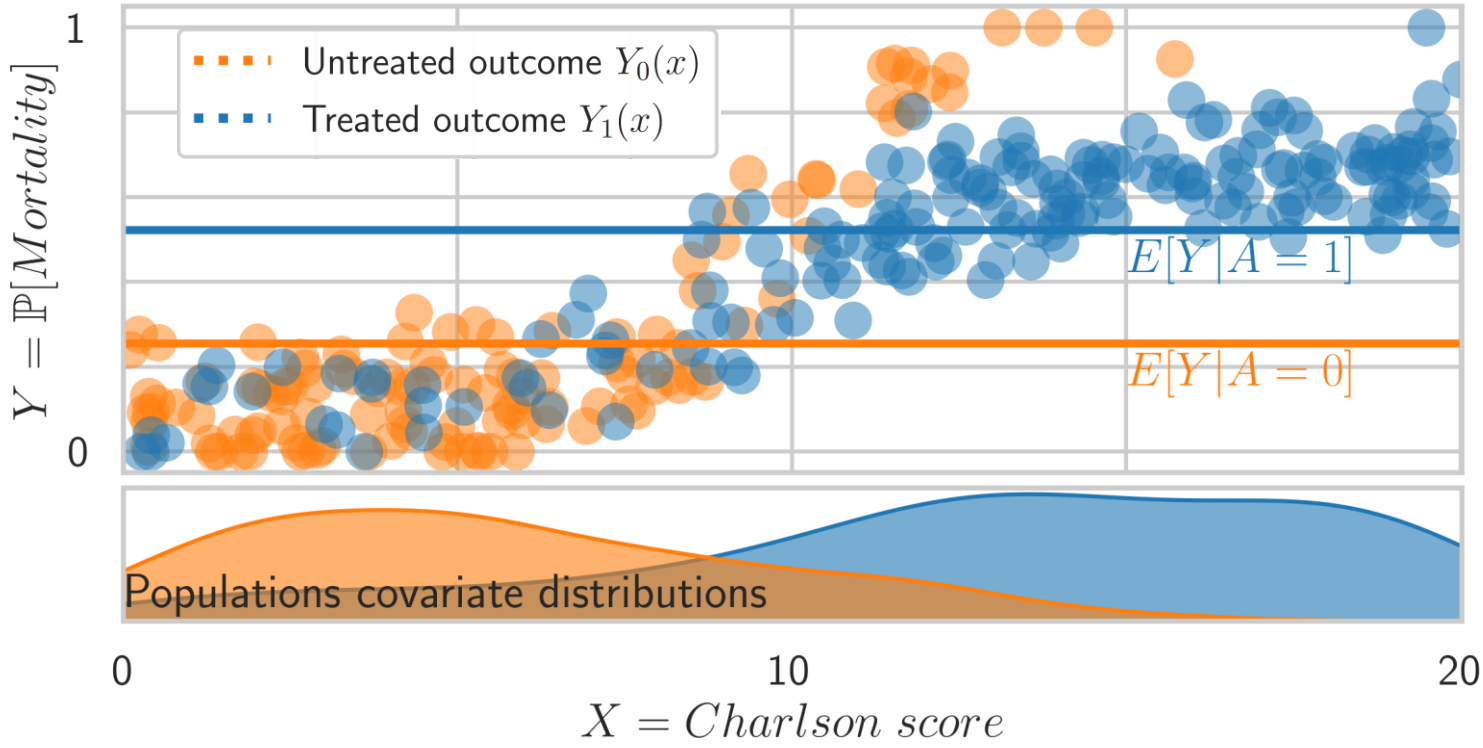


Application with sampled data: treatment and controls



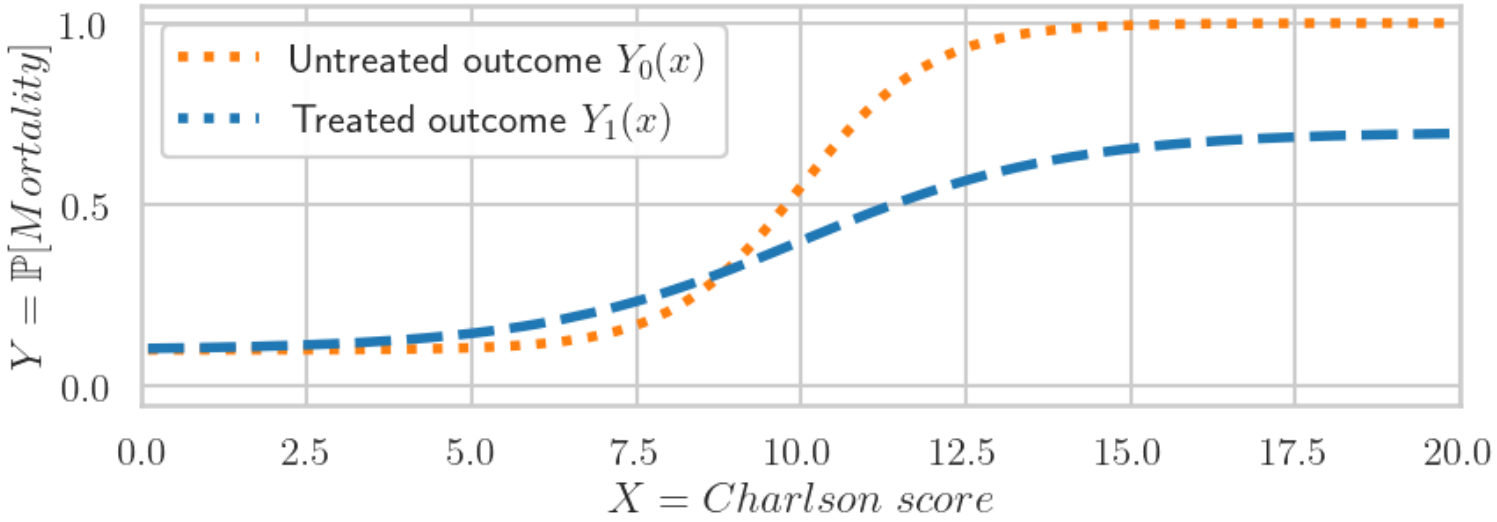
How to estimate the effect of the treatment on the outcome?

A naive (and biased) solution, difference in mean



We are not comparing apples to apples: the treated and the control differ too much

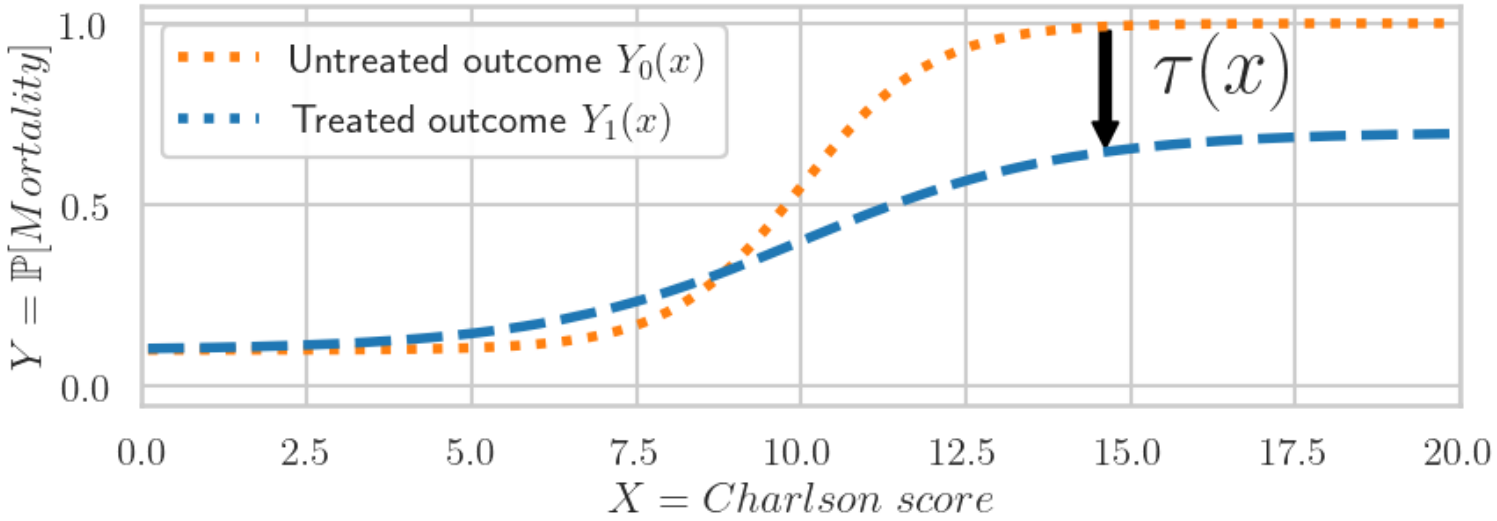
Potential outcomes, a robust statistical methodology



G. W. Imbens; D. B. Rubin (2015): Causal inference in statistics, social, and biomedical sciences. Cambridge University Press

The Neyman-Rubin framework postulates two potential outcomes curves

Potential outcomes, a robust statistical methodology



🎯 Estimates

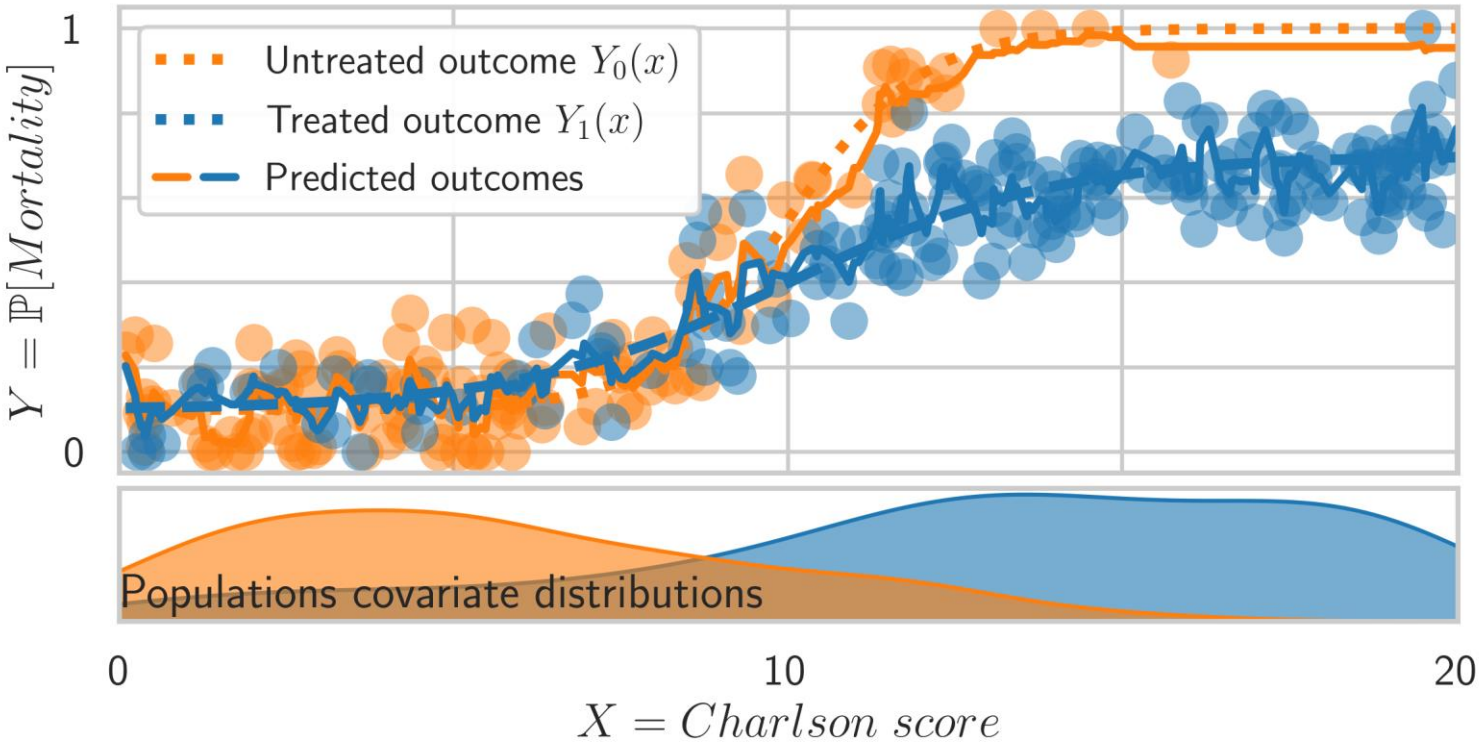
- Average Treatment Effect (ATE)
- Conditional Average Treatment Effect (CATE)

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

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

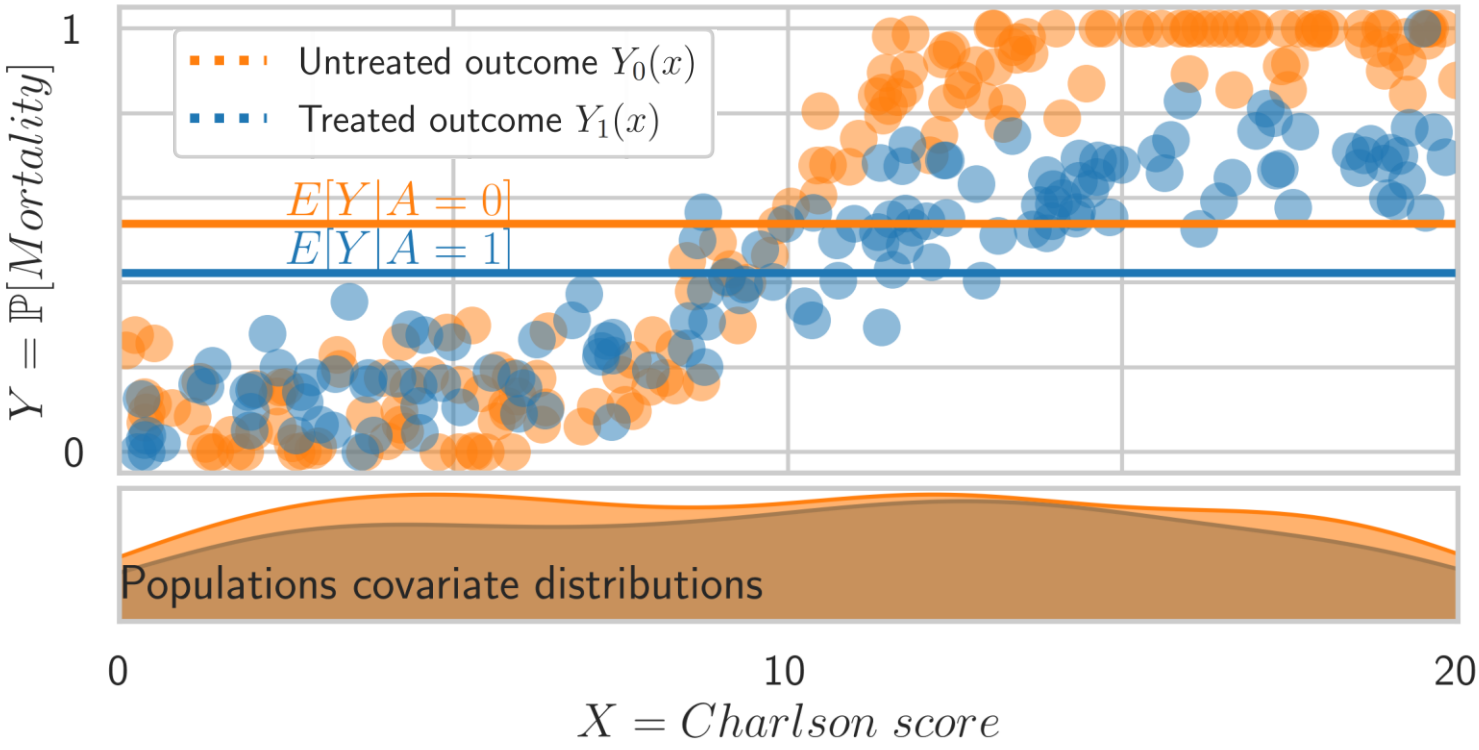
The estimate of the effect is the difference between the two potential outcomes

Model the outcome (G-formula) for tailored decision-making



Machine learning is well suited for the study of subpopulations

Randomized Controlled Trials (RCTs) for ATE



For the population effect, randomizing the treatment balances the populations

Between three worlds

- World 1 – **Epidemiology:** Carefully design the study *(framing)*
- World 2 – **Causal Inference:** Control the confounders *(identification)*
- World 3 – **Machine Learning:** Select the model *(estimation)*

What ingredients from these three worlds do we need?

How to build robust decision-making algorithm from routine care data?

A causal framework comparing the three sources of bias

1) Framing – study design
2) Identification – list confounders
3) Estimation
4) Vibration analysis: Compare different reasonable choices for the average treatment effect (ATE)
5) Conditional Average Effect: Go beyond population effect, study heterogeneity of the effect (CATE)

Calibrate the analysis thanks to the gold standard result before look into heterogeneity

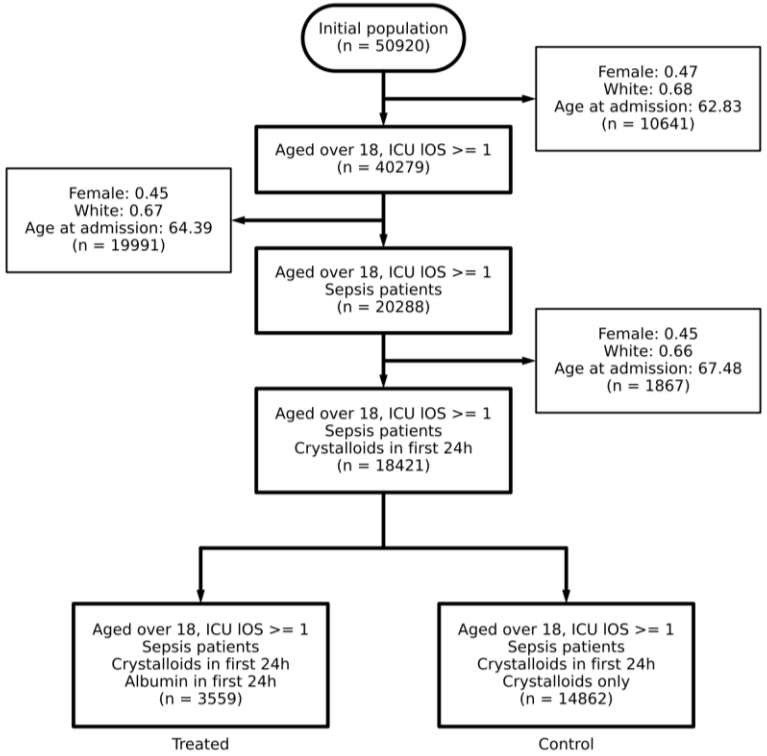
Case study with routine care data

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

Question: In patients with sepsis, what is the effect of albumin in combination with crystalloids compared to crystalloids alone on 28-day mortality?

