université
PARIS-SACLAY

Representations and inference from time-varying routine care data

Représentations et inférence à partir de données de santé temporelles collectées en routine

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique mathématique
Graduate School: Informatique et sciences du numérique,
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche Inria Saclay-Île-de-France (Université Paris-Saclay, Inria), sous la direction de **Gaël Varoquaux**, directeur de recherches à l'Inria, le co-encadrement de **Claire Morgand**, Dr à l'Agence Régionale de Santé Ile-de-France, et la co-supervision de **Pierre-Alain Jachiet**, Dr. à la Haute Autorité de Santé

Thèse soutenue à Paris-Saclay, le 20 novembre 2023, par

Matthieu DOUTRELIGNE

Composition du jury

Membres du jury avec voix délibérative

Pierre Zweigenbaum Directeur de Recherche, Laboratoire Interdisciplinaire des sciences du numérique de Paris Saclay	Président du jury
Florence Tubach Professeur des Universités - Praticien Hospitalier, Sorbonne Université	Rapporteur & Examinatrice
Peter Szolovits Professor, Massachusetts Institute of Technology	Rapporteur & Examinateur
Emmanuel Chazard Professeur des Universités - Praticien Hospitalier, Université de Lille	Examinateur
Marzyeh Ghasseimi Assistant Professor, Massachusetts Institute of Technology	Examinatrice
Etienne Audureau Professeur, Paris Est Créteil	Examinateur

Thèse de doctorat

NNT: 2023UPASG073

Abstract

Real World Databases are increasingly accessible, exhaustive and with fine temporal details. Unlike traditional data used in clinical research, they capture the routine organization of care. These day-to-day records of patients care open the door to new research questions, notably concerning the efficiency of interventions after market access, the heterogeneity of their benefits in under-served populations or the development of personalized medicine. On the other hand, the complexity and large-scale nature of these databases pose a number of challenges for effectively answering these questions. To remedy these problems, econometricians and epidemiologists have recently proposed the use of flexible models combining causal inference with high-dimensional machine learning.

We first illustrate with three examples the current tension between these new sources of data, machine learning and modern public health issues. These examples motivate the main research question of this work: How flexible models can help delivering appropriate treatment to each and every patient to improve her health? In order to gain a better understanding of the modern infrastructures for collecting and analyzing Electronic Health Records (EHRs), we summarize semi-structured interviews conducted as part of a national case study of the clinical data warehouses (CDWs) of the 32 French regional and university hospitals. Acknowledging the difficulty to access large sample sizes and computational power to develop generalizable predictive models, we explore a complexity gradient in representation and predictive algorithms for EHRs. We then turn to causal thinking, detailing key elements necessary to robustly estimate treatment effect from time-varying EHR data. We illustrate the impact of methodological choices in studying the effect of albumin on sepsis mortality in the Medical Information Mart for Intensive Care database (MIMIC-IV). EHRs are high-dimensional databases. For such settings, the selection of hyper-parameters for the causal model is crucial to avoid under- or over-learning. In a simulation and three semi-simulated datasets, we show that the usual machine learning risk are not adapted to the causal setting and that the doubly robust R-risk outperforms other existing causal risks.

Résumé en français

Les bases de données de vie réelle sont de plus en plus accessibles, exhaustives, avec des détails temporels précis. Contrairement aux données utilisées dans la recherche clinique traditionnelle, elles capturent l'organisation routinière des soins. Ces données de soins quotidiens ouvrent la porte à de nouvelles questions de recherche, notamment en ce qui concerne la qualité des soins, l'efficacité des interventions après leur mise sur le marché, l'hétérogénéité de leurs bénéfices dans les populations mal desservies ou le développement de traitements personnalisés. D'un autre côté, la complexité et la nature à grande échelle de ces bases de données posent un certain nombre de défis pour une utilisation efficace. Pour remédier à ces problèmes, les économètres et les épidémiologistes ont récemment proposé l'utilisation de modèles flexibles combinant l'inférence causale et l'apprentissage automatique en grande dimension.

