TREATMENT EFFECT ESTIMATION FROM EHR: FRAME, PROCESS,

INFER AND ESTIMATE

April 13, 2023

ABSTRACT

# Introduction

1.1 Motivation

Data science tasks in healthcare can be classified in three categories (Miguel A Hernán, J. Hsu, and Healy, 2019; OHDSI, 2021; Doutreligne et al., 2023): description, prediction, counterfactual prediction.

Approriate care is a decision making problem: the right care by the right provider to the right patient, at the right time

Association, 2015. This calls for proper methodology to rate appropriateness by counterbalancing expected health benefit linked to an intervention with its negative consequences Brook et al., 1986. We will thus focus on counterfactual prediction, more precisely on treatment effect estimation, letting aside other causal problems such as causal discovery. However, the same preprocessing framework can be useful for descriptive and predictive tasks.

Electronic Health Records (EHR) are receiving increasing attention for estimating causal effects Despite Randomized Control Trials (RCT) remaining the gold standard to evaluate treatment effect, EHR are interesting opportunities for: effectiveness of treatments for routine care data, generalizability, first insights at smaller costs than RCTs (Black, 1996; Bosdriesz et al., 2020), exploring treatment heterogeneity on subgroups Mant, 1999. Moreover EHR data is more adapted to measure the impact of health policies or the diffusion of professional guidelines. However, these types of interventions are more subtle to define than drug effects. Reasonning on such interventions thus requires a clear framework to avoid multiple pitfall inherent to observational studies.

1.2 Related work

Tutorials on causal inference from a data centered perspective TODO: sumz this up. Main idea is no tutorial on causal inference from EHR data (temporal).

Sharma, 2018 is close to shalit’s one.

Shalit and Sontag, 2016 is a tutorial on causal inference from observational data.

Moraffah et al., 2021 is a review on treatment effect and causal discovery from time series data.

Causal inference using EHR data Sofrygin et al., 2019 implement Targeted Maximum Likelihood Estimation in a longitudinal setup applied to dynamic regimes in diabetes cares. This is the only study to our knowledge that study the effect of different choices of time unit for features aggregation.

Using Stanford Cancer Institut Research Database, Zeng et al., 2022 augment survival analysis with unstructured text.

Sonabend et al., 2020 studied treatment choices in sepsis in an Off-line Reinforcement learning framework, with a bayesian perspective. Their results illustrate well different overlap scenarii but they do not provide a wider sense on the evaluation of their methods. They also do not study variation on the treatment effect due to data representation.