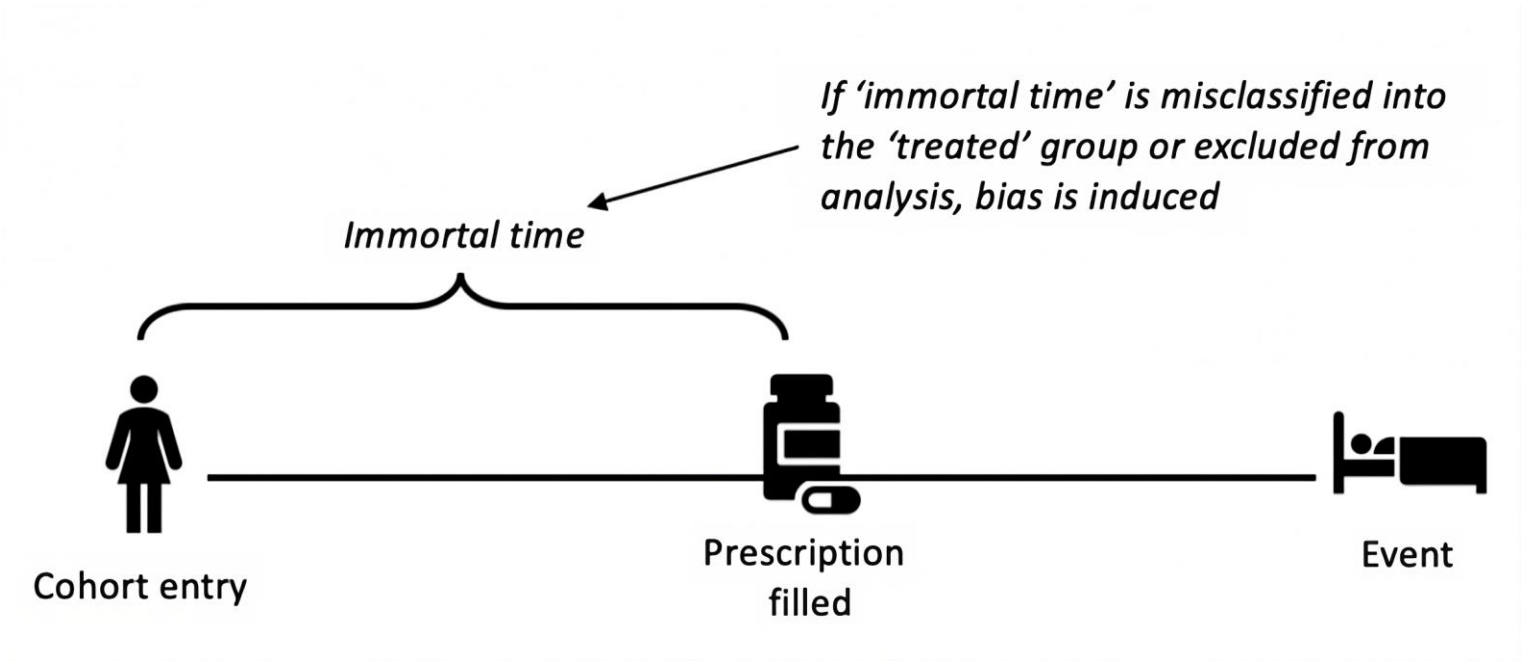
Cohort: 3,559 treated and 14,862 controls

Gold standard RCT: No effect



Caironi et al.(2014). "Albumin replacement in patients with severe sepsis or septic shock". New England Journal of Medicine

Step 1 – Poor study design

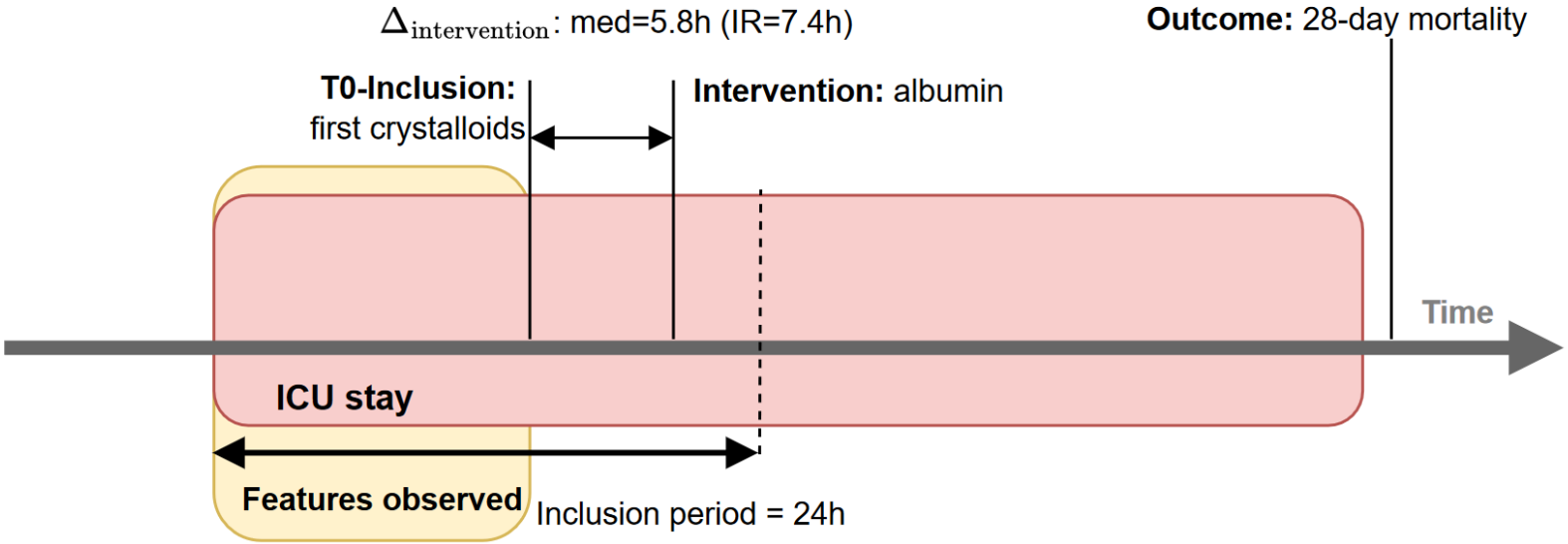


Lee, H. and D. Nunan (2020). Immortal time bias, Catalogue of Bias Collaboration. https://catalogofbias.org/biases/immortal-time-bias/

Step 1 – Poor study design

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

Example: During 24 first hours of hospitalization



Immortal time bias is introduced because treatment and control are not aligned

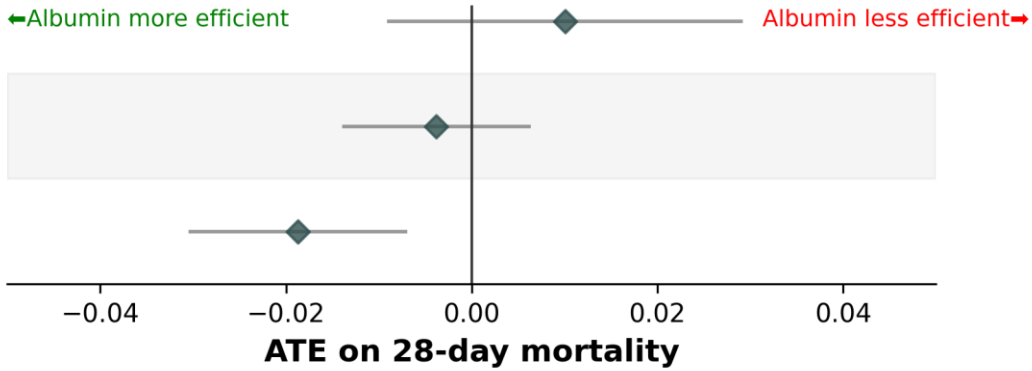
Step 1 – Poor study design

ATE (95% bootstrap confidence interval)

Observation period: 6h 0.01(-0.01 to 0.03)

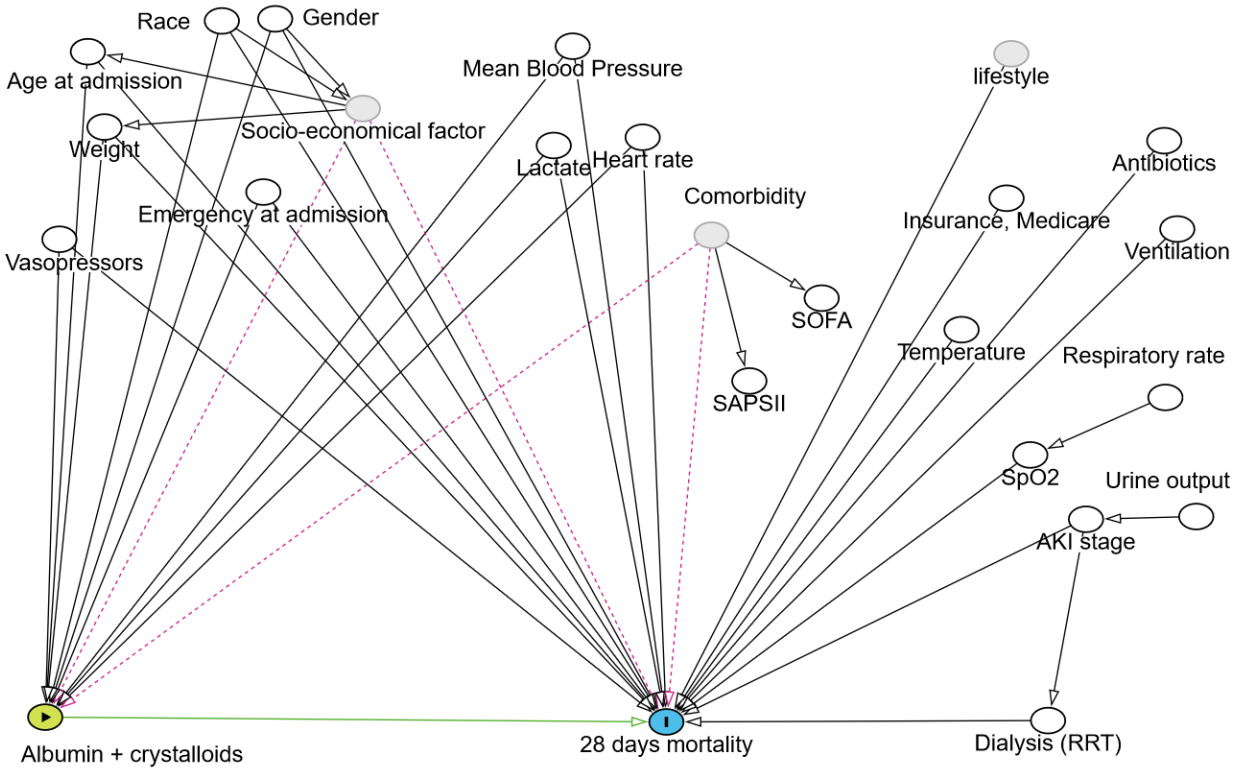
Observation period: 24h -0.00(-0.01 to 0.01)

Observation period: 72h -0.02(-0.03 to -0.01)



Poor design can lead to erroneous conclusions

Step 2 – Identification

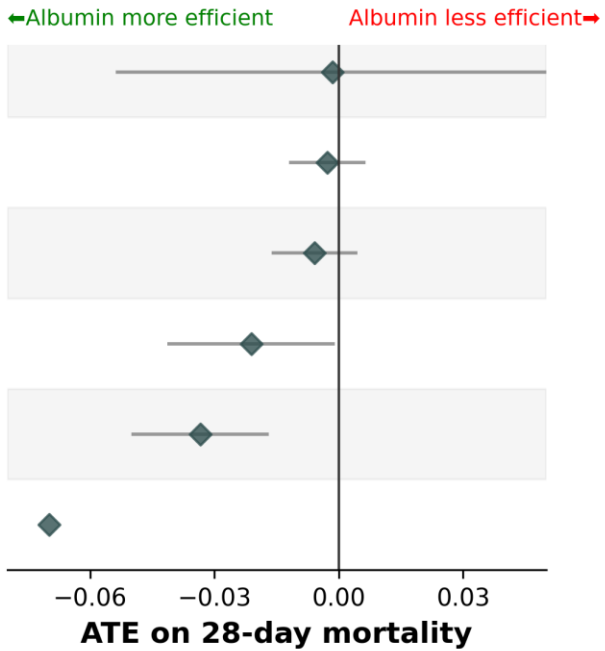


List confounders to answer the question with a Directed Acyclic Graph

Step 2 – Compare less informed sets of confounders

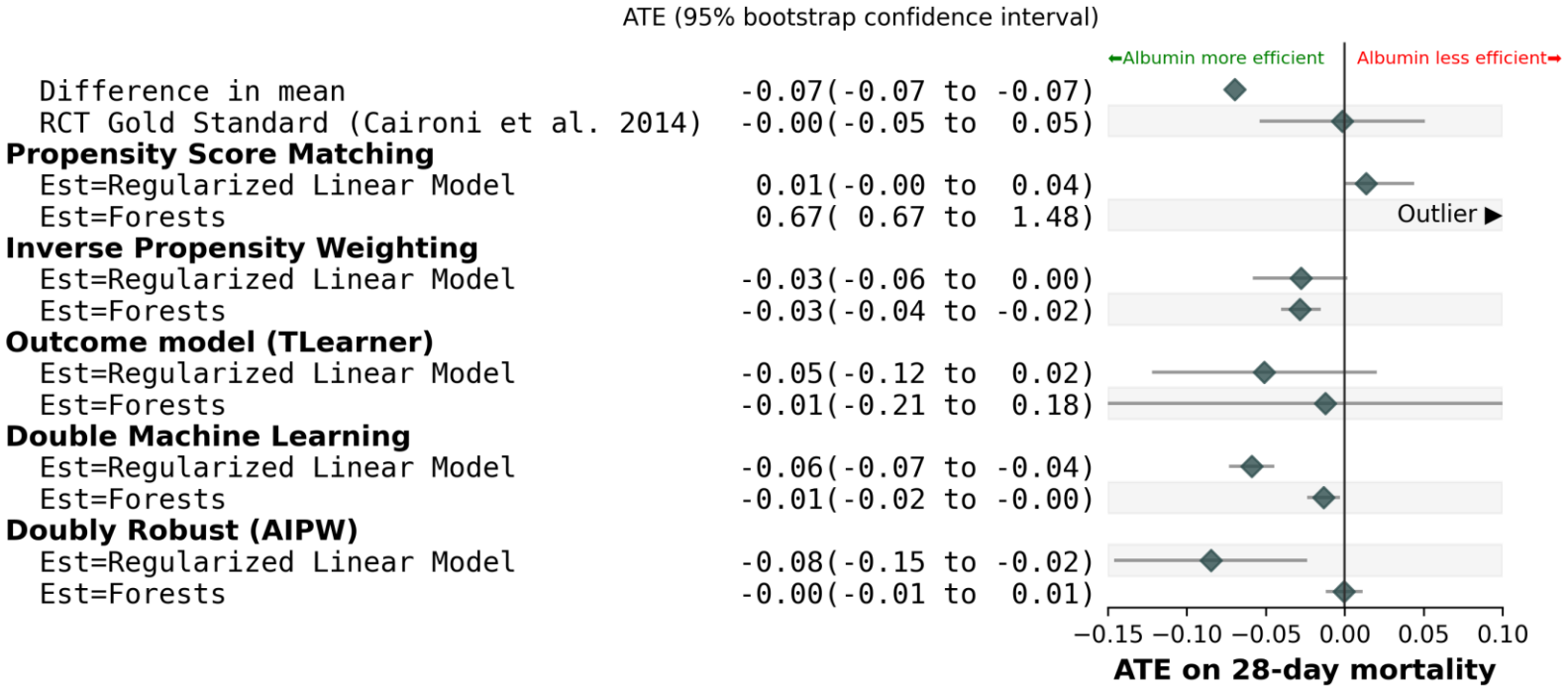
	ATE (95% bootstrap confidence interval)
RCT Gold Standard (Caironi et al. 2014)	-0.00(-0.05 to 0.05)
All confounders (24 features)	-0.00(-0.01 to 0.01)
Without drugs (19 features)	-0.01(-0.02 to 0.00)
Without measurements (13 features)	-0.02(-0.04 to -0.00)
Socio-demographics (5 features)	-0.03(-0.05 to -0.02)
Unajusted risk difference	-0.07(-0.07 to -0.07)

Less and less informed confounder sets



An imperfect DAG including the main confounders still reduces bias

Step 3 – Compare different causal and statistical estimators



Random forests estimators and doubly robust methods retrieve the true effect

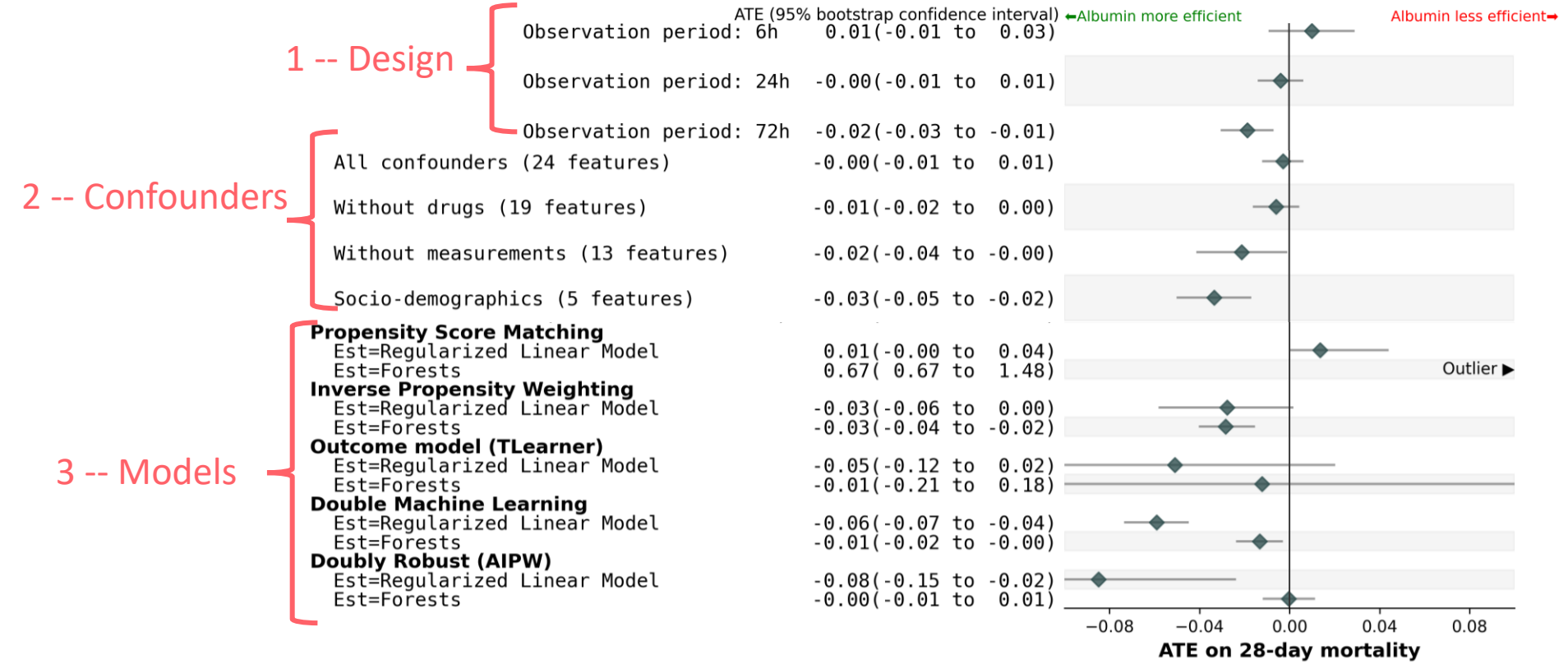
The choice of the causal estimator is important

Step 4 – Comparing the biases of all three steps

1 -- Design

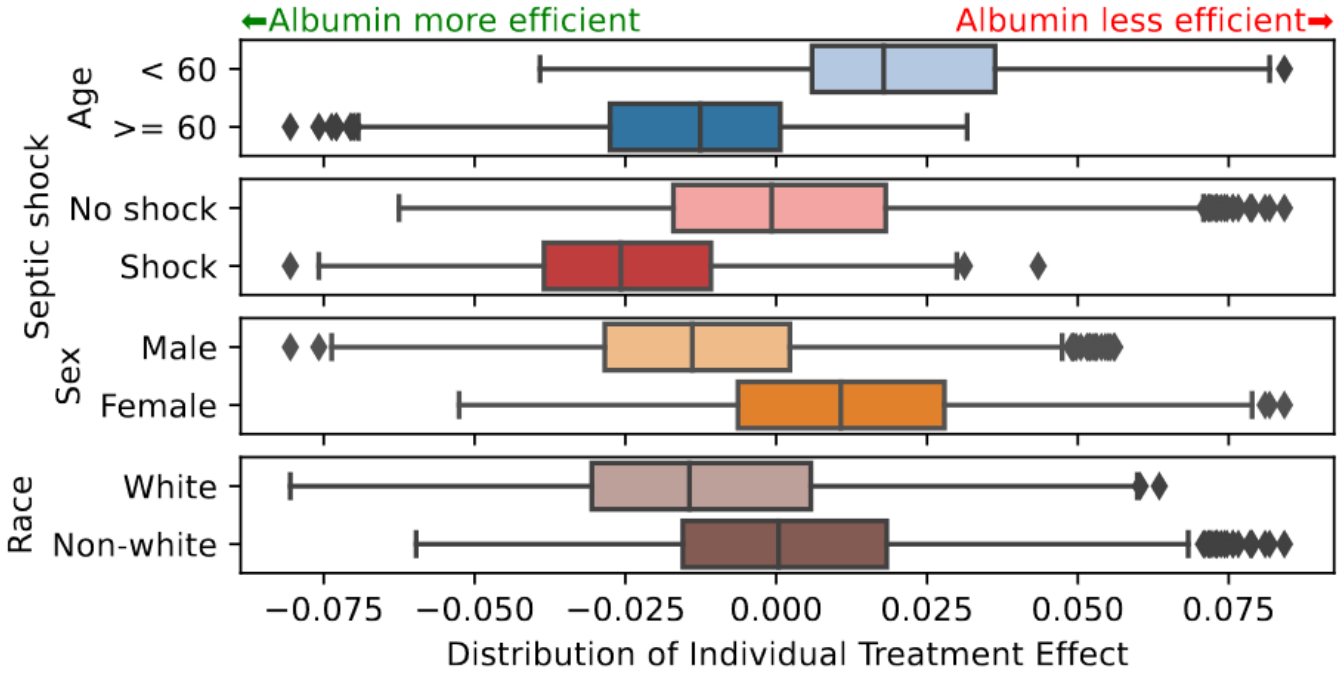
2 -- Confounders

3 -- Models



All steps are equally important for reasonnable analytical choices.