Dans un premier temps, nous illustrons par trois exemples la tension actuelle entre ces nouvelles sources de données, l'apprentissage automatique et des problématiques modernes de santé publique. Ces exemples motivent notre principale question de recherche : Comment des modèles flexibles peuvent-ils aider à fournir un traitement approprié à chaque patient afin d'améliorer sa santé ? Afin de mieux comprendre les infrastructures modernes de collecte et d'analyse des dossiers patients informatisés (DPI), nous faisons la synthèse d'entretiens semi-structurés menés dans le cadre d'une étude de cas nationale portant sur les entrepôts de données cliniques des 32 hôpitaux régionaux et universitaires français. Reconnaissant la difficulté d'accéder à des échantillons de grande taille et à la puissance de calcul pour développer des modèles prédictifs généralisables, nous étudions un gradient de complexité dans les représentations et les algorithmes prédictifs sur DPI. En se tournant vers le cadre causal, nous détaillons ensuite les éléments clés nécessaires pour estimer de manière robuste l'effet du traitement à partir de données de DPI variant dans le temps. Nous documentons l'impact de différents choix méthodologiques pour l'étude de l'effet de l'albumine sur la mortalité dans des cas de septicémie avec la base de données MIMIC-IV (Medical Information Mart for Intensive Care). Les DPIs sont des bases de données à grandes dimensions. Pour de tels problèmes, la sélection d'hyperparamètres pour les modèles causaux est cruciale afin d'éviter le sous-apprentissage ou le sur-apprentissage. Grâce à une simulation et trois ensembles de données semi-simulées, nous montrons que le risque usuel en apprentissage statistique n'est pas adapté au cadre causal et que le risque R doublement robuste surpasse d'autres risques causaux existants.

Remerciements

Avant tout, merci à toi, Gaël. Merci pour ta confiance, merci pour ta présence continue et bienveillante, merci pour ton énergie et ton enthousiasme communiquants. Merci pour les conseils, les encouragements, les relectures attentives. Merci de m'avoir pousser à chercher les bonnes questions et à montrer pourquoi celles-ci sont importantes. Je n'aurais pas pu rêver d'un meilleur directeur de thèse. J'ai l'impression d'avoir énormément appris à tes côtés. J'espère à l'avenir, faire honneur à cette formation. Un grand merci également à Claire d'avoir été là. Ca n'a pas été facile de garder contact à distance, mais nos discussions et tes conseils ont toujours été très précieux pour ma réflexion.

Un immense merci à toi Pierre-Alain. Tu as été le manager parfait pour cette garde partagée : bienveillant, attentif et un très grand exemple de rigueur. Merci pour tes encouragements et tes conseils. J'espère continuer à partager ton enthousiasme et ton recul critique pour tous les sujets de données de santé.

Merci à tous les membres du jury d'avoir accepté de revoir et sanctionné mes travaux. Thanks to Prof. Peter Szolovits to have taken the time to review my manuscript. Merci Prof. Florence Tubach d'avoir accepté d'être rapporteuse pour cette thèse avec quelques aspects très quantitatifs. Thank you Prof. Marzyeh Ghassemi, Prof. Emmanuel Chazard, Prof. Etienne Audureau and Prof. Pierre Zweigenbaum for being part of the committee. I feel honored to present my work in front of great researchers who inspire me.

Merci à toutes les personnes de l'Inria dans ces brillantes équipes MIND – SODA/scikit-learn. Merci à tous les PI, post-docs et ingénieurs: Judith, Marine, Jill-Jênn, Thomas (team café moulu), Olivier, Guillaume, Franck, Riccardo, Arturo, Jun. Et merci également à tous les occupants occasionnels de l'open-space doctorant: Alexis, Alex, Thomas, Julie, Jovan, Cédric, Benoît, Lillian, Loic, Vincent, Léo, Samuel. Une pensée particulière pour Béné, que j'ai pu réellement rencontré à l'Inria six ans seulement après notre arrivée commune sur le plateau de Saclay. Merci pour les discussions passionnantes que ce soit sur les plans scientifiques ou juste de café de comptoir. Merci Théo pour les discussions et les soutiens galères sur l'entrepôt de l'AP. Je réalise la chance que j'ai eu de travailler dans un environnement aussi stimulant et bienveillant.

Merci à toute l'équipe de Data Science de l'AP-HP pour leur accueil chaleureux les jeudis et pour les chouquettes : Romain, Thomas, Perceval, Ariel, Adam, Charline, Etienne, Christelle, et ceux que j'oublie.