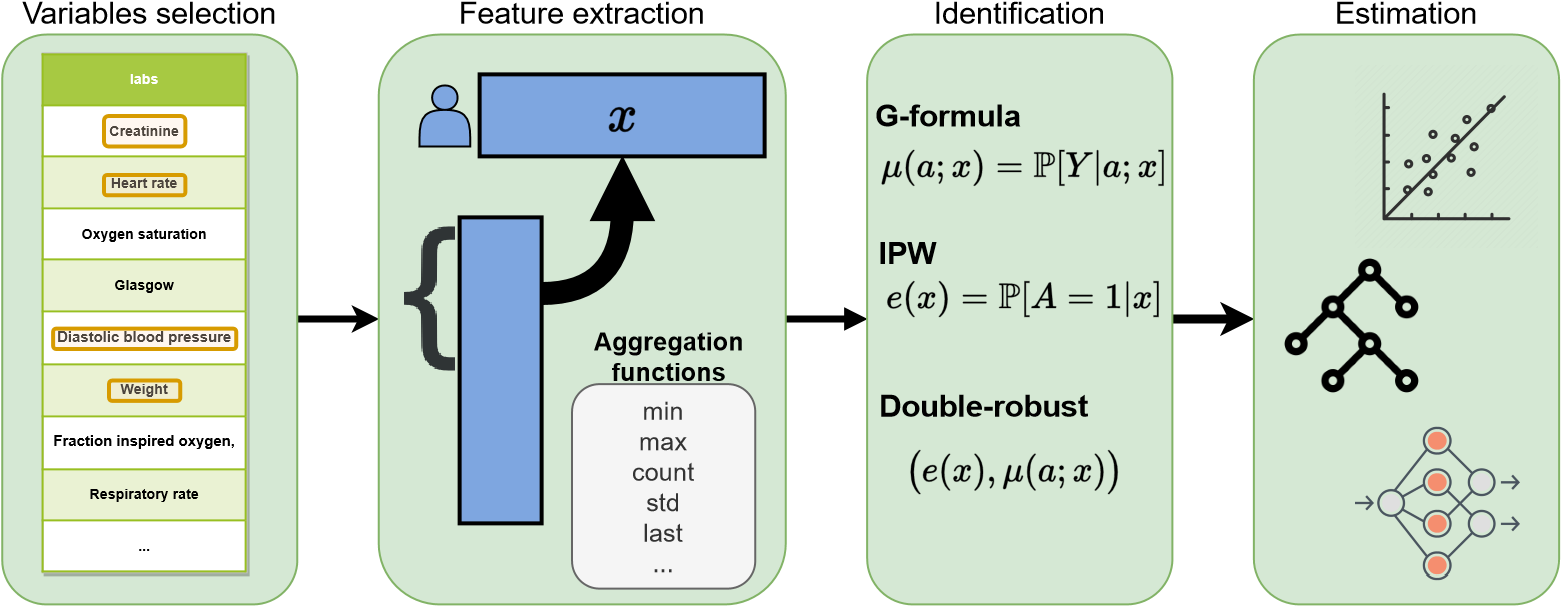


Figure 1: The complete inference pipeline confronts the applied scientist with many potentially valid choices. Some of them can be guided by experts, other by data. All of them benefit from a clear framework, separating the different kind of choices.

These studies rarely describe the engineering choices and challenges of transforming raw EHR data into analysis formats, nor the effects of these transformation on inference.

Sensitivity analysis is focused on variable selection, putting aside that modern statistical models do not operate over variables but over data representations, which are often abitratry decided before the analysis. Illustrating the different steps of an observational study focusing on couterfactual prediction, we will present on a concrete example of the effects of data representation on the inference.

Causal inference on Mimic The Mimic database has served several causal inference studies such as: the effect of indwelling arterial catheters on mortality (D. J. Hsu et al., 2015), the effect of transthoracic echocardiography on 28-day mortality (Feng et al., 2018), the effect of high-flow oxygen therapy on patients with hypoxexmia after extubation on 28-day mortality (T. Liu, Zhao, and Du, 2021), the effect of liberal vs conservative oxygenation on total mortality for mechanically ventilated patients (Gani et al., 2023), the effect of fluid-limiting treamtent stragies among sepsis patients on 30 day-mortality (Shahn et al., 2020), the effect of statin use prior to ICU admission on 30-day mortality for patients

with sepsis (Chinaeke et al., 2021).

1.3 The inference flow

This tutorial goes throught the different steps from framing the treatment effect question up to the estimation of the effect. We insist first on the formalization of the medical question as a target trial, a key step that, if overlooked endangers the validity of the inference (Miguel A Hernán, 2021). We then focus on the processing of the data, the most tedious step, underlying the compromise between residual counfounders and lack of overlap.

Matt: Might be too ambitious to "explore the compromise" which would need an entire study on semisimulated data

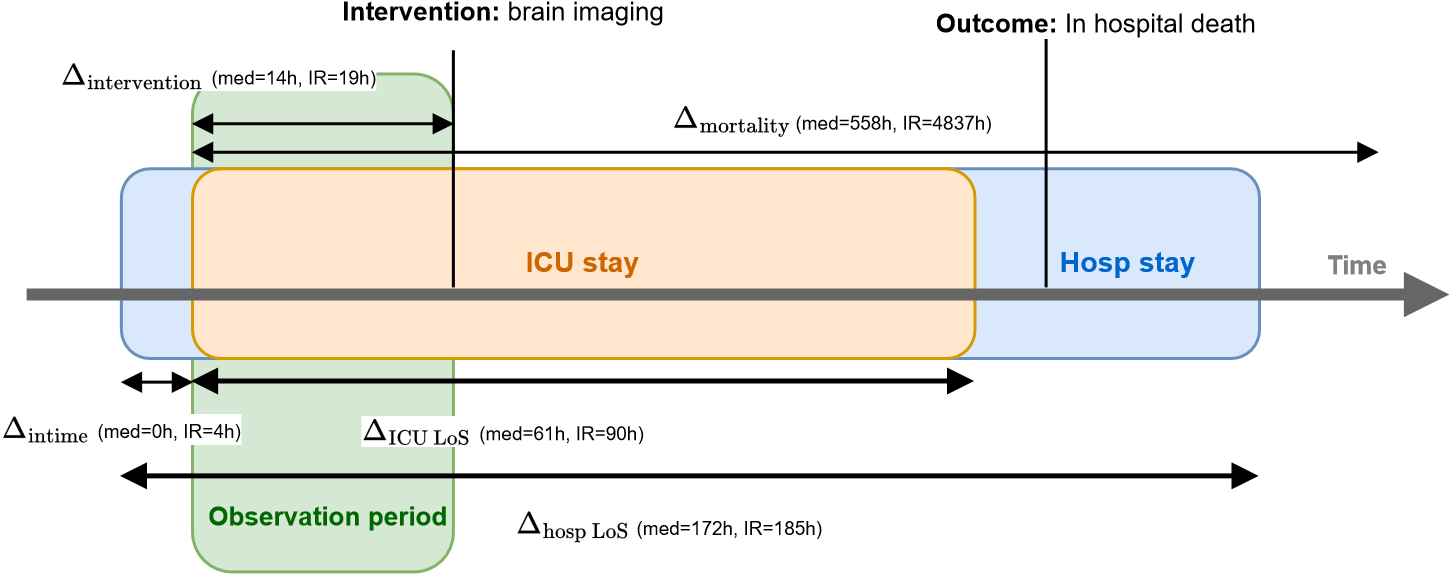
We illustrate these key steps, with the Mimic-IV database, on a specific use case: *What is the effect of performing neuro-imaging evaluation of on critical care patients with stroke related symptoms ?*