Step 5 – Beyond population effect: Heterogeneity of effect



Causal inference can suggest respondant subpopulations for tailored interventions

How to build robust decision-making algorithm from routine care data?

💡 Study design, confounders and estimator choices are **all equally important** to reduce bias

💡 Valid conclusions can be obtained even if one of this step is not perfect, but **ignoring completely one of them endangers the study validity**

💡 **Adjusting the parameters** thanks to a vibration analysis and a **gold standard trial** allows to:

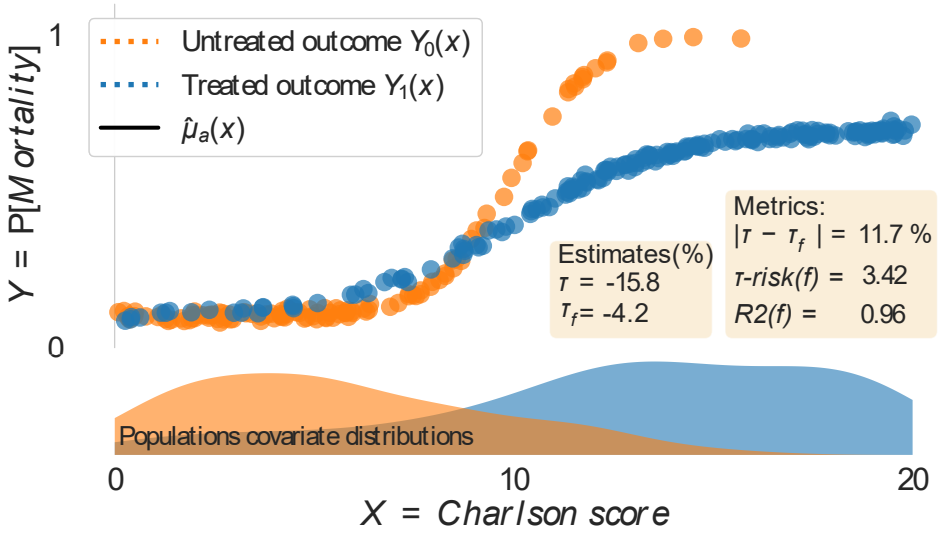
- catch some biases
- study the heterogenity of the effect

How to select predictive models for tailored decision making?

Model selection: Toy example

Select model with small error between the outcome and the prediction on Out-Of-Samples

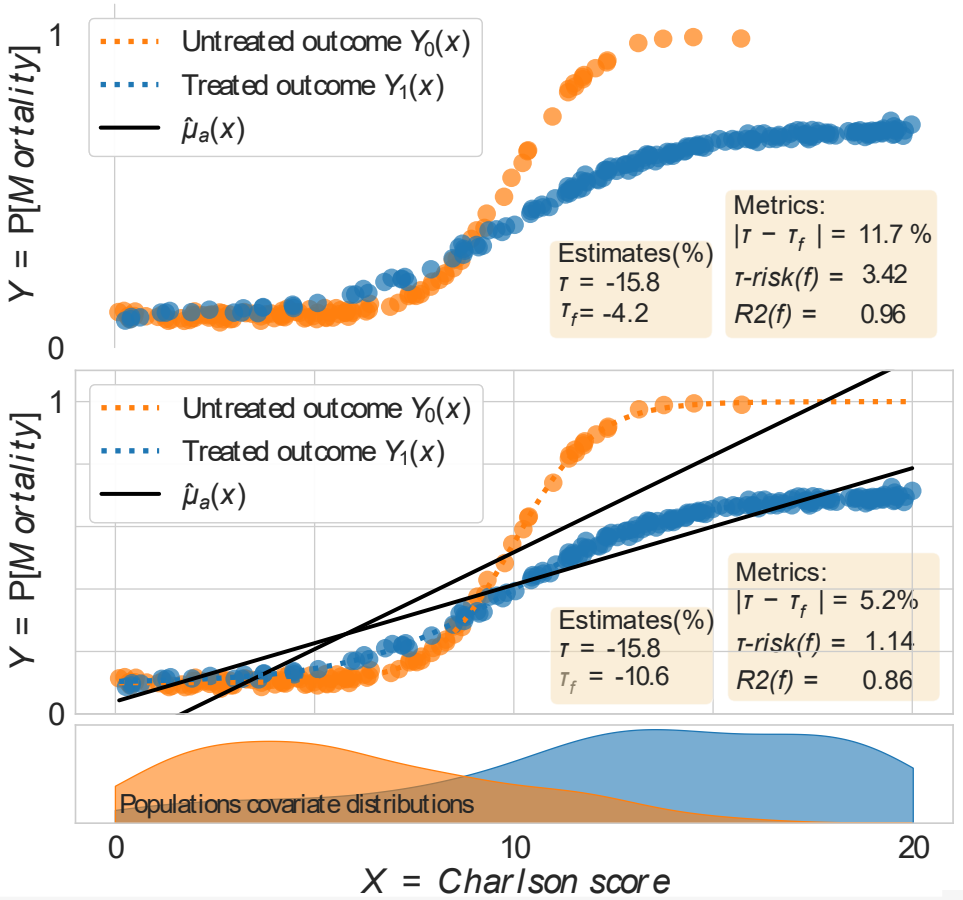
- **Random Forest**
  - **Almost perfect prediction (R^2)**
  - **Bad effect estimation (τ-*Risk)***



Model selection: Toy example

Select model with small error between the outcome and the prediction on Out-Of-Samples

- **Random Forest**
  - **Almost perfect prediction (R^2)**
  - **Bad effect estimation (τ-*Risk*)**
- **Linear model**
  - **Worse prediction (R^2)**
  - **Good effect estimation (τ-*Risk*)**



An estimator can give a good estimate of the effect but predict poorly the outcome

Explore different metrics for model selection

Metric	Equation
$mse(\tau(x), \tau_f(x)) = \tau\text{-risk}(f)$	$\mathbb{E}_{x\sim p(X)}[(\tau(x) - \hat{\tau}_f(x))^2]$ *Oracle: Not observable*
$mse(y, f(x)) = \mu\text{-risk}(f)$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[(\ y\ -\ f(x;a)\ )^2\right]$ *Machine learning: Mean Squared Error*

Explore different metrics for model selection

Metric	Equation
$mse(\tau(x), \tau_f(x)) = \tau\text{-risk}(f)$	$\mathbb{E}_{x\sim p(X)}[(\tau(x) - \hat{\tau}_f(x))^2]$
$mse(y, f(x)) = \mu\text{-risk}(f)$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[\left(\ y\ -\ f(x;a)\ \right)^2\right]$
$\mu\text{-risk}^*_{IPW}$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[\left(\frac{a}{e(x)} + \frac{1-a}{1-e(x)}\right)\left(\ y\ -\ f(x;a)\ \right)^2\right]$
$R\text{-risk}^* = \tau\text{-risk}_R$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[\left(\left(\ y\ -\ m(x)\ \right) - (a - e(x))\ \hat{\tau}_f(x)\ \right)^2\right]$

Explore different metrics for model selection

Metric	Equation
$mse(\tau(x), \tau_f(x)) = \tau\text{-risk}(f)$	$\mathbb{E}_{x\sim p(X)}[(\tau(x) - \hat{\tau}_f(x))^2]$
$mse(y, f(x)) = \mu\text{-risk}(f)$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[(y - f(x;a))^2\right]$
$\mu\text{-risk}^*_{IPW}$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[\left(\frac{a}{e(x)} + \frac{1-a}{1-e(x)}\right)(y - f(x;a))^2\right]$
$R\text{-risk}^* = \tau\text{-risk}_R$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}[((y - m(x)) - (a - e(x))\,\hat{\tau}_f(x))^2]$

New theoretical result:

Propensity weight

$$R\text{-risk}^*(f) = \int_x \mathrm{e(x)(1\text{-}e(x))}\,(\tau(x) - \tau_f(x))^2 p(x)dx + \tilde{\sigma}^2_B$$

Oracle metric

Bayes error (noise)

The R-risk is a weighted version of the oracle metric

Explore different metrics for model selection

Metric	Equation
$mse(\tau(x), \tau_f(x)) = \tau\text{-risk}(f)$	$\mathbb{E}_{x\sim p(X)}[(\tau(x) - \hat{\tau}_f(x))^2]$
$mse(y, f(x)) = \mu\text{-risk}(f)$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[(\ y\ -\ f(x;a)\ )^2\right]$
$\mu\text{-risk}^*_{IPW}$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[\left(\frac{a}{e(x)} + \frac{1-a}{1-e(x)}\right)(\ y\ -\ f(x;a)\ )^2\right]$
$R\text{-risk}^* = \tau\text{-risk}_R$	$\mathbb{E}_{(y,x,a)\sim\mathcal{D}}\left[\left(\left(\ y\ -\ m(x)\right) - (a - e(x))\ \hat{\tau}_f(x)\ \right)^2\right]$

How well a metric is ranking different treatment effect models compared to the **oracle** *τ-Risk?*

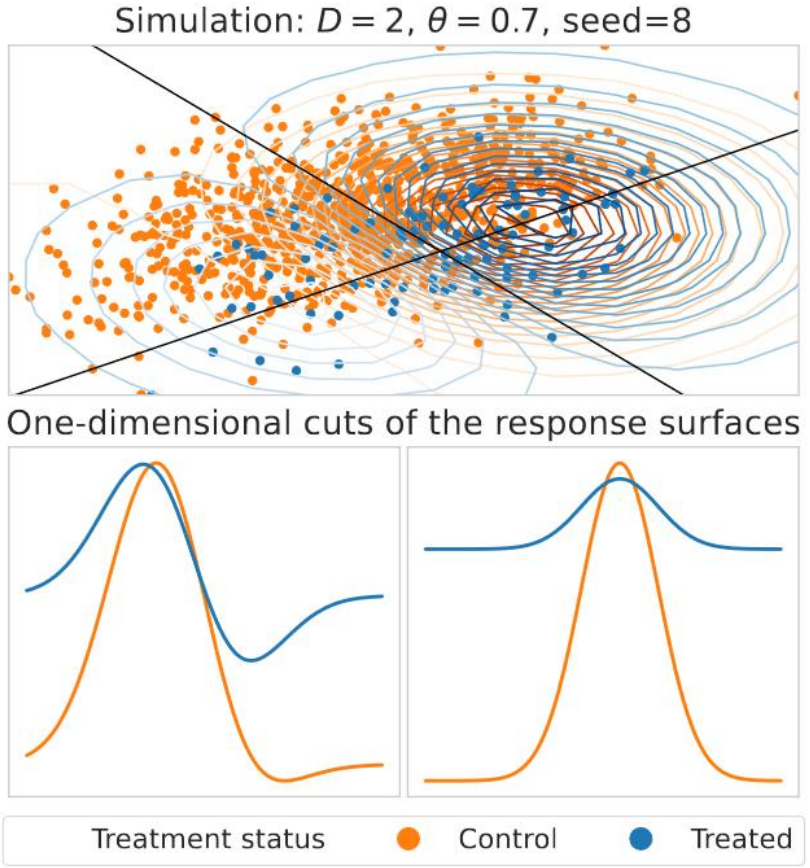
Empirical study

Simulated dataset, *Caussim*:

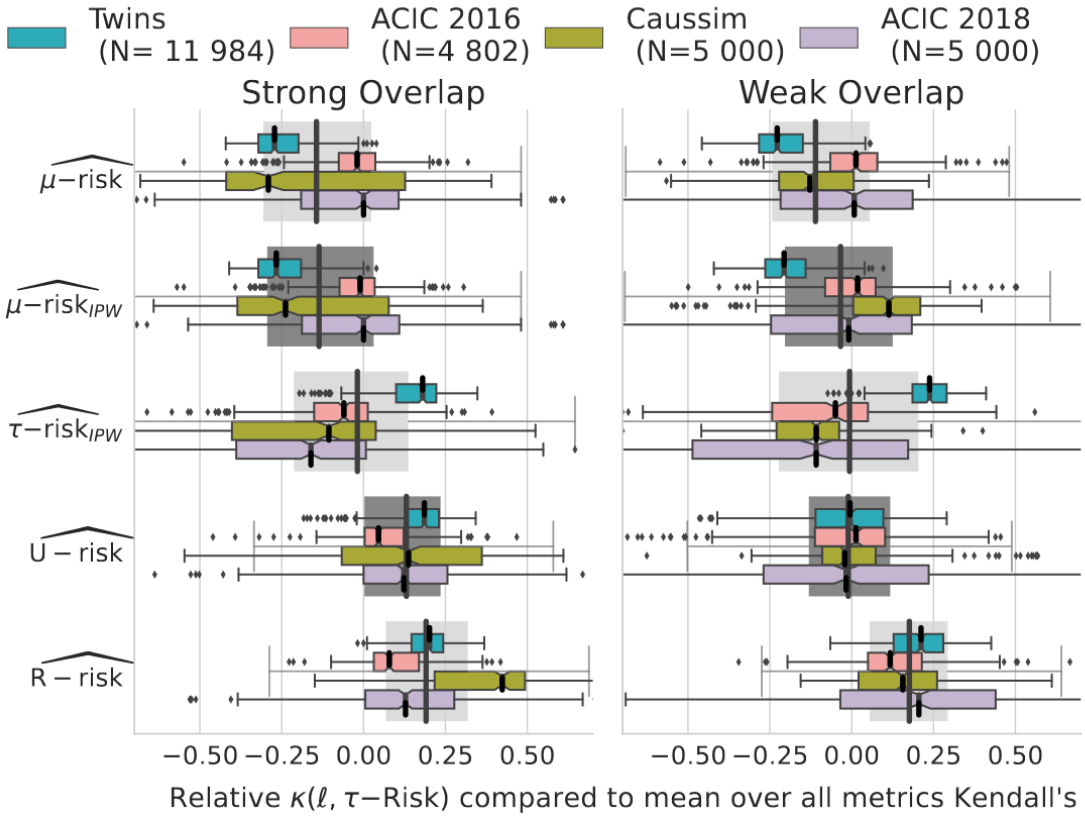
- **Covariates** with basis extension
- **Overlap** between treated and controls
- **Potential outcomes Y(0) and Y(1)**

Three semi-simulated datasets used in the causal inference literature:
Real covariates, simulated treatment and outcomes

- *ACIC2016*
- *ACIC2018*
- *Twins*



Empirical study



The *R-risk* is the best metric for model selection: estimating nuisances is beneficial

How to select predictive models for tailored decision making?

- Selecting a model for intervention should **use a different metric than for prediction**
- The R-risk is a **reweighted version of the oracle metric**
- **Estimation of nuisances** reduces bias for many different settings

Some perspectives

Where do we need to improve?

1) **Framing – study design:** Data quality must improve. MIMIC-IV is not perfect but good enough thanks to open documentation, easy access and a strong link with the data collection since 20+ years.

2) **Identification – list confounders:** Medical and statistical expertise required. Bring together the different communities with events focused on practical questions.
Matos, J., et al. MIT Critical Datathon 2023: a MIMIC-IV Derived Dataset for Pulse Oximetry Correction Models.

3) **Estimation:** Many existing methods. Text-based models have interesting potentials.
Jiang, L. Y., et al. (2023). Health system-scale language models are all-purpose prediction engines. Nature

4) **Vibration analysis:** Elaborate models require huge amounts of compute. If this is the right direction, we need to change the collocation of compute and data.
Jiang, L. Y., et al. (2023) used one of the biggest computing cluster of the east cost.

5) **Conditional Average Effect:** Great opportunity for research. Nested trials will bring interesting insights.
Dahabreh, I. J., & Hernán, M. A. (2019). Extending inferences from a randomized trial to a target population. European journal of epidemiology.

Where could causal inference methodology benefit the most?

- **Not suitable for evaluating drug efficacy**
  Less robust than randomized control trials (maybe for drug life cycle)

- **Interesting to evaluate cares with poor fundings for trials**
  Such as public health interventions or procedures
  (national claims might be relevant)

- **Improve machine learning with causal reasoning**
  Identify responders and design tailored care pathways

🥹Thank you for your attention !

I. Exploring a complexity gradient in representation and predictive models for EHRs

II. Prediction is not all we need: Causal thinking for decision making on EHRs
M. Doutreligne, T. Struja, J. Abecassis, C. Morgand, L Celi, G. Varoquaux, https://arxiv.org/abs/2308.01605

III. How to select predictive models for causal inference?
M. Doutreligne, G. Varoquaux, https://arxiv.org/abs/2302.00370

IV. Potential and challenges of Clinical Data Warehouse, a case study in France
M. Doutreligne , A. Degremont, P.A. Jachiet, A. Lamer, X. Tannier
https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000298

Supplementary slides for motivation

Worldwide initiatives to collect, organize and study health data

- German Medical Informatics Initiative, 2016

- English NHS funded OpenSAFELY platform, 2020

- US NIH funded CHoRUS network, 2022

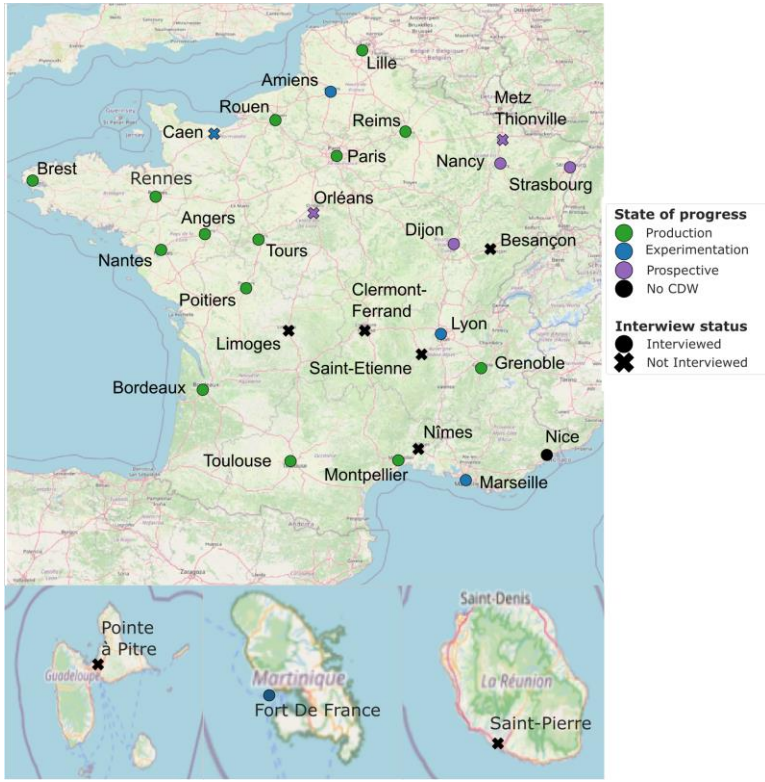
Worldwide initiatives to collect, organize and study health data

- German Medical Informatics Initiative, 2016

- English NHS funded OpenSAFELY platform, 2020

- US NIH funded CHoRUS network, 2022

- French fundings for Clinical Data Warehouse , 2023



***Doutreligne, M.**, Degremont, A., Jachiet, P. A., Lamer, A., & Tannier, X. (2023). Good practices for clinical data warehouse implementation: A case study in France.*

What interventions are the most effective with constrained medical resources?

RCTs are costly and study of supopulations is difficult due to small samples

"Fewer than half of the clinical guidelines for the nine most common chronic conditions consider older patients with multiple comorbid chronic conditions." *(Parekh; Barton, 2010)*

RCTs measure **efficacy** (ideal conditions) rather than **effectiveness** (usual practices)

"Only 6% of asthmatics would have been eligible for their own treatment RCTs " *(Travers et al., 2007)*

Is it possible to leverage routine care data to complement available evidence?