J'aimerais également remercier tous les membres de la mission Data de la HAS : Timothée pour ses PR magnifiques, Pavel pour Despacito et les discussions NLP, Morgane pour les nouvelles de la campagne Normande, Adeline pour ses conseils ~~trop~~ carrés, Catherine pour les cafés du lundi matin, Christine pour sa bonne humeur, et Viktor pour les Ti Punchs. Malgré une présence en dent de scie de ma part, vous m'avez intégré pleinement dans l'équipe. C'est magique d'avoir vu cette équipe se construire et d'en faire partie aujourd'hui. Merci également aux autres personnes que j'ai cotoyées à la HAS, et qui m'ont accompagné durant cette aventure: Judith Fernandez, Pierre Liot, Agnès Solomniac, Sandrine Morin. Merci à Thomas Wanecq qui a rendu cette thèse possible, fait rare dans le milieu de l'administration.

Merci à Aude-Marie Lalanne Berdouticq et Emmanuel Didier (Institut Santé numérique en Société) pour leur regard de sociologues lors de la revue sur les entrepôts de données de santé.

Merci à ceux de l'APHP, de la Drees et d'ailleurs qui m'ont introduit dans le monde de la donnée de santé (et lors du Covid) : Nicolas, Adrien, Ivan, Xavier, Stéphanie, Albert, Phong, Arnaud, Raphaële, Antoine N., Antoine L.

Thanks to everyone at MIT or Harvard who made my two-month stay in Boston so enjoyable. Thanks to Leo Celi for the kind welcoming and the burpees. Thanks to Tristan for the nice French conversations and his great advices for chapter 4. Thanks to Fredrik for his communicating enthusiasm. Thanks to João, Luis and all the organizing team of the MIT Critical Datathon. I had so much pleasure to participate to the event ! Thanks to Pablo and all the wonderful team 9. Thanks to all other members of the Physionet team for their encouragements and supports. Merci Brice d'avoir été ma caution vin rouge, culture et pickles pendant ce voyage. Merci Clara de m'avoir permis de prendre quelques plombs dans les super voies de BP.

Et merci à tous les ami.e.s qui m'accompagent depuis si longtemps et qui m'ont supporté ces trois dernières années. Merci à Pilou pour m'avoir donné le goût de la geekitude il y a déjà un bon moment. Merci à Théo pour l'avoir renouvelé à l'X. Merci Pierre pour cette amitié qui tient malgré la distance: j'espère que nos discussions académiques et profanes continueront encore longtemps. Merci aux zamishs pour les diners administratifs : Matthieu, Vincent, Albane, Erwan. Merci à tous ceux de Saint-Ex avec qui je m'échappe le moment d'un WE: Dédé, Pôle, Thomas, Lélé, Tiph, Laura, Girou, Cindy, Mathilde et Nils. Merci Nico et Daph, toujours là pour une séance à Bleau. Merci à tous ceux de la grimpe, les semaines en falaises sont inoubliables grâce à vous : Lorraine, Théo, Albane, Théo^2, Lucas, Alice, Louis (promis après ce manuscrit je fais du typst), et Noémie. Merci Orel pour ces soirées de grimpe à Pantin. Merci Philou d'être cet ami attentif. Merci à ceux que je vois de loin en loin: Féf, Anthoine, FH, Paul, Sylvain, Hippolyte, Manu, Gustave.

Merci Nathalie de m'avoir fait visiter il y a longtemps ton labo de biologie, alimentant ma curiosité pour la recherche.

Papa, Maman, Cécile, vous m'avez accompagné depuis la première rentrée des classes jusqu'à l'école d'ingénieur et la thèse. Je sais ce que je vous dois et j'espère ne pas l'oublier. Merci pour votre présence et pour le goût de la lecture, qui mène si loin.

Merci à toi Manon d'exister et d'être là.

Contents

Chapter 1

Introduction: Amazing opportunities of health data?

Outline

1.1 Why focus on health inference from Electronic Health Records

1.1.1 *Data looking for a question [...] and rais[ing] puzzles* (Cox, 2001)

Machine learning has met great success in leveraging large amount of poor quality and weakly labelled data in Natural Language Processing or computer vision. Can other applications benefit from such approach. Because *medical practice, and biomedical research, [are] inherently information-management tasks* (Patel et al., 2009), many researchers foresaw a big potential for improving healthcare in applying machine learning to novel data collections (Topol, 2019; Rajkomar et al., 2019).