* Frame the medical question of interest,
* Preprocessing the data: focus on the event format
* Variable selection
* Variable aggregation
* Identification
* Statistical estimation
* Sensitivity analysis

|  |  |  |  |
| --- | --- | --- | --- |
| PICO component | Description | Notation | Illustration |
|  |  |  | Patients with stroke related symptoms, |
| Outcome | What are the targets that we want to compare between the two groups ? | *Y* (1)*,Y* (0) ∼ *p*(*Y* (1)*,Y* (0)) | Survival measured as instay mortality |

Table 1: PICO components help to clearly define the components of the medical question of interest

.



## ICU stay timeline

Figure 2: Defining the period of observation, from the patient inclusion up to the treatment allocation is crucial to avoid coliders. The definition of the followup period helps avoid temporal biases such as right censoring.

1.4 Framing the medical question

PICO: how to ask treatment effect question ? Miguel A. Hernán and Robins, 2016 recommends to use the concept of *target* trial when designing a comparative observational study.

The PICO framework (Richardson et al., 1995) let the practioner or the analyst write down the four components underlying a sound medical question. Table 1 associate to these steps, the notation of the potential outcome framework (Imbens and Rubin, 2015).

The observational cohort: preventing temporal biases and coliders Refer to OHDSI formalization of the cohort. Different sort of biases arises when framing the medical question and therefore, analysits should carfully review them.

When designing the inclusion criteria –P step–, some inclusion critieria could be missing non at random, which introduces a selection bias on the selected population. The worst case happens if this missingness is associated with the treatment which biases the treatment effect throught a post treatment colider (Weiskopf et al., 2023). TODO: real example with MIMIC such as missing ICD9 codes.

TODO: cite and illustrate some temporal biases such as immortal time bias.

Matt: Update with mimic-iv

1.5 The event format

Healthcare data in EHR and other large healthcare databases such as claims are a collection of event log tables, linked within a star schema to the central patient table. However, treatment effect methods take as input a matrix of shape (*nb*\_*patient*×*nb*\_*confounders*), similar to the randomized control trial *Table 1*.

For EHR data, the event format emerges as a convenient simplification of the complex star schema database (Rajkomar et al., 2018; Beam et al., 2019; Chazard et al., 2022; Bacry et al., 2020). This view allow to easily apply the remaining steps on the data.

**ZCQM008**

t0

**I25**

**A41**

**I08**

**C**

**B01AA**

**A12AX**

t1

t2

**1577 = 200**

**Anna**

**Age = 54**

**Gender = Female**

**Insurance Status = RG**

**Residence = Le Havre**

t3

😔

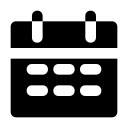
**Inpatient mortality = ?**

😀

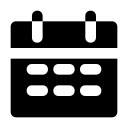
**Inpatient mortality = ?**

**A = 1**

**A = 0**



**Followup date**



**Intervention Date**

**Covariates X (demographics, biology, procedures...)**

💊

**Intervention A**

📈

**Outcome Y**

Figure 3: Simplifying the star schema into an event log makes it easier to include healthcare events of different types as potential confounders.

1.6 Variable selection

Dags, expert knowledged to the rescue, but might not sufficient in complex settings.

1.7 Variable aggregation

1.8 Identification

1.9 Estimation

1.10 Refutation step

Sensitive analysis

Negative control Talk about negative control and find an exemple for our use case.