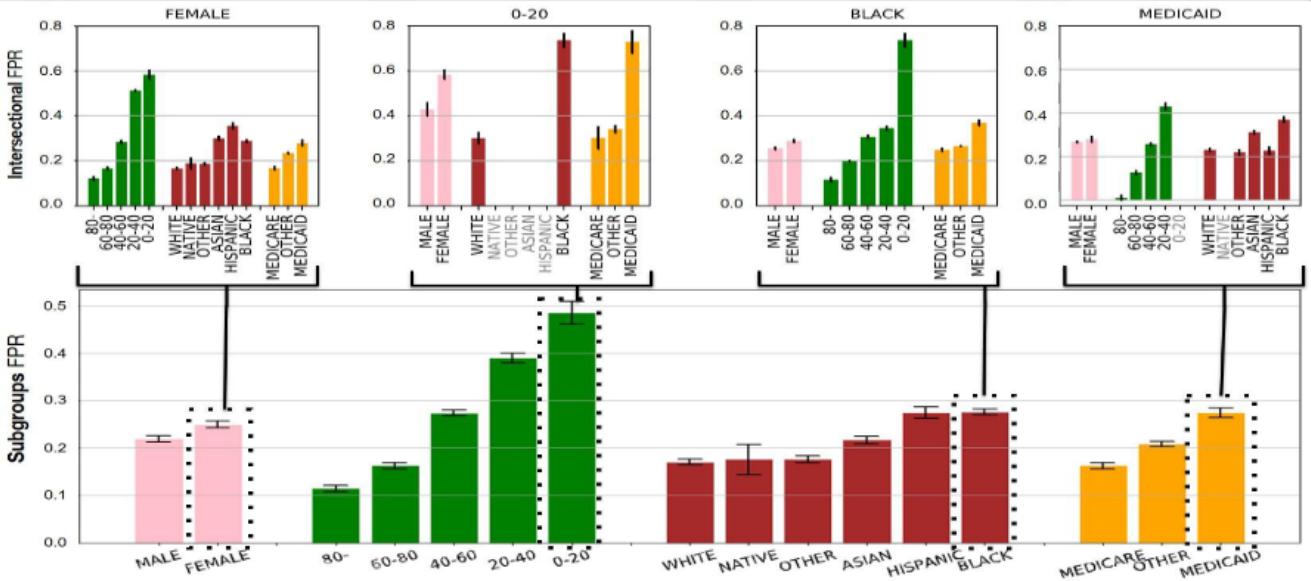
- A. K. Parekh; M. B. Barton (2010): "The challenge of multiple comorbidity for the US health care system". Jama
- J. Travers, S. Marsh, M. Williams, M. Weatherall, B. Caldwell, P. Shirtcliffe, S. Aldington; R. Beasley (2007): "External validity of randomised controlled trials in asthma: to whom do the results of the trials apply?" In: Thorax

Other failure modes of machine learning…
eg. Exclusion of under-served populations for chest X-ray diagnosis

Automating CheXclusion With EHR + ML

Largest underdiagnosis rates in:

- Female
- 0-20
- Black
- Medicaid insurance



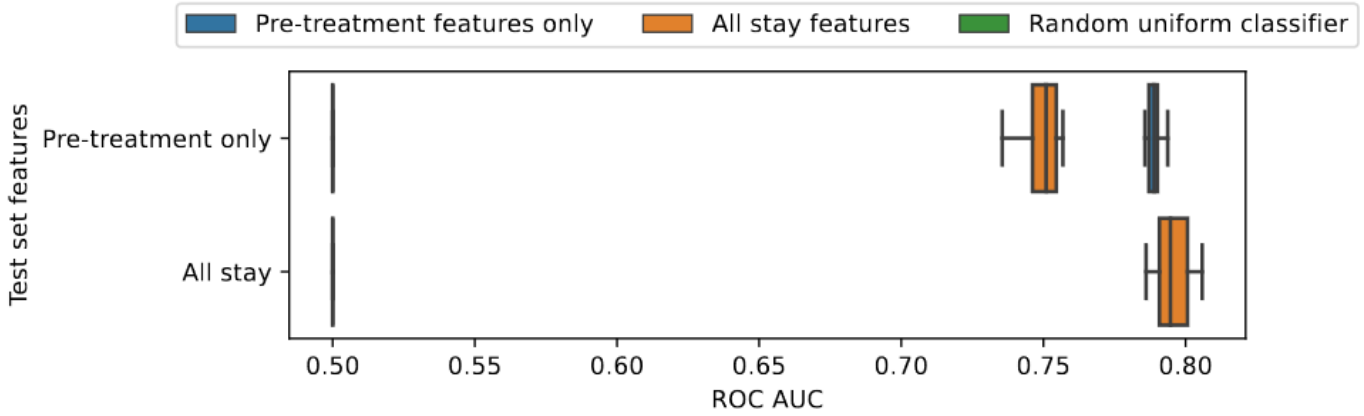
Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi.
"Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations" Nature Medicine 2021.

Yet failure modes: example in intensive care

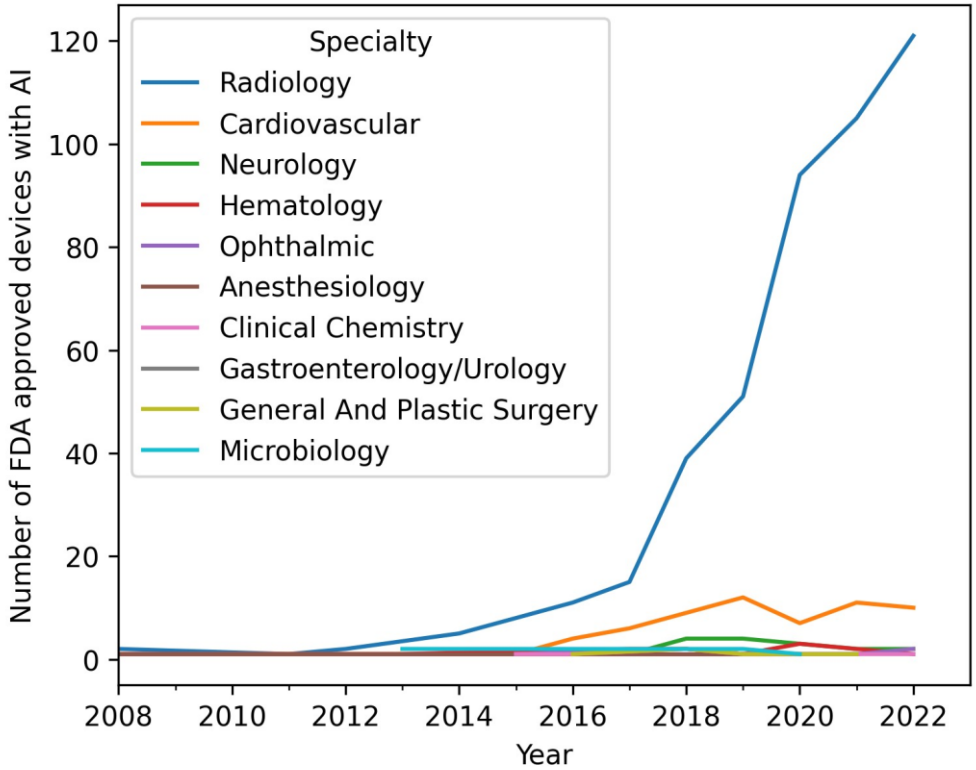
- Predict 28-day mortality, *interested in fluid rescusitation treatment*
- Train with post-treatment variables
- Evaluate on a clinically useful dataset with only pre-treatment variables

Yet failure modes: example in intensive care

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

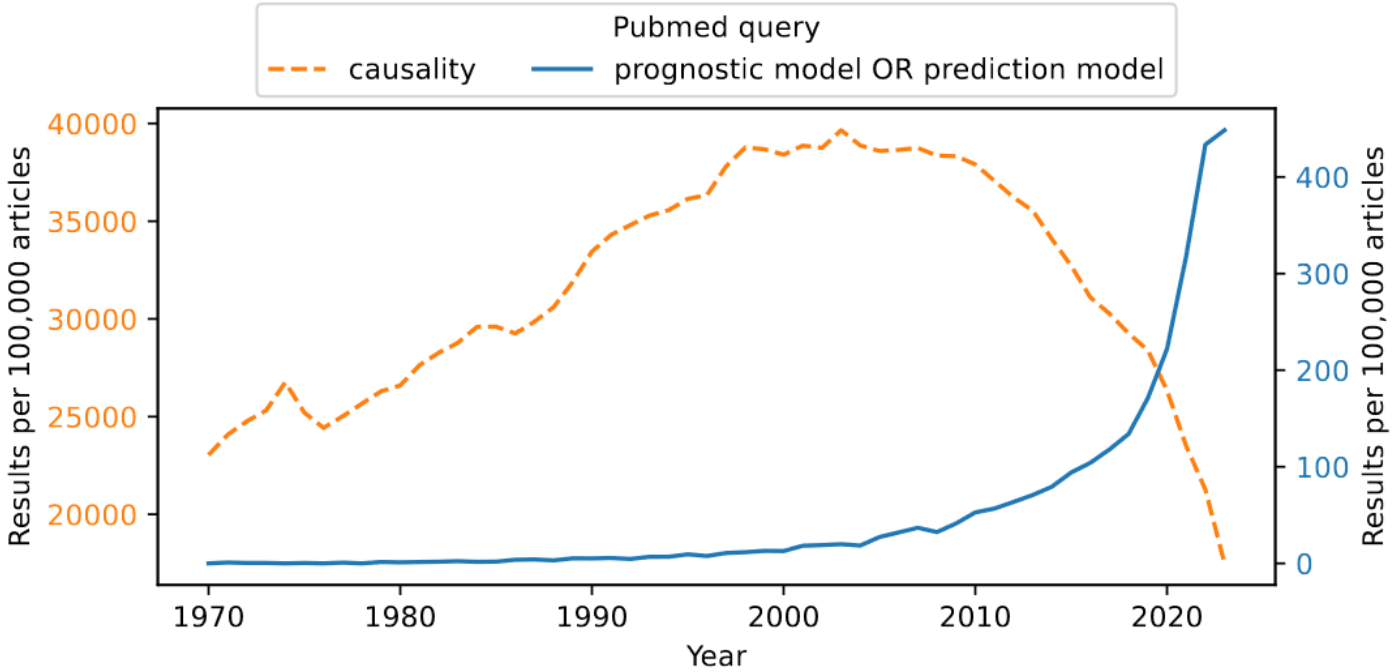


Slow adoption of medical devices with AI outside radiology



Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ digital medicine, 3(1), 118.

Prediction or causation



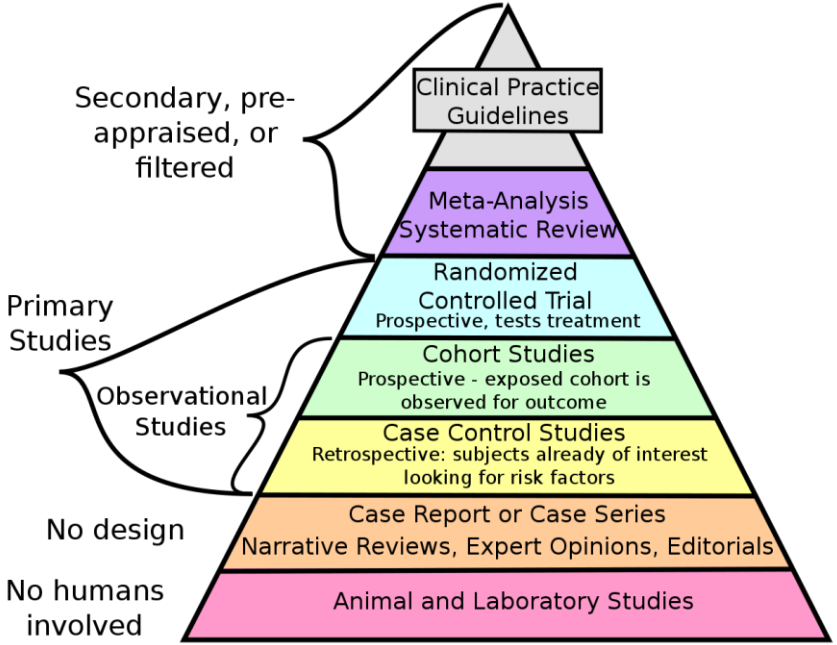
Proportion of articles by year in Pubmed returned by queries on causality or predictive modeling.

What interventions are the most effective with constrained medical resources?

First, measure efficacy (ideal conditions)

Medical guidelines built thanks to the scientific literature to recommend ideal care trajectories

Higher degree of evidence relies on meta-analyses of Randomized Controlled Trials (RCTs)

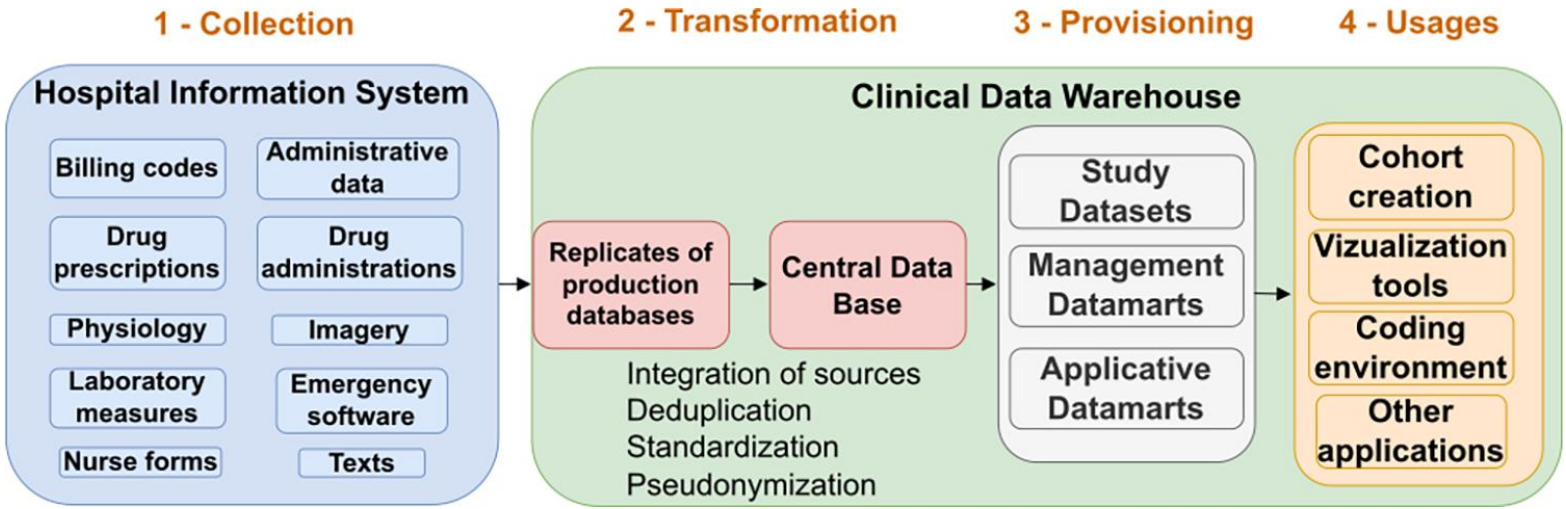


Hierarchy of evidences

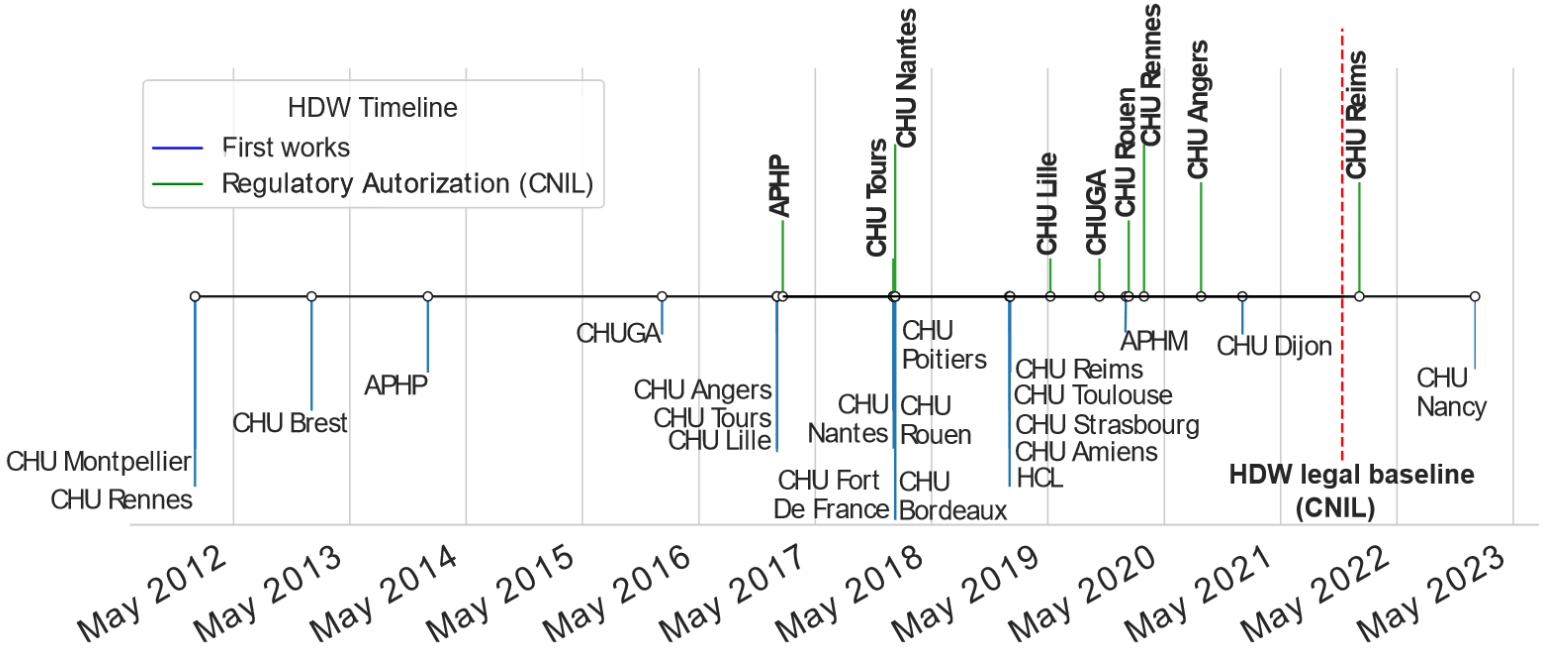
Supplementary slides for: Clinical Data Warehouse

Clinical data warehouse

Technical and organizational infrastructures pooling data from several medical information systems to homogeneous formats, for management, research or care reuses



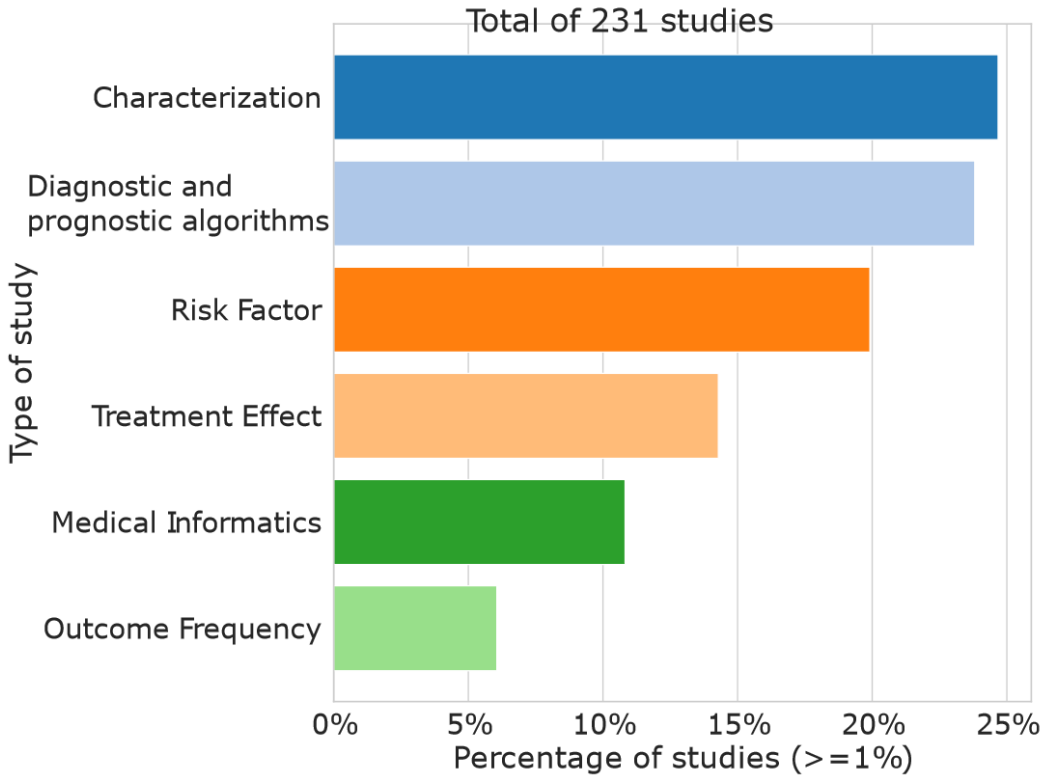
Timeline of CDWs implementation in french university hospitals