There is currently a tension between massive routine care data collections such as claims or Electronic Health Records (EHRs) (presented in Section 1.2), a new statistical framework (machine learning (Breiman, 2001b) described in Section 1.3), and pressing analytical questions in public health (detailed in Section 1.4). To understand the importance and the challenges of health inference from EHRs, let us discuss three concrete examples below.

1.1.2 Learning statistics during the Natural Language Processing revolution

In the 2010's, Halevy et al., 2009 proposed to take advantage of regularities present in large piles of data to automatically design features relevant for multiple application tasks. Consider a model trained to classify images from a massive collection of labelled pictures. During training, it progressively learns internal representations of natural images such as feet, faces, or hands. These intermediary representations can be reused –transferred– to applicative

tasks with different objectives such as predicting the severity of a traumatism from a photo. This paradigm –called pre-training– has been very successful in applied domains such as computer vision (Krizhevsky et al., 2012), then Natural Language Processing (NLP) (Devlin et al., 2018) and structural biology (Jumper et al., 2021). A valuable question is whether pre-training could be applied to other fields with vast amounts of complex data such as healthcare.

1.1.3 Paris hospitals, a large scale repository of clinical notes

In 2017, the emerging healthcare data warehouse of the Paris Hospitals (AP-HP), acknowledging the continuous improvement of NLP, ambitioned to leverage its vast repository of clinical notes for research. Yet, routine care data is not a familiar material for medical research. For engineers and NLP researchers, this routine care data are different from traditional research data collection because it requires new tools –e.g., from NLP– to deal with the novel complexity and scale of the data collection process. For epidemiologists and physicians, the major difference lies in the opposition between experimental and observational data (presented in Subsection 1.2.3).

1.1.4 Billing claims: new data for public health?

In 2018, the direction of statistics of the French ministry of health extracted billions of national claims from the aging national health platform into a dedicated server. To improve and accelerate analytics on this new platform, they built upon the work from Bacry et al., 2020 to leverage parallel computing and developed collaborative documentation (HDH, 2023a). However, describing sequences of cares remained challenging due to the high number of different medical events used in this data. This problem known as the *curse of dimensionality* has been introduced in operational research (Bellman, 1957) and is well known in statistics (Breiman, 2001b): The required number of samples to train predictive algorithms typically grows exponentially with the number of dimensions. Sequences of claim events can be viewed as a sequence of tokens, just as natural sentences. Thus, a trend from medical informatics, draws from NLP techniques, to create low dimensional representations –embeddings– of medical concepts (Beam et al., 2019). The relationships between events captured by these embeddings are strikingly close to known associations as shown in Figure 1.1. However, the downstream utility of such representations remains unclear. Public health policy is not concerned with knowledge representation or even predictive accuracy. Public services seek to understand the heterogeneities in healthcare consumptions: what is an appropriate care and how to measure it from the data (Canadian Medical Association, 2015)? This question is motivated by the growing necessity to adapt healthcare funding to a resource-constrained system (McGinnis et al., 2013; Aubert et al., 2019).

The following sections review new data collection mechanisms, highlights the different stance taken by machine learning over traditional statistics and precises critical questions in public health that could benefit from innovative methods.