# Use case on MIMIC-IV

2.1 Data description

2.2 Defining the population, intervention, control and outcomes

* P: The population of interest is defined as patient aged over 18, hospitalized in ICU with a billing diagnoses of stroke or TIA (ICD9 codes 430-439) as main diagnosis which have at least 24 hours of observation. We hypothetize that there is no missing stroke billing code, at least no missingness associated to the treatment. This assumption is equivalent to say that wheras a patient with stroke has a brain imaging or not, it will be coded with TIA/stroke.

Matt: Is this reasonnable ? Do we opportunistically discover a stroke by CT-scan ? Or are the symptoms sufficiently clear for the initial diagnosis to be independant.

The 24 hour required survival is an attempt to rule out extremely severe patients that would have require more urgent care than a brain imaging, for example if the circulatory or respiratory is defaillant at ICU admission.

Matt: update on mimic-iv

At this step, the cohort comprises 2 801 ICU stays with 2 646 distinct patients.

* T: We search for brain imaging during the first 24 hours (or 12 first hours ?).
* C: No brain imaging during the first 24 hours.
* O: Mortality 28-days after ICU-admission. The followup starts after the first 24 hours of survival.

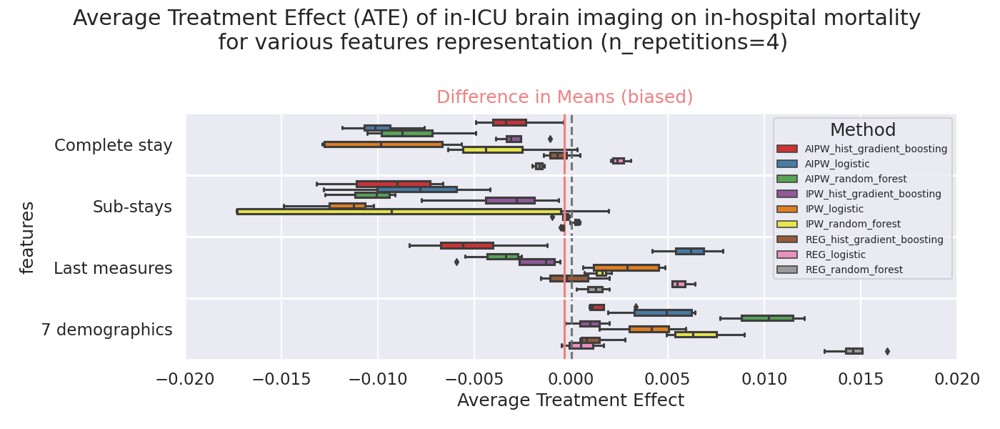


Figure 4: The different choices made in the inference pipeline have significant impact on the treatment effect results, requiring to question each model causal assumptions such as residual confounding or population overlap at the variable representation step, model specification at the estimation step.

* 1. Preprocessing confounders to event format
  2. Inference pipline Feature selection
* 13 baseline measures: We are using the same 17 ICU measured clinical variables as in (Harutyunyan et al. 2019): Capillary refill rate (binary), Diastolic blood pressure, Fraction inspired oxygen, Glascow coma scale eye opening (ordinal, 1 to 4), Glascow coma scale motor response(1 to 6), Glascow coma scale total (1 to 15), Glascow coma scale verbal response (1 to 5), Glucose, Heart Rate, Height, Mean blood pressure, Oxygen saturation, Respiratory rate, Systolic blood pressure, Temperature, Weight, pH. A summary table of the input features on the target population is given in Appendix B.4.
* Expert set
* All available measures

Feature preprocessing As a baseline for treatment effect estimation, we consider summary statistics of the stays as implemented by (Harutyunyan et al. 2019). Their script computes 6 different statistical features for a given time series (min, max, mean, sd, skew and number of measurements) for 7 subsequences of the input sequence (the full time series, the first 10, 25, 50This preprocessing of raw features leads to a vector of dimension 714 describing each ICU stay. We normalize and impute with Z-score the resulting data.

2.5 Sensitivity results

Matt: Update with mimic-iv

# References

Association, Canadian Medical (2015). “Appropriateness in Health Care”. In: *Canadian Medical Association Policy*,

p. 9. URL: [https://policybase.cma.ca/media/PolicyPDF/PD15-05.pdf.](https://policybase.cma.ca/media/PolicyPDF/PD15-05.pdf)

Bacry, Emmanuel et al. (2020). “SCALPEL3: a scalable open-source library for healthcare claims databases”. In: *International Journal of Medical Informatics* 141, p. 104203.

Beam, Andrew L et al. (2019). “Clinical concept embeddings learned from massive sources of multimodal medical data”. In: *Pacific Symposium on Biocomputing 2020*. World Scientific, pp. 295–306.