Type of data in CDWs

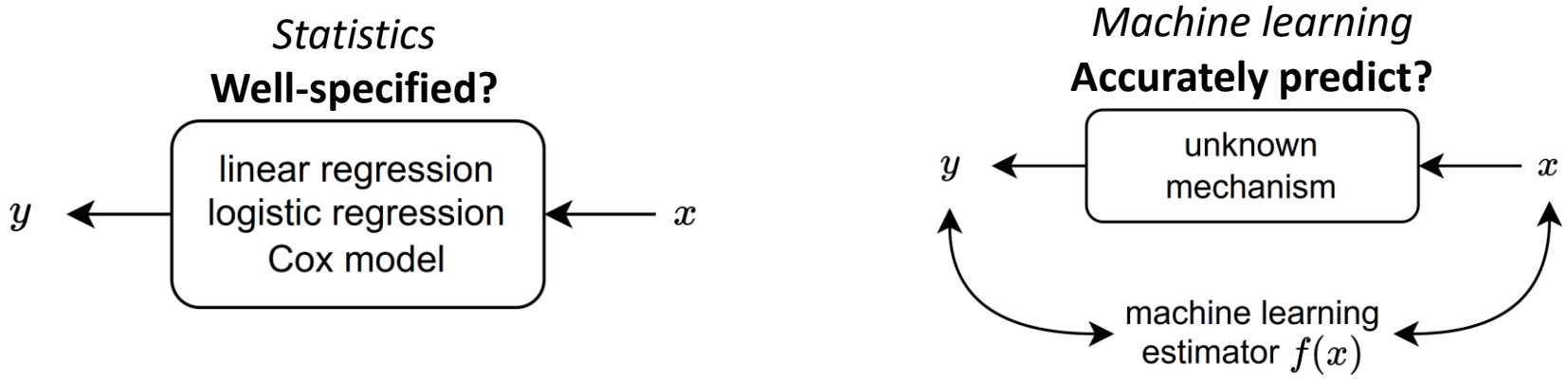
Category of data	Number of CDW	Ratio
Administrative	21	100 %
Billing Codes	20	95 %
Biology	20	95 %
Texts	20	95 %
Drugs	16	76 %
Imagery	4	19 %
Nurse Forms	4	19 %
Anatomical pathology	3	14 %
ICU	2	10 %
Medical devices	2	10 %

Objective of studies

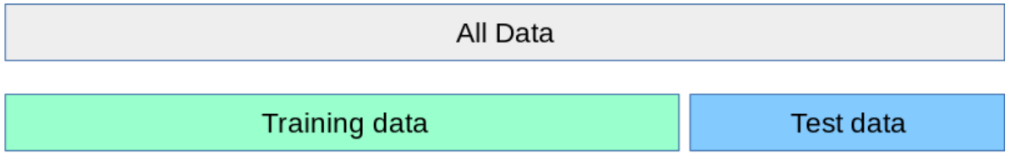


Supplementary slides for: predictive algorithms for EHR

Machine learning *estimators* are selected for their predictive accuracy



Out-of-samples validation avoids overfitting



L. Breiman, 2001, Statistical modeling:The two cultures (with comments and a rejoinder by the author)". Statistical science

Machine learning is not concerned with the data generation mechanism

Machine learning predicts well in multiple domains

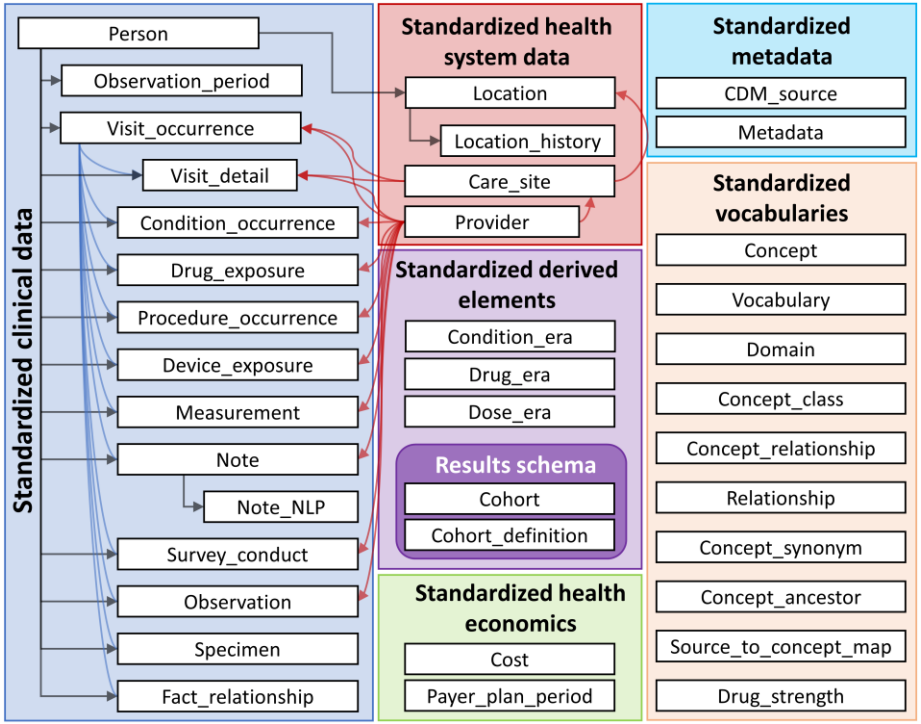
Images

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Text

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Routine care data?



OMOP standard data model

What is the complexity of this healthcare data?

Machine learning literature for routine care data

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˜	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.

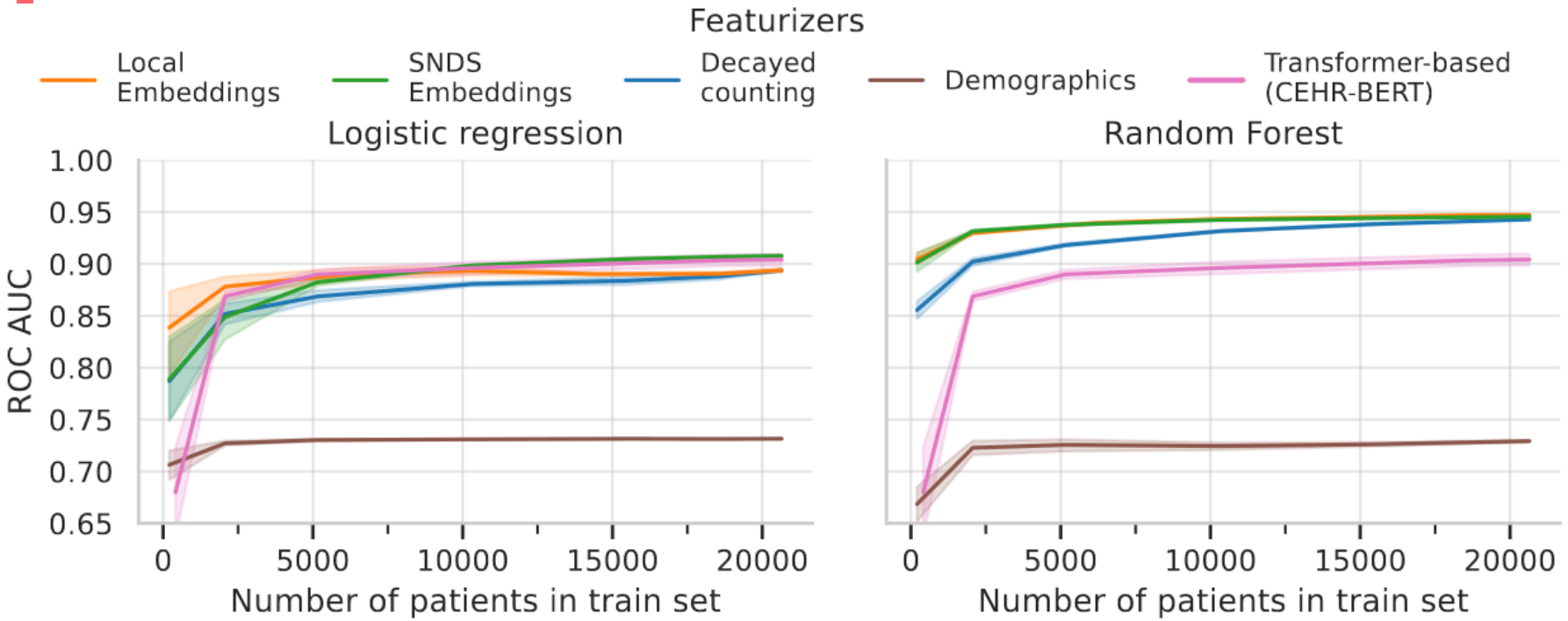
Source: **High-performance medicine: the convergence of human and artificial intelligence** Eric Topol, Nature Medicine Jan 2019

Machine learning predicts well a variety of medical endpoints

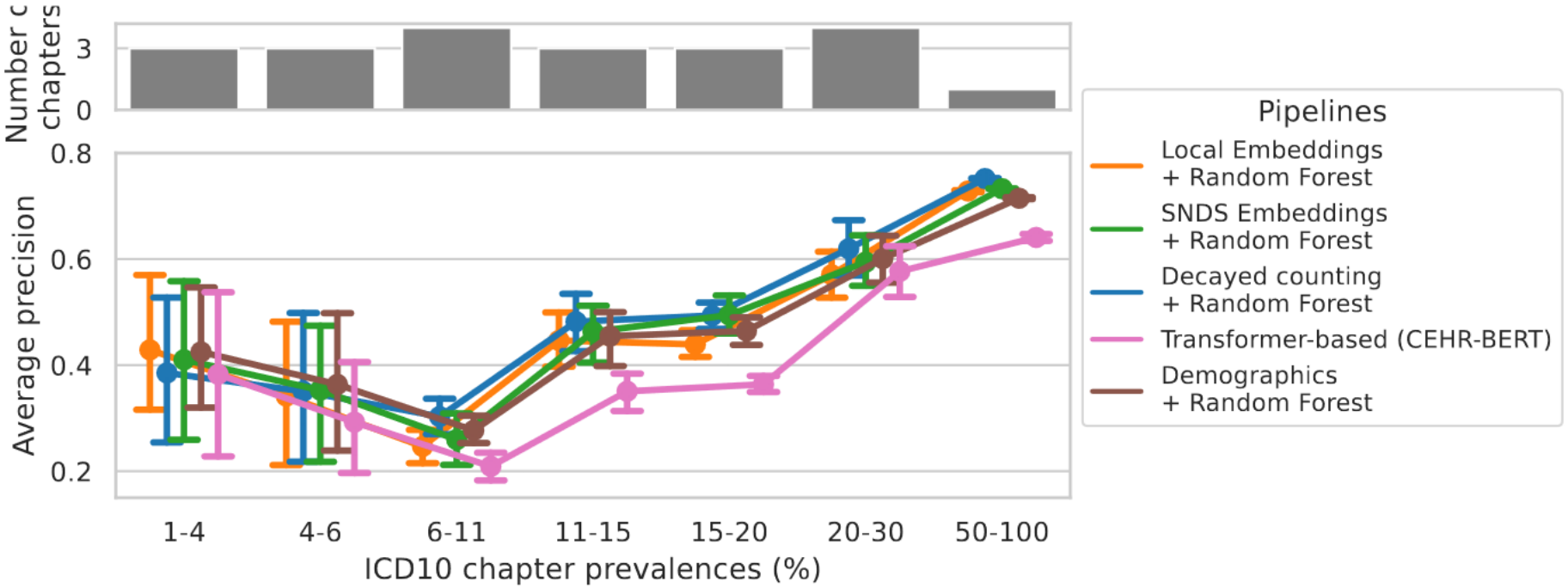
Length of stay results

	Long length of stay
Description	Long stay classification (longer than 7 days)
Task	Binary classification
Cohort Size	27,053
Prevalence	23.1%
Number of cases	6,249

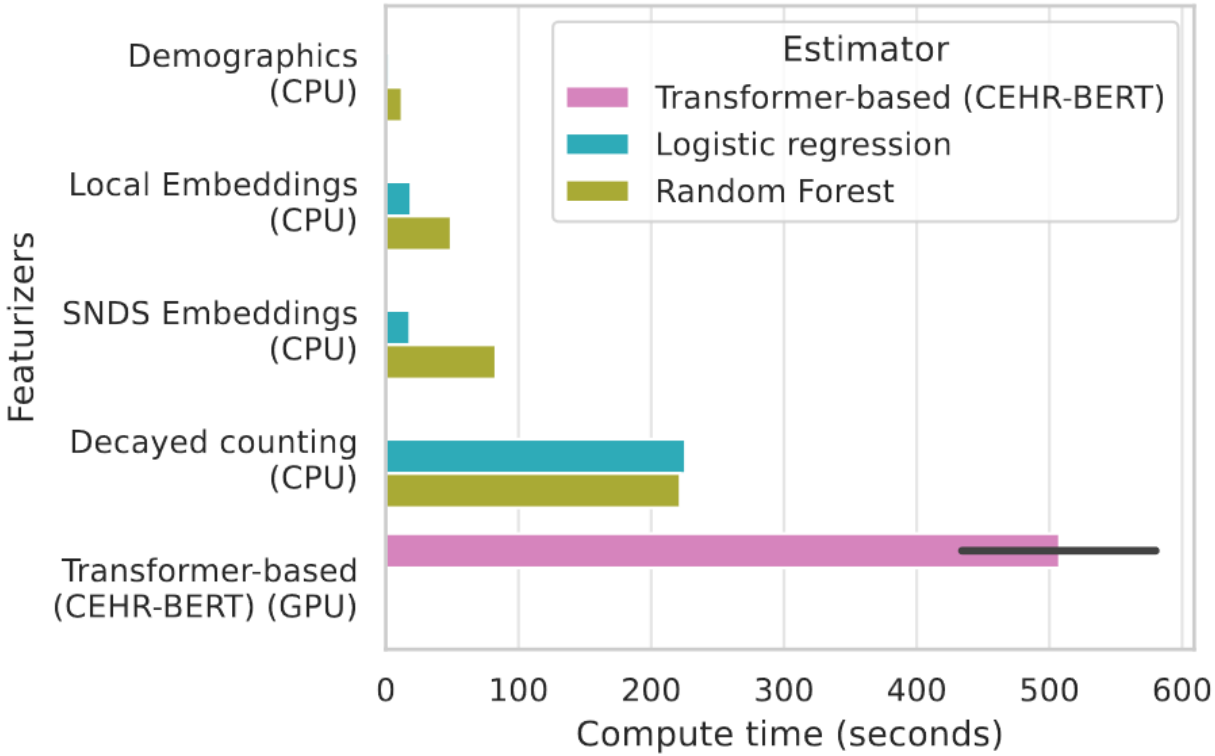
Length of stay results



High number of cases matters



Static embeddings are quicker to train



The challenge of low prevalence
Health data is big data?

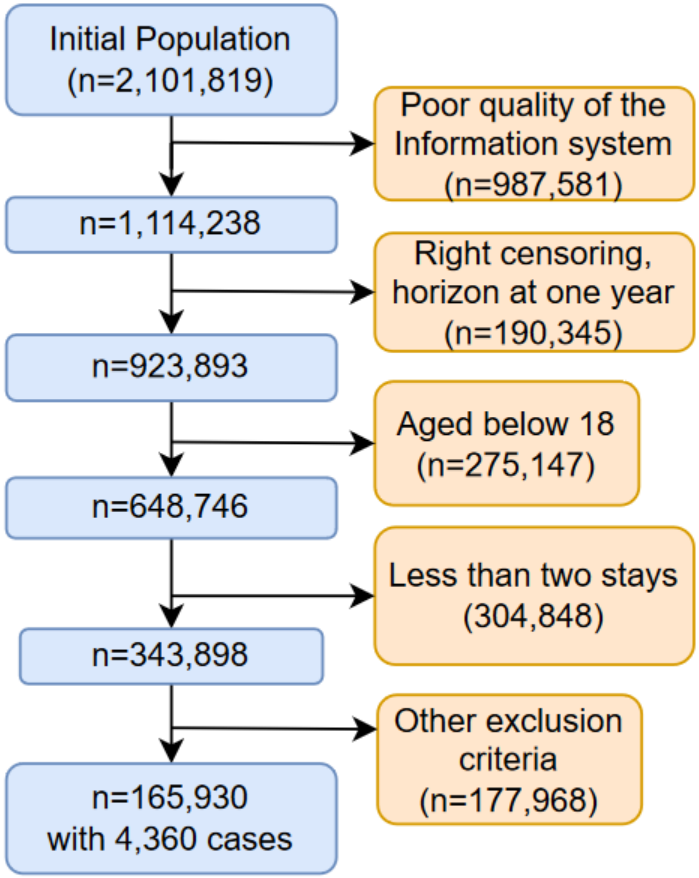
A lot of sample « losses » due to:

- Inclusion criteria

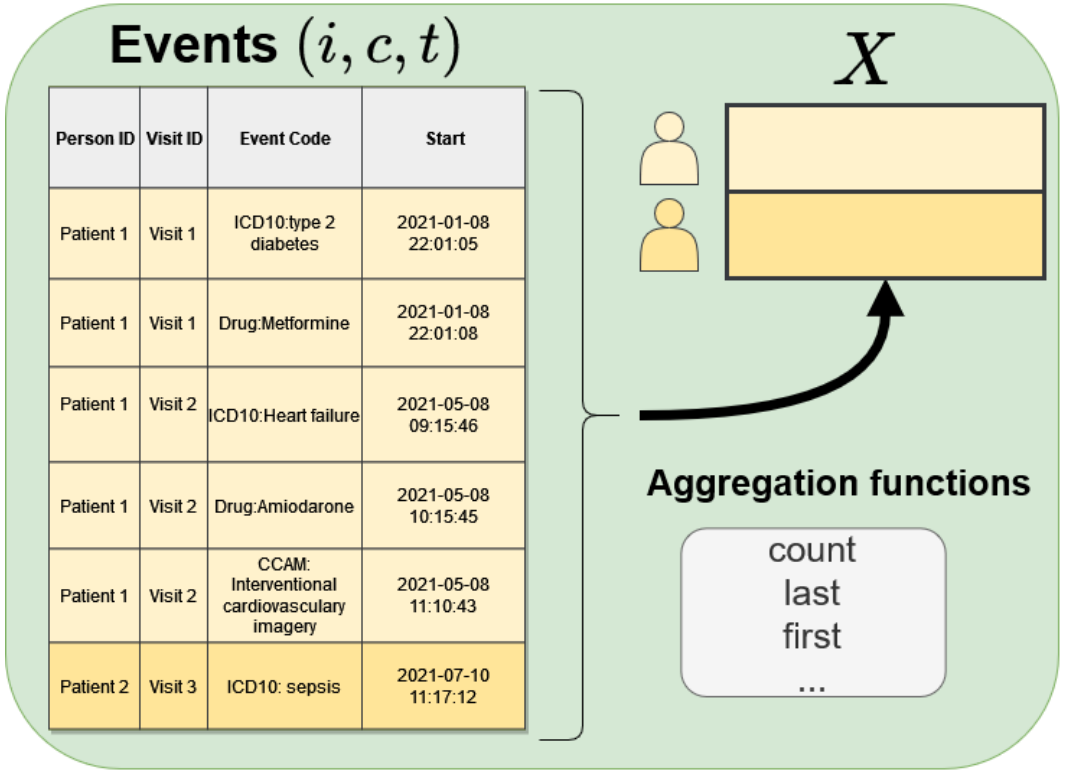
- Information system instability

- Censoring

- Rare outcomes

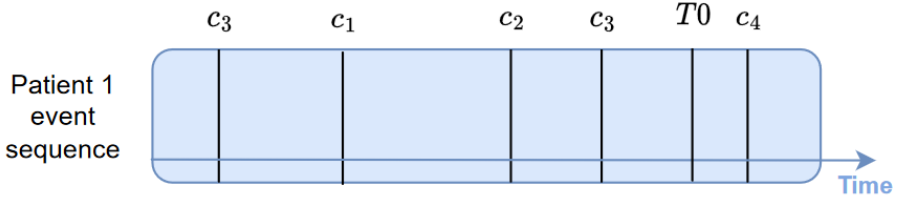


Engineering focus: chain patient representation and predictive model



Computing patient features : count and decay

Patient features without time decay

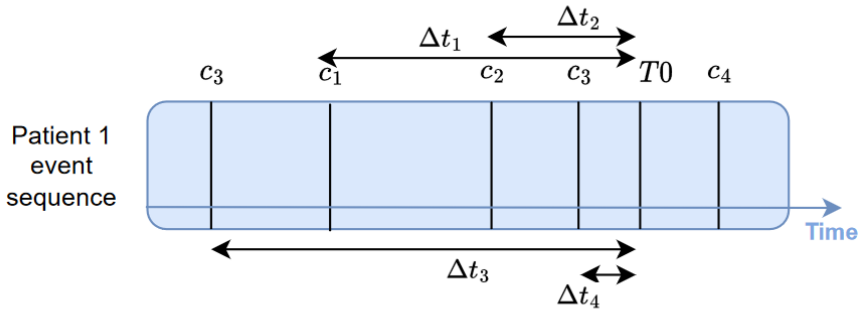


$$C(1, c_1) = 1$$
$$C(1, c_2) = 1$$
$$C(1, c_3) = 2$$
$$C(1, c_4) = 0$$

1	**1**	**3**	**0**
0	1	2	1
1	0	1	0
...	...	...	...