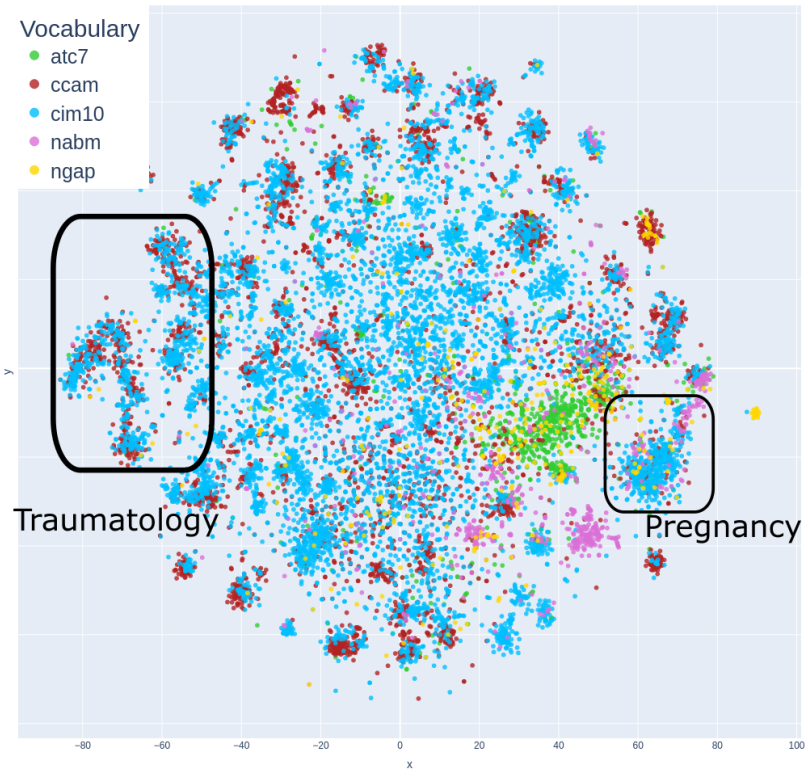


Fig. 1.1. Projection in two dimensions (TSNE) of medical event embeddings. Each point is a projection of the embedded vector for a given medical concept: Embeddings have been built from French Medical Claims (SNDS). Colors correspond to different medical vocabularies: drugs in green, billing diagnoses in blue, billing procedures in red, biology in pink, general practitioner (GP) activity in yellow. An interactive version of this plot is available at: `https://straymat.gitlab.io/event2vec/visualizations.html`

1.2 The data: Electronic Health Records

Background Information

1.2.1 From research-oriented to large scale routine care data collection

Traditional research data collections The gold standard for generating new evidence in healthcare are Randomized Controlled Trials (RCTs). Based on the randomization of patients to a treatment or a control group, the data collection is designed for one experiment, addressing one precise research question. RCTs are at the heart of *evidence-based medicine*, which promotes a hierarchy of evidence in which randomized experiments are superior to natural –uncontrolled– experiments or expert opinion (Guyatt et al., 1995). The modern trial methodology has been shaped by the large-scale International Studies of Infarct Survival (ISIS) experiments (ISIS-1 Collaborative Group, 1986). Two other common research data collections are cohorts –any designated group of individuals followed or traced over a period of time (Porta, 2014)– and registry –covering exhaustively a well defined clinical population. All these types of data are costly to collect and most of the time cover small samples of carefully selected patients.

More data collected routinely – Real World Data Healthcare data is increasingly collected from electronic information systems used in routine care (Jha et al., 2009; Sheikh et al., 2014; Kim et al., 2017; Esdar et al., 2019; Kanakubo; Kharrazi, 2019; Liang et al., 2021; Apathy et al., 2021). The term Real World Data (RWD) has been coined to define these new kind of data, not primarily collected for research (FDA, 2021a; HAS, 2021; Kent et al., 2022). Two major sources of RWD are insurance claims and EHRs.

1.2.2 Two major routine care data collection

Insurance claims – Good population coverage, low granularity Healthcare insurance systems collect massive amount of data, such as the coding of diagnoses and procedures for hospitals reimbursements, or patient prescriptions in city care. They usually have a good temporal and geographic coverage of the population –especially in countries with universal healthcare insurance. However, they fall short of clinical features, exam results, social background or reason for seeking care (Ziegler et al., 2022). Finally, billing optimization processes endanger the validity of billing variables by over-representing well-reimbursed cares (Juven, 2013).

Clinical data – EHRs and Hospital Information System EHRs are defined as the longitudinal collection of health data in an electronic information system (IS) (Gunter; Terry, 2005). As of 1990, the computerization of paper-based patient records incentivized by national foundings led to the wide adoption of EHR solutions. For example, in the United States 80% of hospitals (Adler-Milstein et al., 2017) and close to 90% of office-based practices (Quick-Stat, 2023) have now adopted EHRs. Used routinely by clinicians, EHRs enable them to record and interrogate clinically relevant information for patient care mainly through notes. EHRs can be specific to an institution, or shared between several actors (Hoerbst; Ammenwerth, 2010). This central system is accompanied by other business applications, such as patient administrative management, computerized prescribing, biology software, reanimation software, and imaging software. Together, these softwares make up the Hospital Information System (HIS). Depending on the degree of maturity of the HIS, the various data sources communicate more or less well with each other.