Black, Nick (1996). “Why we need observational studies to evaluate the effectiveness of health care”. In: *Bmj* 312.7040, pp. 1215–1218.

Bosdriesz, Jizzo R et al. (2020). “Evidence-based medicine—when observational studies are better than randomized controlled trials”. In: *Nephrology* 25.10, pp. 737–743.

Brook, Robert H et al. (1986). “A method for the detailed assessment of the appropriateness of medical technologies”. In: *International journal of technology assessment in health care* 2.1, pp. 53–63.

Chazard, Emmanuel et al. (2022). “Book Music Representation for Temporal Data, as a Part of the Feature Extraction Process: A Novel Approach to Improve the Handling of Time-Dependent Data in Secondary Use of Healthcare Structured Data”. In: *Studies in Health Technology and Informatics* 290, pp. 567–571.

Chinaeke, Eric E et al. (2021). “The impact of statin use prior to intensive care unit admission on critically ill patients with sepsis”. In: *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 41.2, pp. 162–171.

Doutreligne, Matthieu et al. (2023). “Good practices for clinical data warehouse implementation: a case study in France”. In: *arXiv preprint arXiv:2302.07074*.

Feng, Mengling et al. (2018). “Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database”. In: *Intensive care medicine* 44, pp. 884–892.

Gani, Md Osman et al. (2023). “Structural causal model with expert augmented knowledge to estimate the effect of oxygen therapy on mortality in the icu”. In: *Artificial Intelligence in Medicine* 137, p. 102493.

Hernán, Miguel A (2021). “Methods of public health research—strengthening causal inference from observational data”. In: *New England Journal of Medicine* 385.15, pp. 1345–1348.

Hernán, Miguel A, John Hsu, and Brian Healy (2019). “A second chance to get causal inference right: a classification of data science tasks”. In: *Chance* 32.1, pp. 42–49.

Hernán, Miguel A. and James M. Robins (Apr. 2016). “Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available”. In: *American Journal of Epidemiology* 183.8.

Hsu, Douglas J et al. (2015). “The association between indwelling arterial catheters and mortality in hemodynamically stable patients with respiratory failure: a propensity score analysis”. In: *Chest* 148.6, pp. 1470–1476.

Imbens, Guido W and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Liu, Taotao, Qinyu Zhao, and Bin Du (2021). “Effects of high-flow oxygen therapy on patients with hypoxemia after extubation and predictors of reintubation: a retrospective study based on the MIMIC-IV database”. In: *BMC Pulmonary Medicine* 21.1, pp. 1–15.

Mant, David (1999). “Can randomised trials inform clinical decisions about individual patients?” In: *The Lancet* 353.9154, pp. 743–746.

Moraffah, Raha et al. (2021). “Causal inference for time series analysis: Problems, methods and evaluation”. In: *Knowledge and Information Systems* 63, pp. 3041–3085.

OHDSI (2021). *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI. URL: [https:](https://ohdsi.github.io/TheBookOfOhdsi/)

[//ohdsi.github.io/TheBookOfOhdsi/.](https://ohdsi.github.io/TheBookOfOhdsi/)

Rajkomar, Alvin et al. (2018). “Scalable and accurate deep learning with electronic health records”. In: *NPJ digital medicine* 1.1, p. 18.

Richardson, W Scott et al. (1995). “The well-built clinical question: a key to evidence-based decisions”. In: *Acp j club* 123.3, A12–A13.

Shahn, Zach et al. (2020). “Fluid-limiting treatment strategies among sepsis patients in the ICU: a retrospective causal analysis”. In: *Critical Care* 24.1, pp. 1–9.

Shalit, Uri and David Sontag (2016). URL: [https://docplayer.net/64797211-Causal-inference-forobservational-studies.html.](https://docplayer.net/64797211-Causal-inference-for-observational-studies.html)

Sharma, Amit (2018). *Tutorial on causal inference and counterfactual reasoning*. URL: [https://causalinference.](https://causalinference.gitlab.io/kdd-tutorial/)

[gitlab.io/kdd-tutorial/.](https://causalinference.gitlab.io/kdd-tutorial/)

Sofrygin, Oleg et al. (2019). “Targeted learning with daily EHR data”. In: *Statistics in Medicine* 38.

Sonabend, Aaron et al. (2020). “Expert-supervised reinforcement learning for offline policy learning and evaluation”. In: *Advances in Neural Information Processing Systems* 33, pp. 18967–18977.

Weiskopf, Nicole G et al. (2023). “Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse”. In: *Journal of the American Medical Informatics Association*, ocad013.

Zeng, Jiaming et al. (Feb. 2022). “Uncovering interpretable potential confounders in electronic medical records”. In: 13.