N patients

Patient features with time decay δ



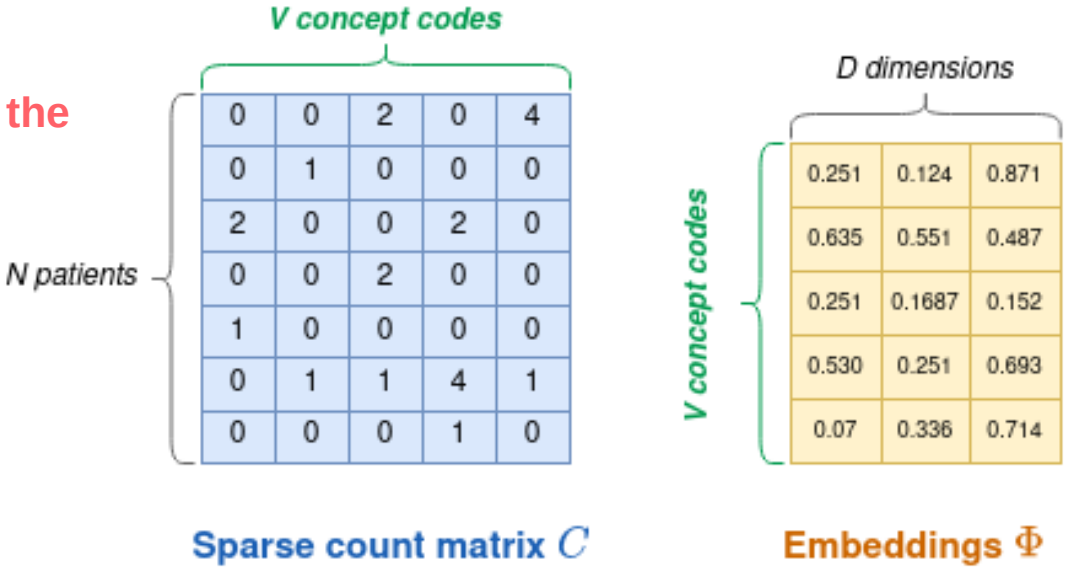
$$C(1, c_1) = \exp\left(-\frac{\Delta t_1}{\delta}\right)$$
$$C(1, c_2) = \exp\left(-\frac{\Delta t_2}{\delta}\right)$$
$$C(1, c_3) = \exp\left(-\frac{\Delta t_3}{\delta}\right) + \exp\left(-\frac{\Delta t_4}{\delta}\right)$$
$$C(1, c_4) = 0$$

0.02	**0.1**	**0.801**	**0**
0	0.2	0.1	0.001
0.1	0	0.8	0
...	...	...	...

Computing patient features : adding the embeddings

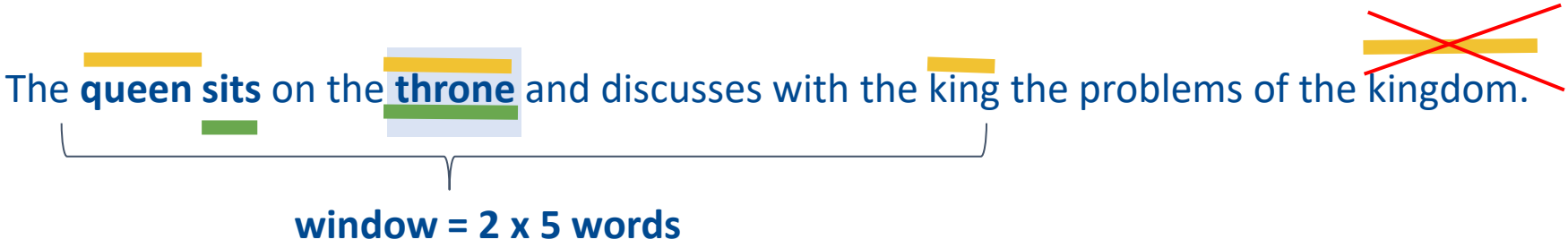
Obtain patient features by collapsing the vocabulary dimension:

$$X = [C \cdot \Phi, C_{decay} \cdot \Phi]$$



Inspirated from word2vec

Distributional hypothesis *(Haris, 1954)*: **Two words have close meaning iif they appear in similar contexts.**



Proximity in the embedding space is forced by **proximity in the corpus**.

Event2vec, a package to compute concept embeddings

A python package available on pypi

- A pyspark version for big data (>500m rows)

- polars for medium sized datasets (up to 100m rows)

- Sklearn compatible transformers

- Quick start and step by step guides: *https://straymat.gitlab.io/event2vec/tutorials/_0_tuto_event2vec.html*

Load events

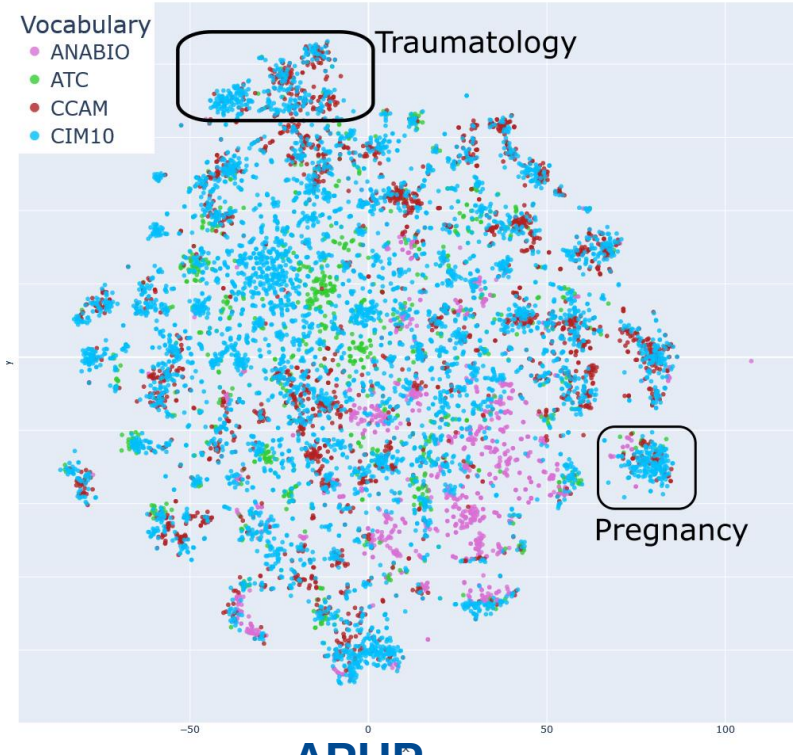
	person_id	start	event_source_concept_id
0	1	2018-11-08 19:24:15	CIM10:N182
4	1	2018-12-20 19:24:15	CCAM:JVJB01
8	2	1993-01-26 07:22:42	CIM10:E12
12	3	2009-04-25 10:14:21	CIM10:N182
9	2	2020-01-26 07:22:42	CIM10:E12

Build embeddings

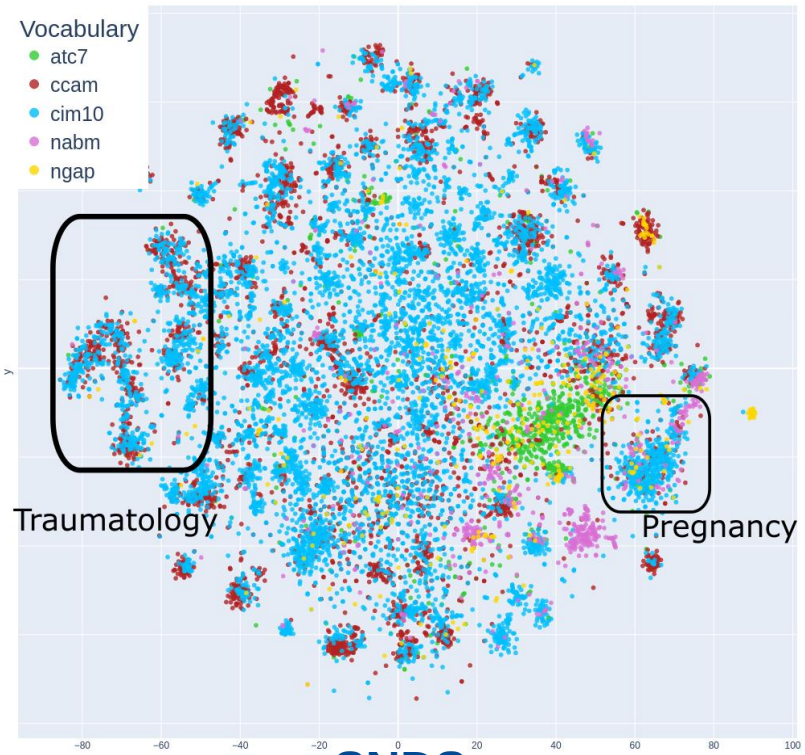
```
alpha = 0.75
k = 1
d = 3

embeddings = event2vec(
    events=events,
    output_dir=output_dir,
    colname_concept="event_source_concept_id",
    window_orientation="center",
    window_radius_in_days=30,
    d=d,
    smoothing_factor=alpha,
    k=k,
    backend="pandas",
)
```

Qualitative results: https://straymat.gitlab.io/event2vec/visualizations.html



APHP
(200K random patients)



SNDS
(3M random patients)

Nearest neighboors

Induced distance in embedding space with cosine similarity

```
# ## Compute the 10 closest concepts
Run Cell | Run Above | Debug Cell
# %%
k = 50
source_concept_code = "I50"  # Heart failure
# For SNDS
top_k_concepts = get_closest_nn(
    source_concept_code=source_concept_code,
    embedding_dict=snds_embeddings,
    concept_labels=concept_labels,
    k=k,
)
```

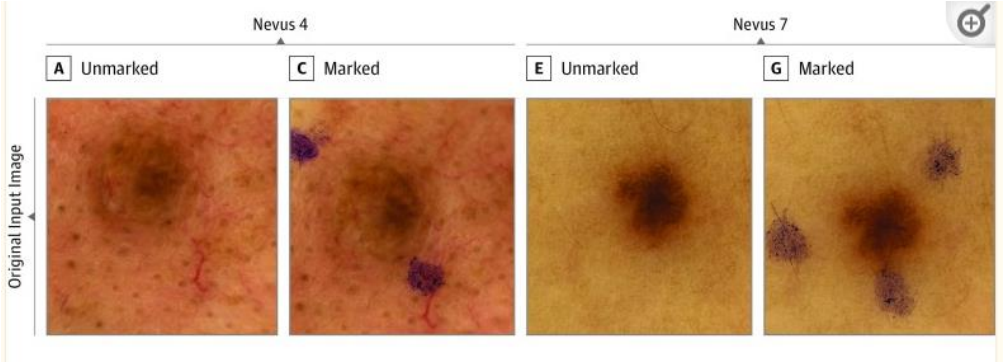
concept_code	concept_name	similarity
cim10:I50	Insuffisance cardiaque	1.000
cim10:I080	Atteintes des valvules mitrale et aortique (rhumatismales)	0.608
cim10:I442	Bloc auriculoventriculaire complet	0.595
atc7:D05BB02	Acitretin	0.582
ccam:YYYY105	Imagerie pour acte de radiologie interventionnelle ou de cardiologie interventionnelle niveau 1, réalisée en salle d'imagerie	0.577
cim10:I071	Insuffisance tricuspidienne (rhumatismale)	0.575
atc7:C01BD01	Amiodarone	0.570
cim10:I341	Prolapsus (de la valvule) mitral(e)	0.570
ccam:DERP003	Choc électrique cardiaque transcutané [cardioversion externe], en dehors de l'urgence	0.567
ccam:DELF007	Implantation souscutanée d'un stimulateur cardiaque définitif, avec pose d'une sonde intraatriale ou intraventriculaire droite par voie veineuse transcutanée	0.564
atc7:C01AA05	Digoxin	0.563
ccam:DEQP008	Restitution tridimensionnelle informatisée de l'activité électrophysiologique cardiaque [cartographie cardiaque tridimensionnelle], au cours d'une intervention sur le système cardionecteur	0.561
cim10:I453	Bloc trifasciculaire	0.555
cim10:I253	Anévrisme du coeur	0.554
ccam:DELF005	Implantation souscutanée d'un stimulateur cardiaque définitif, avec pose d'une sonde intraatriale et d'une sonde intraventriculaire droites par voie veineuse transcutanée	0.543
cim10:I340	Insuffisance (de la valvule) mitrale (non rhumatismale)	0.542

Supplementary slides for: Decision making with EHR

Failures because of shortcut features

Prediction: malignent melanoma
Intervention: excision of nevi

Benign nevi

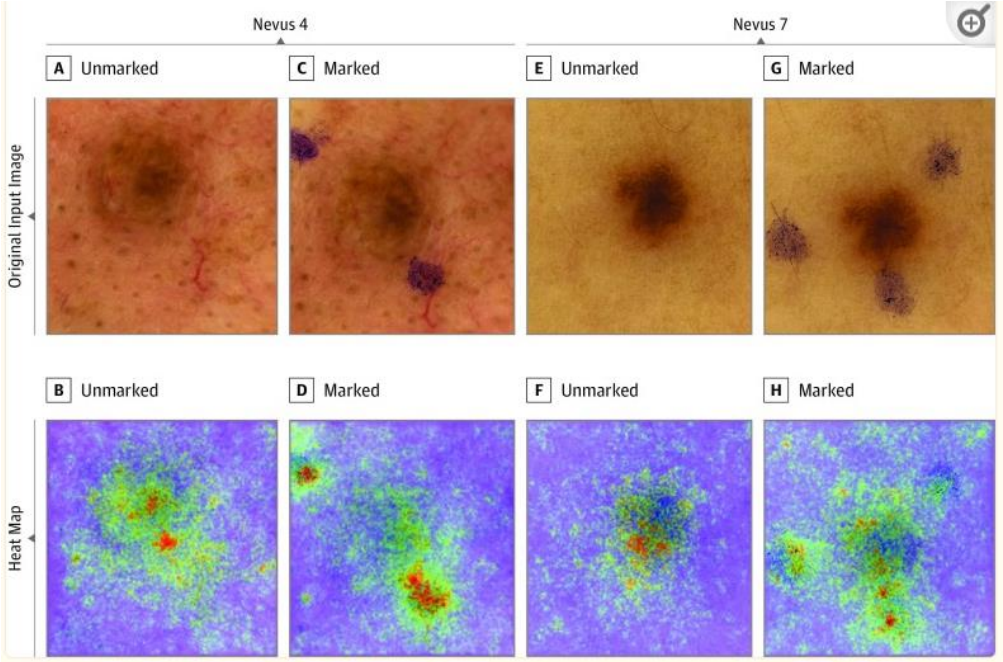


Winkler, Fink, Toberer, Enk, Deinlein, Hofmann-Wellenhof, Thomas, Lallas, Blum, Stolz, et al. (2019). "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition". In: JAMA dermatology

These failures occur because of shortcut features

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

Benign nevi

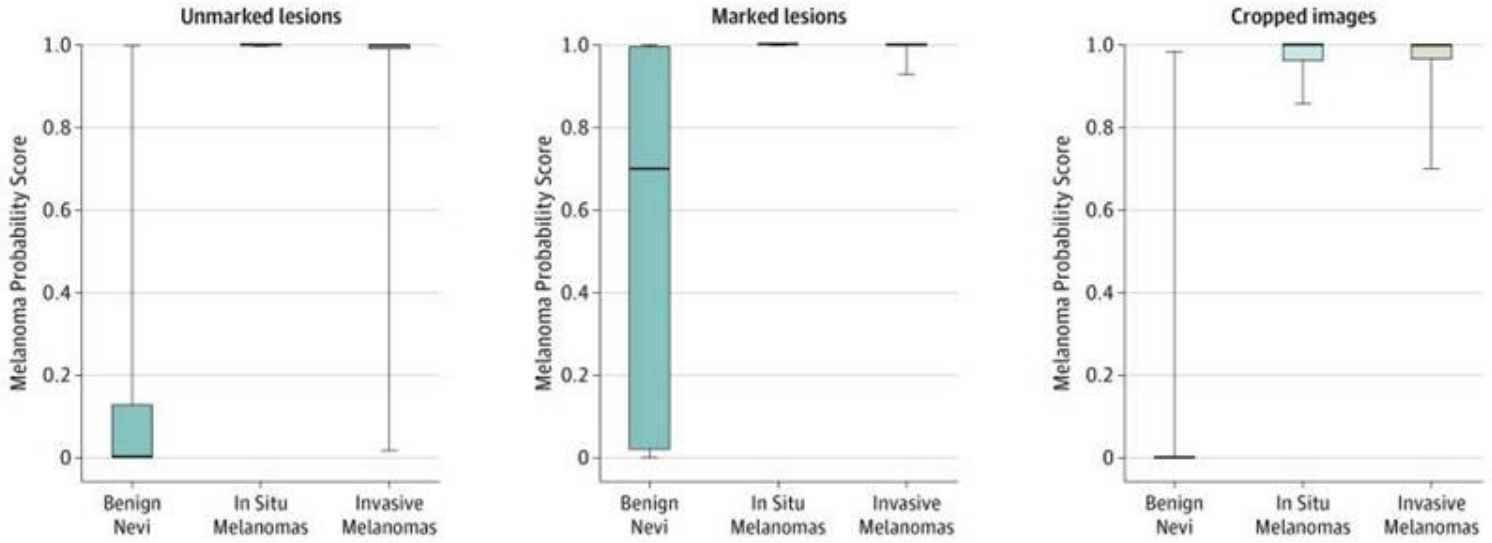


Winkler, Fink, Toberer, Enk, Deinlein, Hofmann-Wellenhof, Thomas, Lallas, Blum, Stolz, et al. (2019). "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition". In: JAMA dermatology

Predictive models fail because of shortcut features

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

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". JAMA dermatology

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**

Formal problem: Is an intervention effective?

📖 **Quantify** the effect of a (binary) intervention **A** on an **outcome Y**

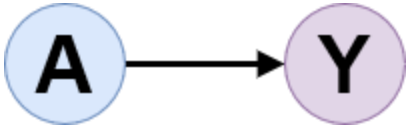
Example:

For patients with sepsis in the ICU requiring fluid rescuscitation

Should I give a combination of crystalloids and albumin

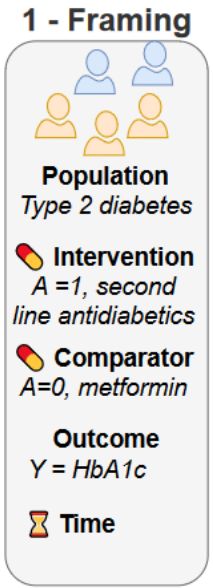
Or crystalloids only

To improve 28-day survival



Causal graph

Causal Framework: Study design

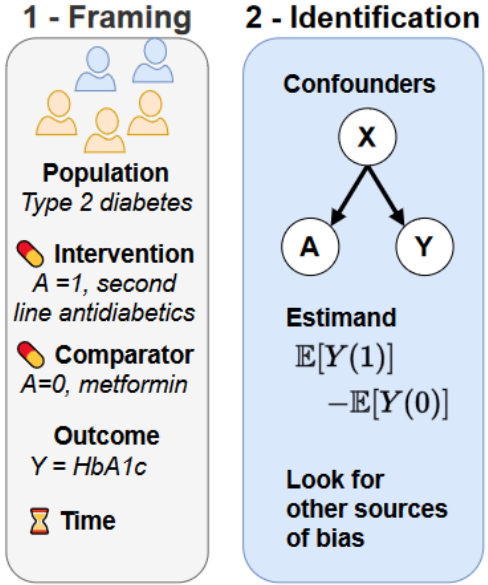


Emulate the **ideal trial** that you would conduct if you could recruit patients

Hernan, 2021.