As early as 1990, the artificial intelligence in medicine community emphasized the importance of EHR (Shortliffe, 1993). These systems are a precious collection of natural experiments, allowing to complete the understanding of genotype-environment interactions with phenotypes (Butte; Kohane, 2006; Patel et al., 2009). Interest in EHR studies developed during the early 2000s highlighting their increasing rich data modalities. EHRs also provide a potential alternative to costly traditional data collections with decreasing financial supports. Casey et al., 2016 estimated **the average cost per participant in traditional studies of cardio-vascular disease risk factors between US$2,700 and US$17,700 compared to US$0.11 for a recent EHR.** Today's interest of EHR for research has been acknowledged in artificial intelligence in medicine (Yu et al., 2018), clinical research (Cowie et al., 2017) and epidemiological studies (Casey et al., 2016; Gianfrancesco; Goldstein, 2021). Opportunities and challenges of EHRs are further discussed and illustrated with an overview of the French situation in Chapter 2.

1.2.3 Interventional data vs observational data

Treatment allocation is a fundamental difference between RCTs and RWD Consider the toy example of assessing if the probability of death is influenced by the administration of a drug for treated patients compared to no administration for control patients. In observational data, interventions are far from random but focused on patients requiring care. In observational data –illustrated in Figure 1.2a), the drug would be given to patients with more comorbidities –assessed for illustration purposes by the Charlson score (Charlson et al., 1987) shown on the x axis. Treated and control populations depart from each other in such way that it is doubtful whether the difference in results can be attributed to the

Background Information

treatment alone. Such population discrepancies call for dedicated estimation methodologies –presented in Section 4.2.3.

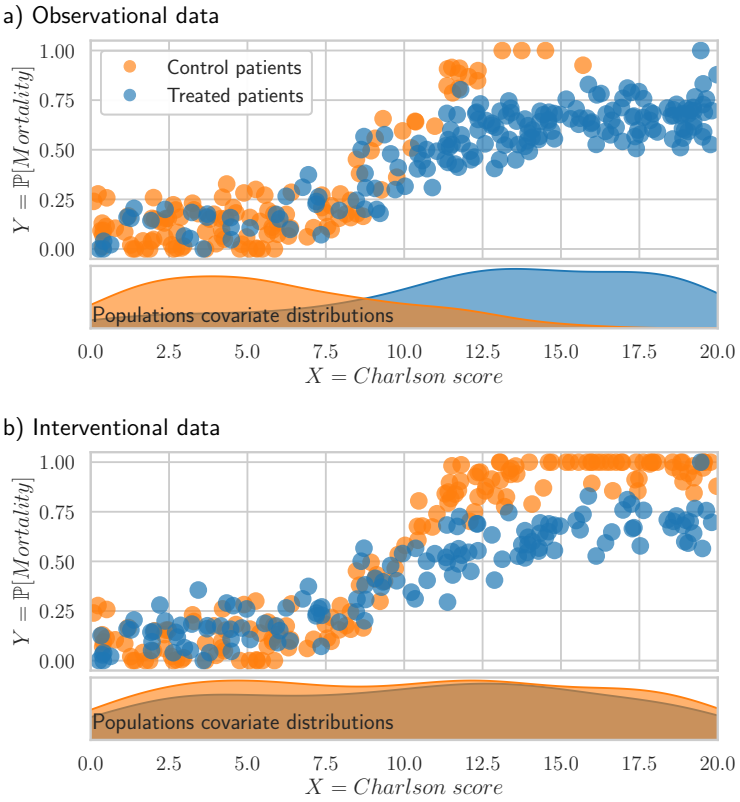


Fig. 1.2. a) In routine care, treatments allocation is not random: priority is often given to patients that will benefit the most from the intervention.
b) In interventional studies, randomization ensures the comparability of treated and control population on average.

For interventional data –shown in Figure 1.2b), the randomization forces the probability of receiving the treatment for every patient –often to 50%. Importantly, it is independent from patient characteristics such as the Charlson score. This favors the comparability between treated patients and control patients: Once this randomization has been repeated over many patients, *on average*, the difference in outcomes can only be attributed to the treatment.

Background Information

Rothman, 2012 distinguishes RCTs from observational data as follows: *In an experiment, the reason for the exposure assignment is solely to suit the objectives of the study; if people receive their exposure assignment based on considerations other than the study protocol, it is not a true experiment.* Epidemiological textbooks highlight the experimental setup since randomization yields excellent *internal validity* (Campbell, 1957): The estimated average treatment effect is close to the true treatment effect if we could repeat the experience indefinitely on the same population –the average estimate is *unbiased*. Statistical assumptions for the RCT methodology are easily met (Colnet et al., 2020, Section 3.1.1) and involve the measurements of only two variables: the treatment and the outcome. This simplicity contributed to its success for clinical evidence generation.

External validity – When RCTs insufficiently inform practices Interventional data are hard to collect: patients must be voluntary, without comorbidities, adhering to the treatment. These requirements contribute to idealized populations recruited in RCTs, endangering the *external validity* of interventional studies (Feinstein; Horwitz, 1997; Concato et al., 2000; Rothwell, 2005): The findings of a study may not generalize to other populations.

As a practical consequence, RCTs may not apply to real world situations since included populations and experimental conditions differ too much from usual practices. **For example, only 6% of asthmatics would have been eligible for their own treatment RCTs** (Travers et al., 2007). An advantage of observational data over RCTs is their description of usual care practices, opening a window on care effectiveness (Cochrane, 1972): How well does a treatment work in practice, outside the ideal circumstances of the experimentation ?

External validity have been raised in economics earlier than in epidemiology (Deaton, 2020). The focus in such problems is probably greater in economics because situations are not well controlled: economic agents act outside of the laboratory. On the contrary, clinical situations are closer to laboratory settings. This difference in data sources and methodology is visible in the divide between clinical and social epidemiology (Zielhuis; Kiemeney, 2001). What about medico-economics ? The present thesis does not address this question, but it is a clear motivation.

Background Information

Heterogeneity of treatment – The link with personalized medicine The previous paragraph discussed average treatment effect. However, heterogeneity of treatment among subgroups is also an object of interest in healthcare (Hernàn; Robins, 2020). It is at the heart of the personalized medicine paradigm –also called precision medicine. Originally anchored into genomics, this concept tries to take individual variability into account for tailored treatment recommendations (Schork, 2015; Topol, 2019).

In sequential decision-making problems, heterogeneity of treatment effect is close to operational research (Schaefer et al., 2004) or reinforcement learning (Bareinboim et al., 2015). These fields are focused on sequential decision making processes with a large action-state space, rather than a one-off treatment as in this thesis. To robustly estimate the probabilities involved in large dimensional state-action spaces, these works often require a simulated environment (Bennett; Hauser, 2013). Simulations are convenient to collect large samples of trajectory implementing the policy to be evaluated but are hard to build for complicated problems such as patient trajectories. Bridges with healthcare and personalized medicine include application in optimizing antiretroviral therapy in HIV (Guez et al., 2008) or management of sepsis in the Intense Care Unit (Komorowski et al., 2018) and is an active area of research (Coronato et al., 2020).

1.3 Two cultures of statistics for health

Breiman, 2001b clearly distinguished two statistical cultures: a predominant community at the time focused on models, and an emerging trend that relies only on predictive accuracy. The latter is called machine learning.

1.3.1 Model-based statistics: oracle domain experts

Biostatistics puts a strong emphasis on model specification, often interpreting model parameters as real mechanisms (Cox, 2001). Maybe because epidemiology seek to estimate causes and effects, it also favors the model-based culture as one can judge from textbooks (Rothman, 2012): *In many epidemiological applications, it is the choice among effect measures that dictates the type of model the investigator ought to use.* In medical journals, Cox model for survival data and logistic regression for binary outcomes have been the standard for

publication. It is tempting to link explicit relations between variables in linear models to claims over the data. But interpretability of the model parameters is a strong assumption (Lipton, 2018), which relies on the well-specification of the model: The underlying variables and their relations built into the models should describe natural laws as in physics.

Despite judging model choice as *crucial to fruitful application* (Cox, 2006), this tedious task is left to experts, that should have an appropriate theoretical model of the problem (Cox, 2001). This practice is prone to circular reasoning where a model is chosen only to justify an established theory. **Crowdsourcing studies asking the same question to different teams of experts showed that the choice of model and features is far from consensual and yield substantially different quantitative and qualitative results.** Botvinik-Nezer et al., 2020 attributed this dispersion to different modeling choice in a crowdsourced analysis of brain imaging data by 70 teams. Schweinsberg et al., 2021 showed that theoretical constructs –how the model input variables are defined from raw data– also contribute to the heterogeneity of results. All of these studies followed the model-based culture, assuming specific linear models and interpreting their inner part as causal.