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

Causal framework in real life: Identification

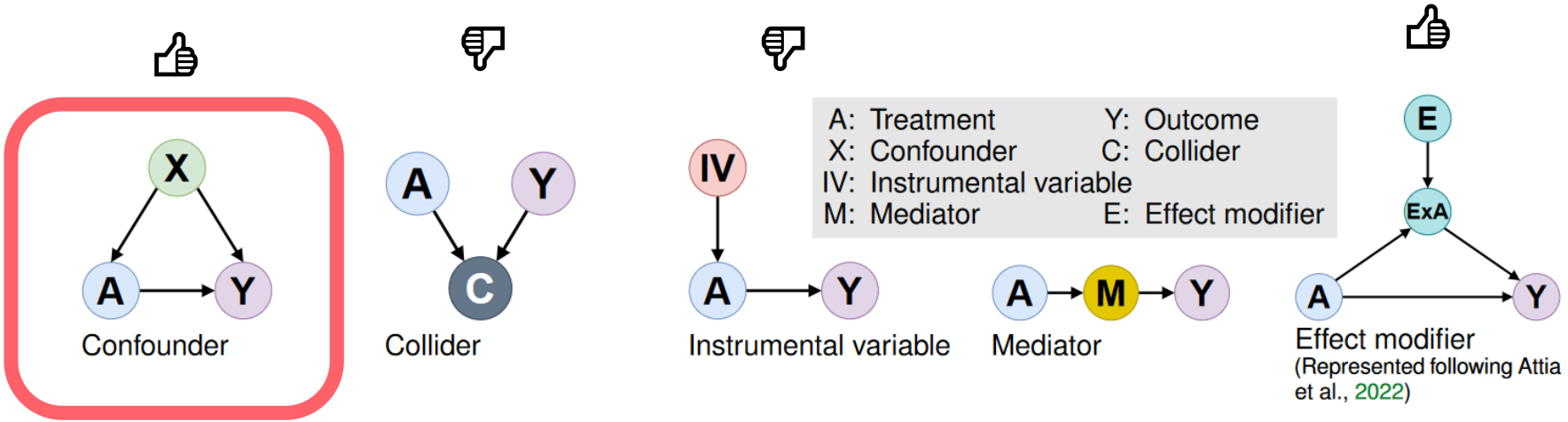


List necessary information to answer the causal question

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

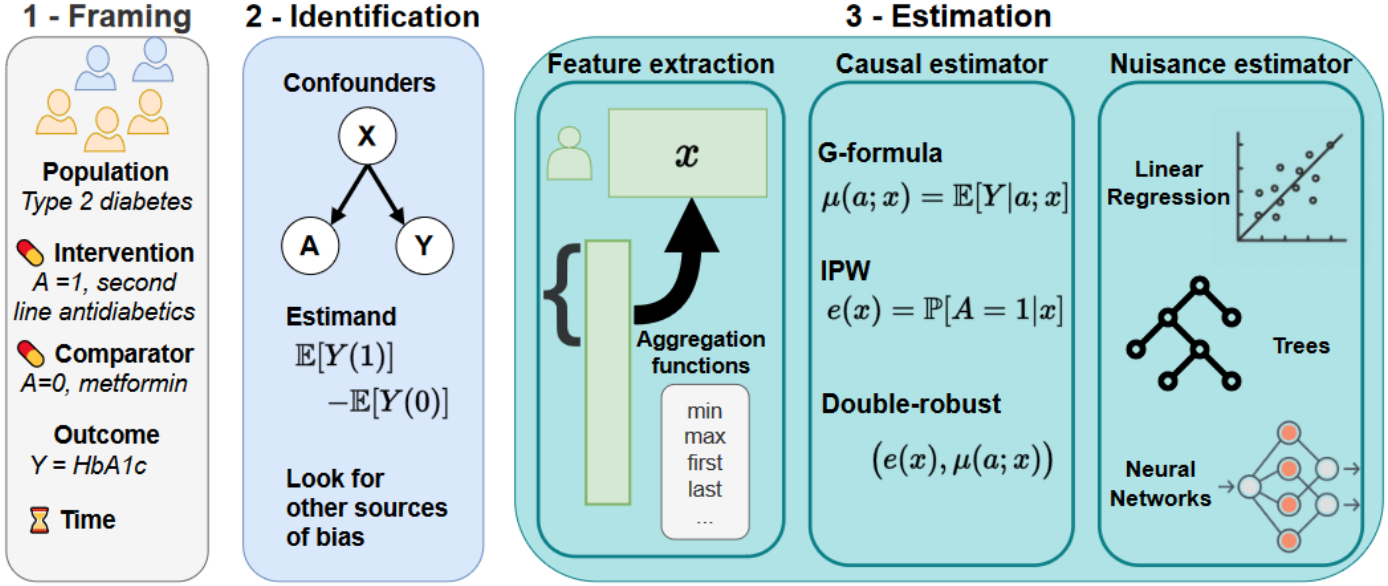
Identification - List necessary information to answer the causal question

Categorize variables in the data base



Focus on confounding

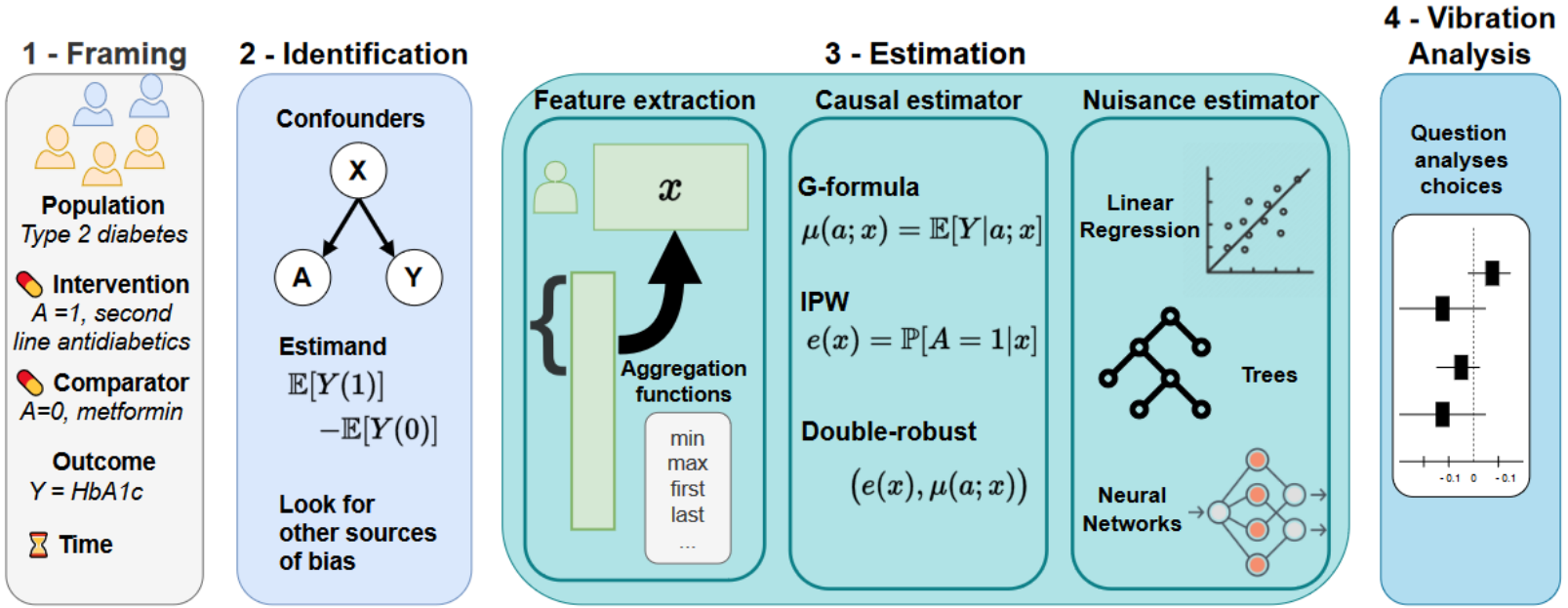
Causal Framework: Estimation



Select appropriate estimators

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

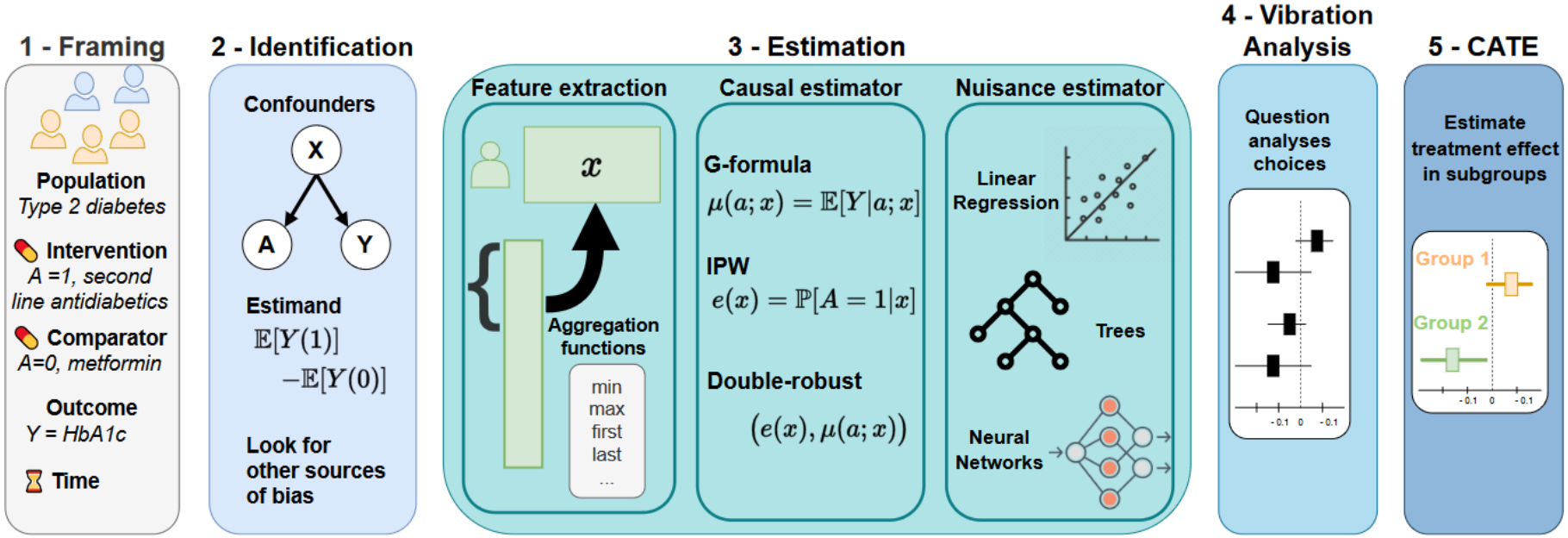
Causal Framework: Vibration analysis



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". Journal of clinical epidemiology

Causal Framework: Treatment heterogeneity



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". American Journal of Epidemiology

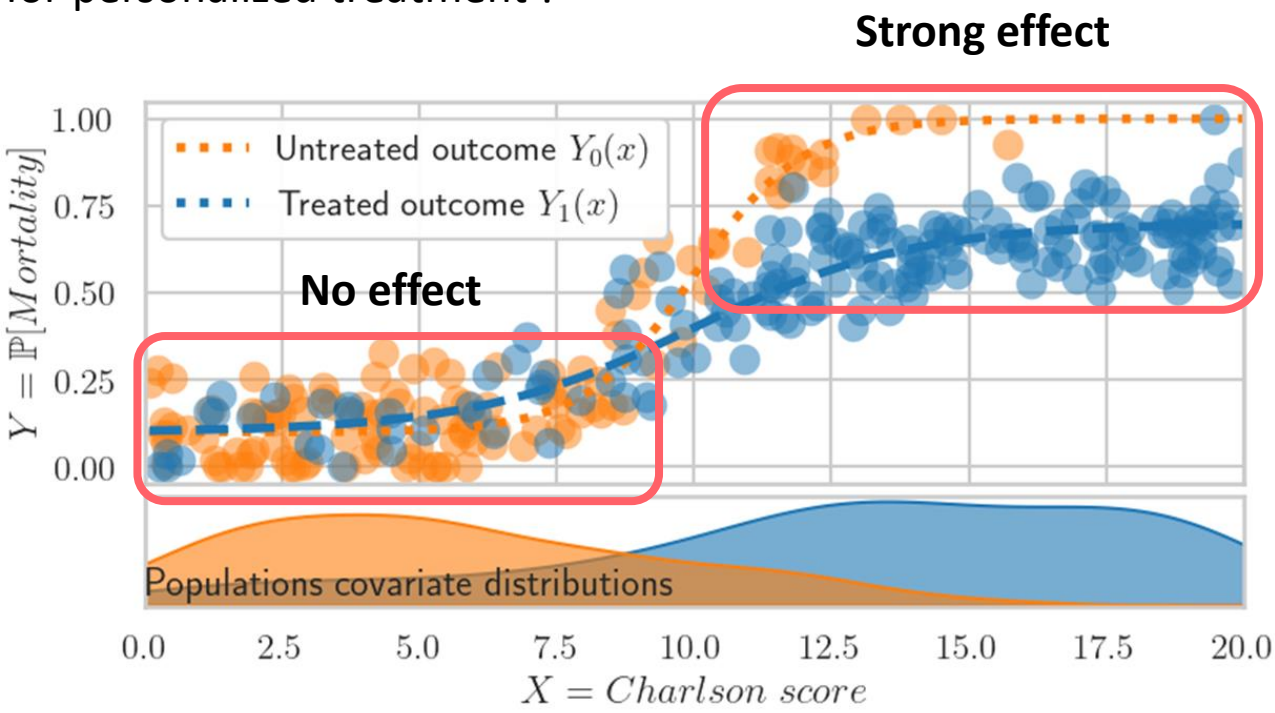
Treatment heterogeneity – Compute treatment effects on subpopulations

Does the effect vary 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



What source of bias dominates ? A practical example

Many possible 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

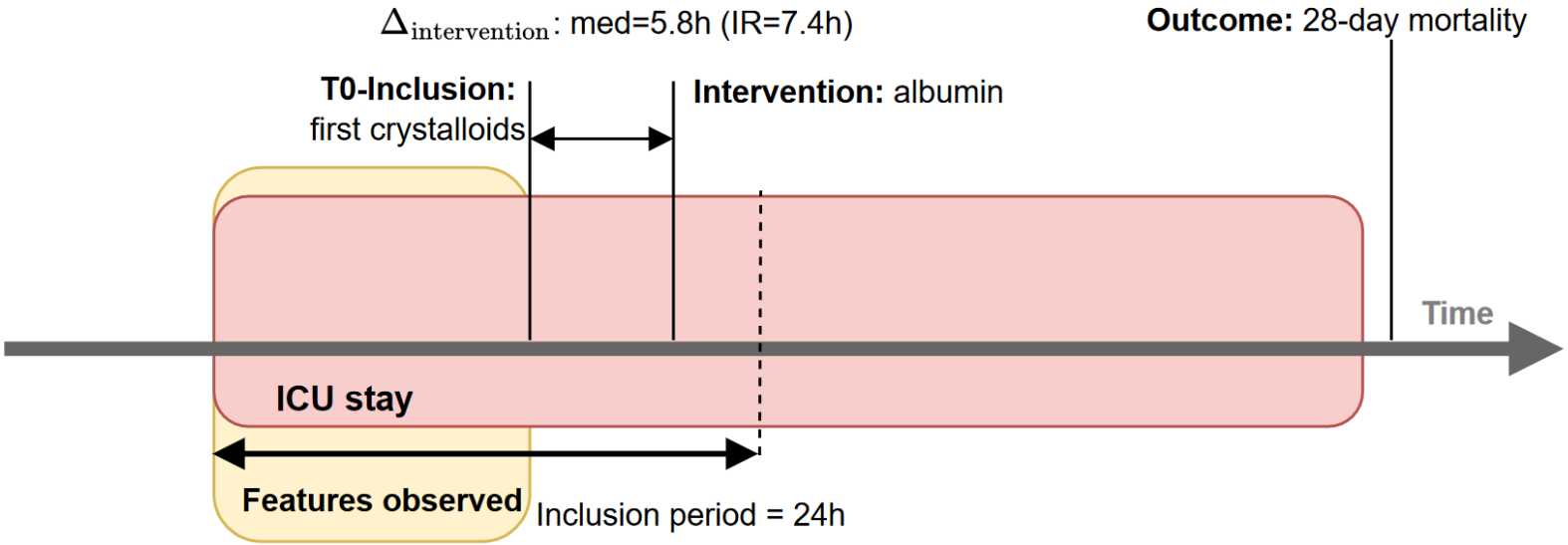
Inverse Propensity Weighting (IPW), outcome modeling (G-formula) with T-Learner, Augmented Inverse Propensity Weighting (AIPW) and Double Machine Learning (DML)

Outcome and treatment estimators: regularized logistic regression and random forest

Study design – focus on the time component

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

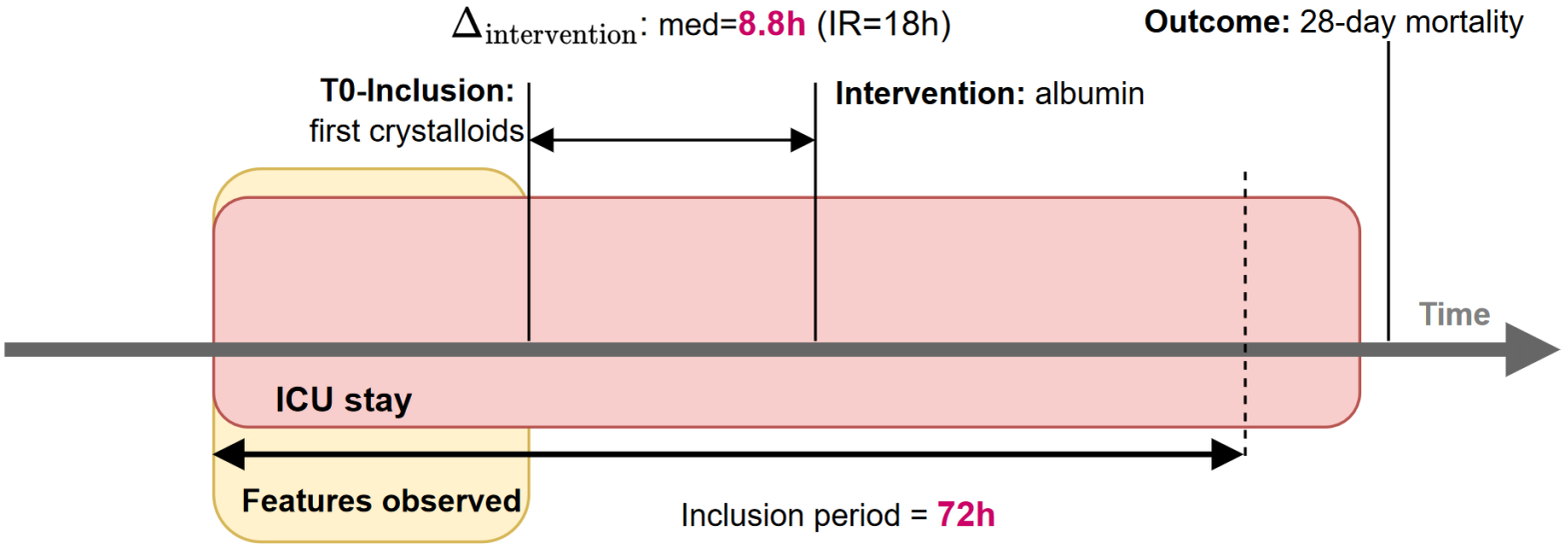
Eg. During 24 first hours of hospitalization



Study design – focus on the time component

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

Eg. During 72 first hours of hospitalization



Immortal time bias introduced with different inclusion times

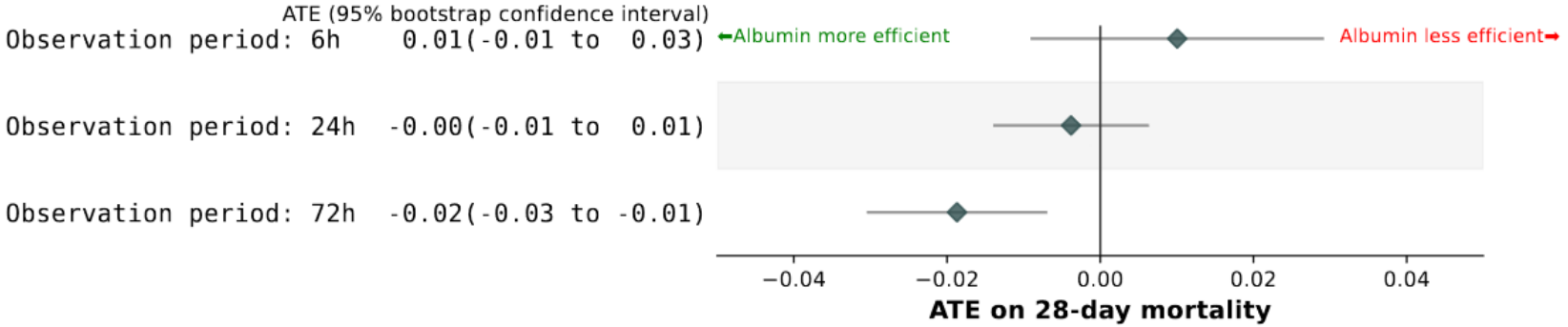
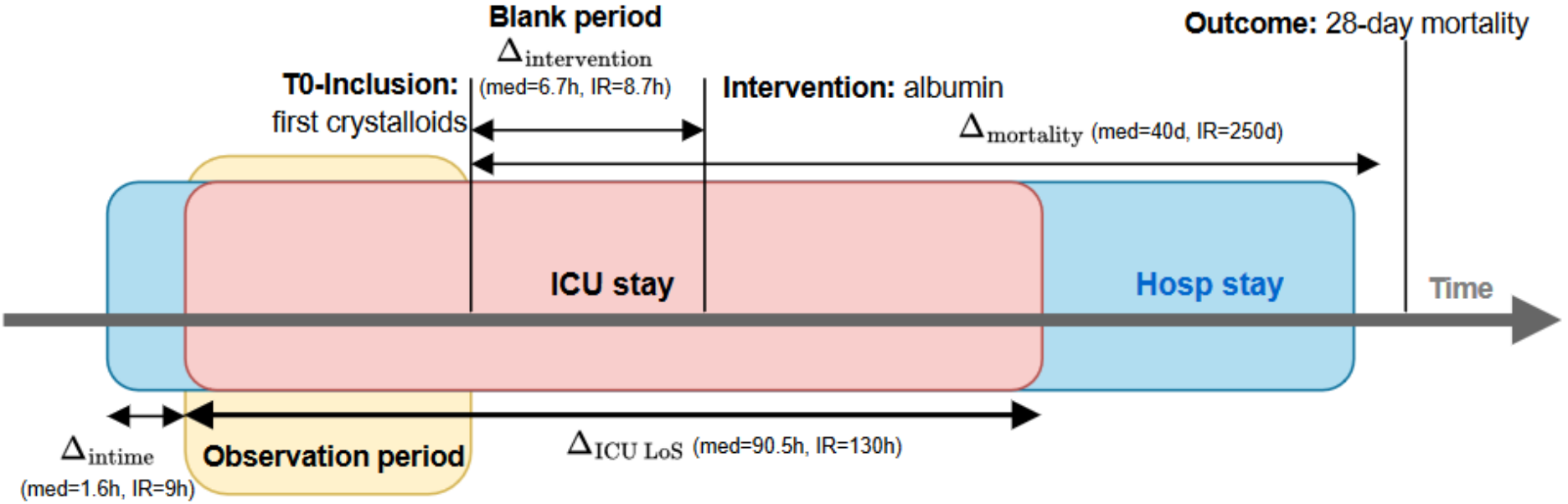


Figure 8: ***Detecting immortal time bias*** *– Increasing the observation period increases the temporal blank period between inclusion and treatment initialization, associating thus patients surviving longer with treatment: Immortal Time Bias. A longer observation period (72h) artificially favors the 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 project in nephrology where ITB was harder to control for: https://soda.gitlabpages.inria.fr/deepacau/#intervention-comparator

Immortal time bias introduced with different inclusion times



Selection flowchart for the usecase

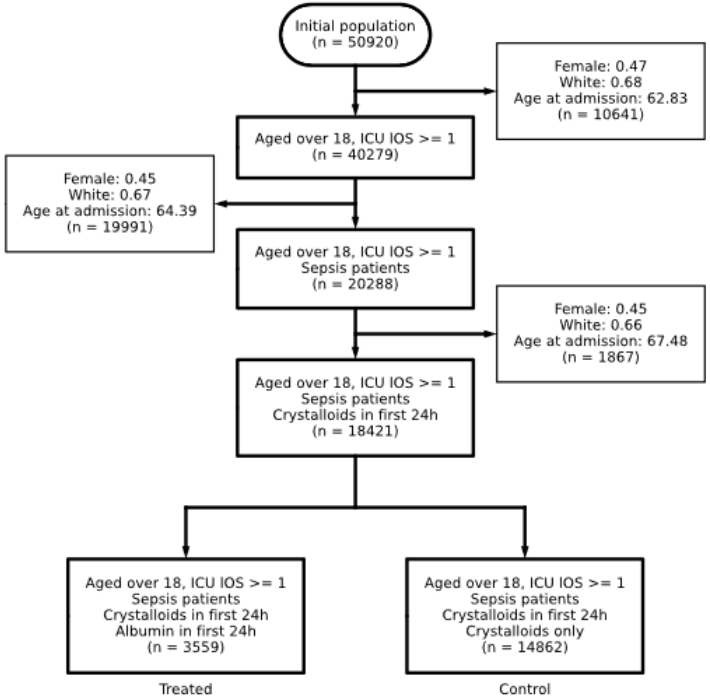


Figure 12: *Selection flowchart on MIMIC-IV for the emulated trial.*

Different choices of aggregation does not change the result

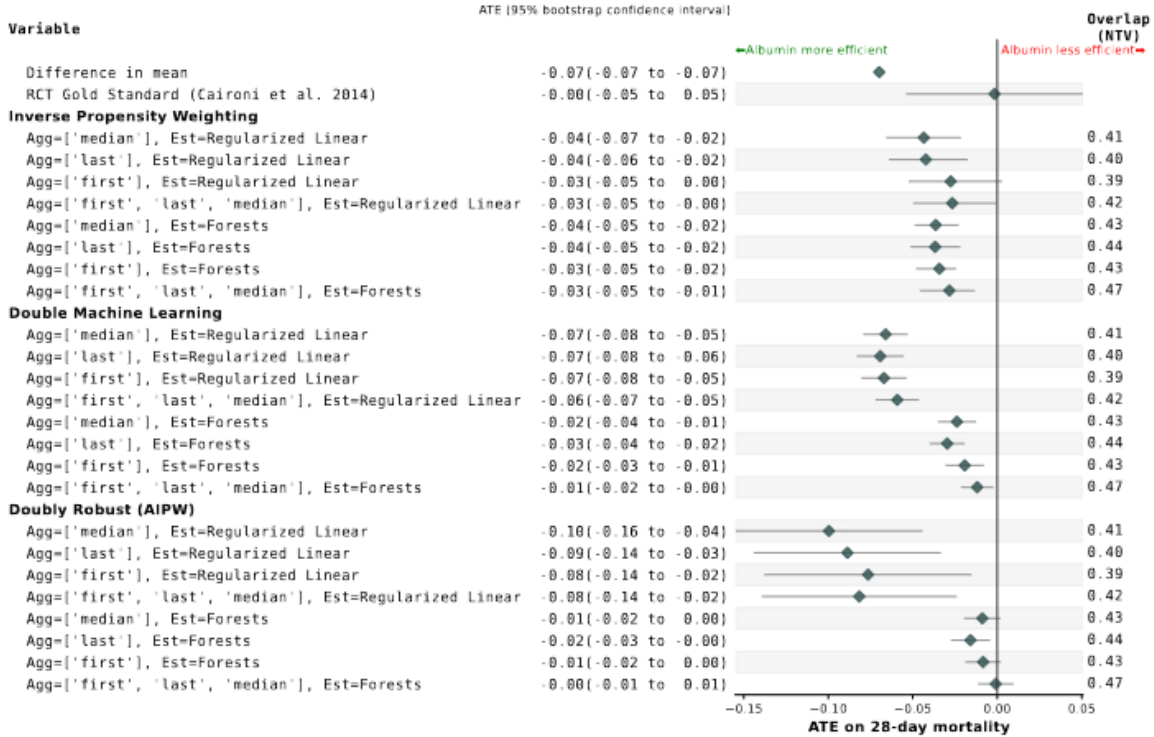


Figure 16: *Vibration analysis dedicated to the aggregation choices. The choices of aggregation only marginally modify the results. When assessed with Normalized Total Variation, the overlap assumption is respected for all our choices of aggregation. The green diamonds depict the mean effect and the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

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

Causal assumptions: 1 – Ignorability / Unconfoudneness

Consider all confounders capturing differences between **treated** and **control** populations impacting the outcome

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

Conditionally on features, treatment allocation is *as random*

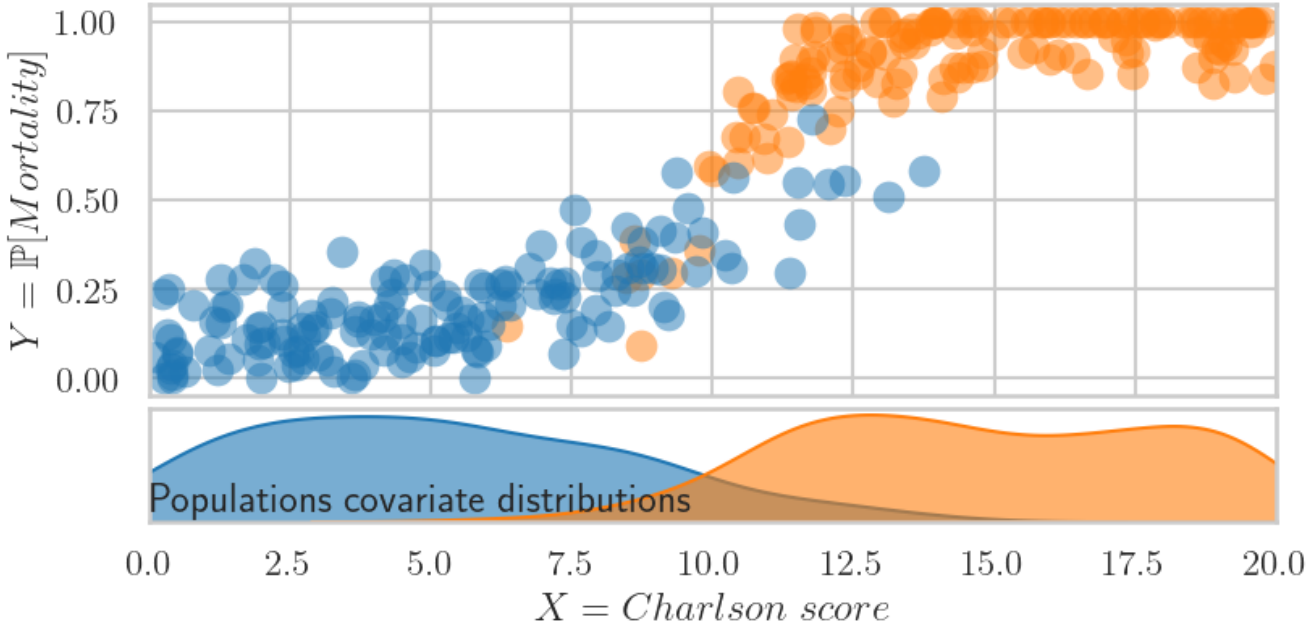
⚠ **Not verifiable with data only: Medical expertise needed** 👩‍⚕️

📋 Legally, **medical records** should contain all information considered for interventions

Causal assumptions: 2 – Positivity / Overlap

Treated and **controls** should be **close enough ie.** **randomness in treatment allocation**

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



Other (weaker) assumptions

3 - Consistancy

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 individuals and there is no interactions.

4 – Identically and independtly distributed observations

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 :

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

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 $\tilde{A} = A - \hat{e}(X)$

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) $<$ 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 $<$ 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	✓	

Lessons learned

- Study design clarifies the question and helps to avoid biases

- Choice of the estimator affect the results, choice of aggregation is less important

- Vibration analysis important to catch some bias

- Event imperfect causal graph reduce bias

- Vibration analysis require software skills (measurement tables is 300M rows in MIMIC)
https://github.com/soda-inria/causal_ehr_mimic/tree/main/caumim
- No python packages for estimation with all best statistical practices and estimators

Supplementary slides for: How to select causal models?

Empirical study: results

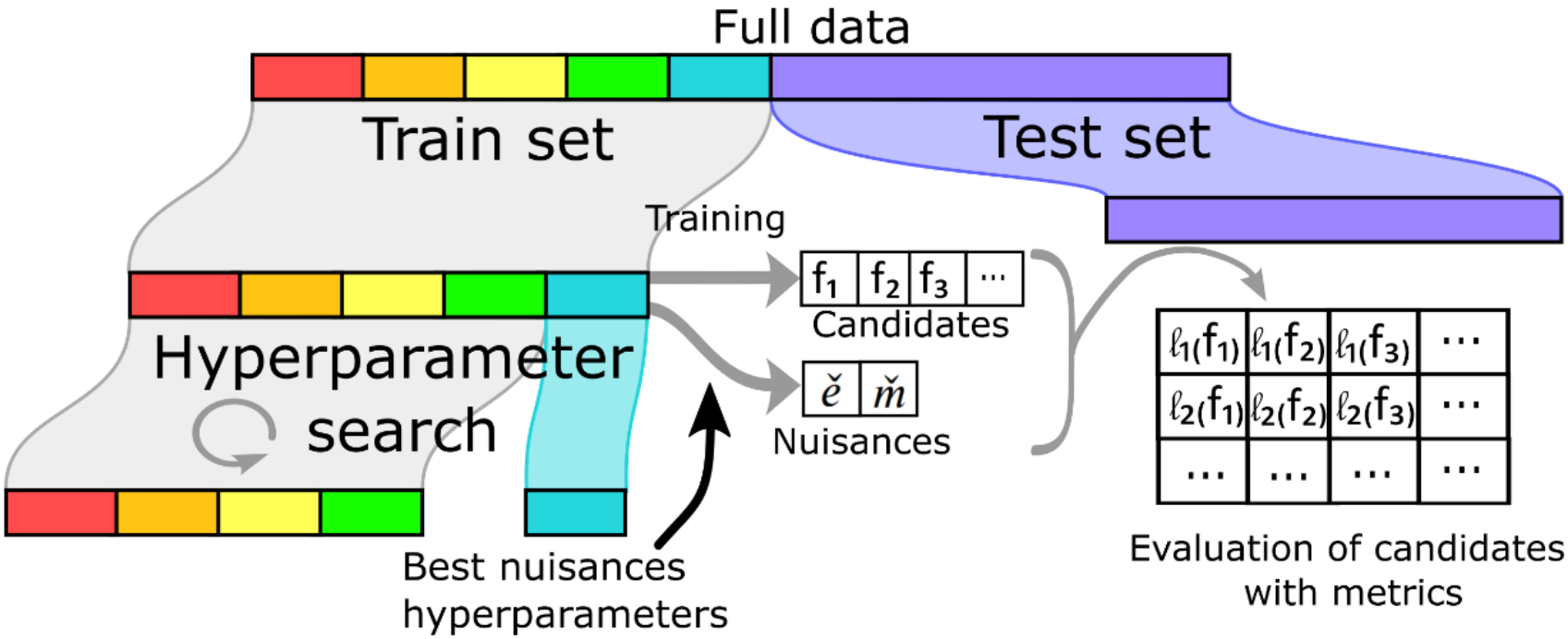
How well a **metric rank models compared to the oracle** τ*-Risk, measured with Kendall* κ

$$\kappa = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})}$$

Remove **inter-dataset variation** by substracting mean Kendall over all metrics

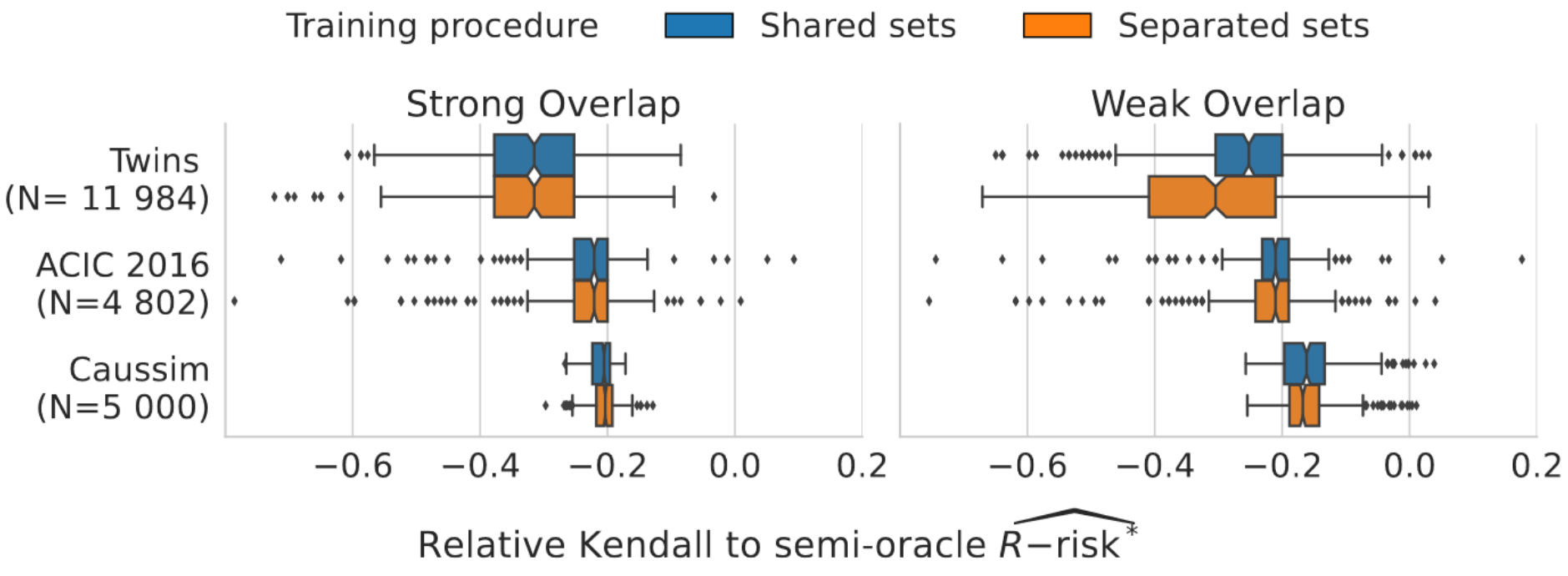
$$\kappa(\ell, \tau{-}\text{risk}) - mean_{\ell}\big(\kappa(\ell, \tau{-}\text{risk})\big)$$

Empirical study: estimation procedure



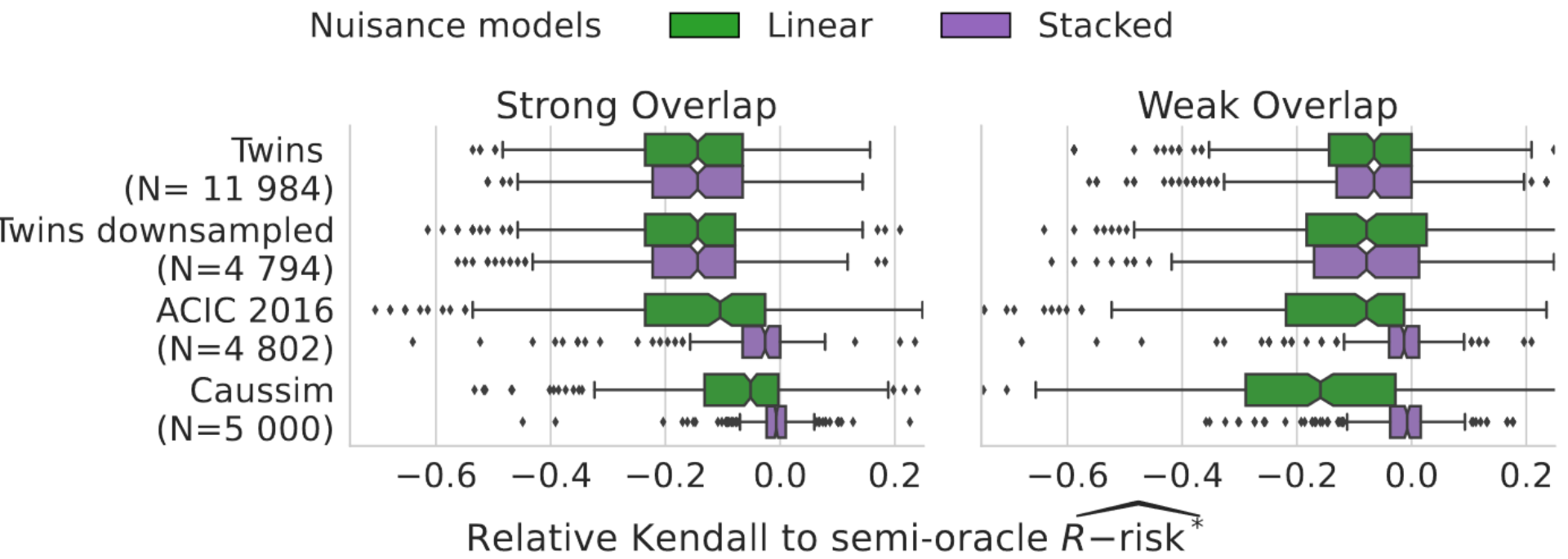
Empirical study: results

Nuisances can be estimated on the same data as outcome model



Empirical study: results

Stacked models are good overall estimators of nuisances



Empirical study: Model selection is harder for low population overlap