1.3.2 Machine learning: black-box predictive ability?

Departing from carefully designed functional forms for statistical models, algorithms emerged in the late 90s to automatically learn patterns in the growing body of complex data. Among them, one can cite random forests, gradient boosting, neural networks or support vector machines.

Supervised learning Statistical learning (Vapnik, 1999) is the theoretical framework underpinning machine learning. It has been introduced in the 1960s and popularized in the 1990s with the development of flexible pattern recognition models. It is concerned with empirical risk minimization: Finding good approximation of the features-outcome association. Appendix A.1 recalls the formal definition of the supervised regression problem. Appendix A.2 details the principles of random forest and gradient boosting as we are using these two algorithms in different chapters of this thesis.

Model selection – How to choose between multiple promising models Choosing the best model among a family of potential estimators is done by comparing their performance on an unseen test dataset. If we were to evaluate a model on the same data that it has been trained on, we would systematically favor flexible models that adapt better to the sampled training data. To avoid this process called *overfitting*, the analyst should separate the data on which an estimator is fitted –the training set– from the data used to chose the estimators –the test set. To avoid data loss, cross-validation (Stone, 1974) proceeds in multiple rounds of training and testing on the same data as illustrated in Figure 1.3. The data is divided in K folds –typically 5 to 10–. Then, each fold is part of the training set K-1 time and is used once as test set. Algorithms are then selected on the average performance over the K test sets.

Note, that despite being valid for model selection –i.e. choosing the best model, cross-validation yields biased estimates to evaluate a given model (Wager, 2020a). Nested cross-validation is a classic way to perform this final model evaluation (Varoquaux et al., 2017).

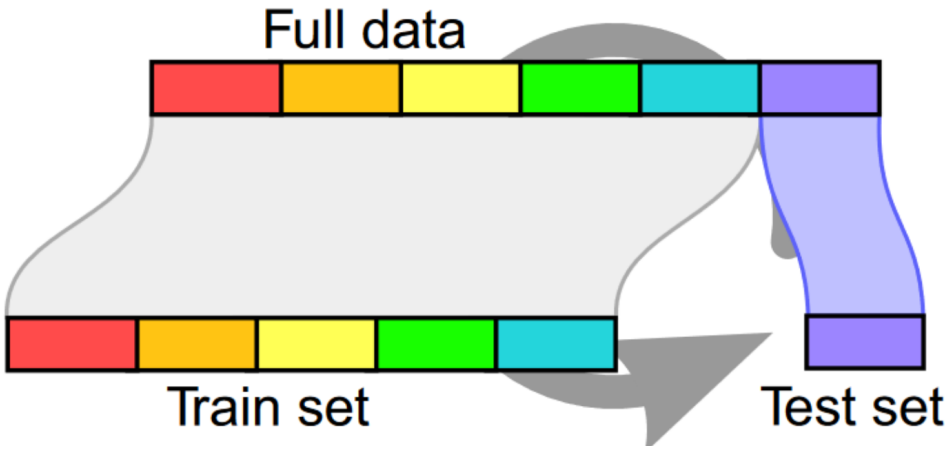


Fig. 1.3. Cross-validation: the data is split into training and testing sets. The training set is used to fit the model. The testing set is used to evaluate the model. The process is repeated multiple times to avoid overfitting. Original figure from Varoquaux et al., 2017.

Background Information

Pre-training – From learning patterns to learning representations As of 2012, the success of deep neural networks in computer vision (Krizhevsky et al., 2012), sparkled interest for representation learning. Grounded in information theory, this subfield of machine learning is concerned in automatically building low dimensional features from raw data that capture *useful information* (Bengio et al., 2013). A good representation should keep information on the output y and loose non-useful information in the input x, making it robust to noise: I refer to Achille; Soatto, 2018 for a detailed formalization. This definition of usefulness highlights the importance of finding pretext tasks to supervise the learning of representations. For example, in computer vision, the pretext task is to label coarse classes of images (e.g., dog, cat, container, ...). In NLP, the pretext task is to predict randomly masked world in a text. These tasks should be weakly-supervised in the sense that they do not require human annotations. The learned representations are then used as input to a supervised task, which should be close enough from the pretext task to leverage information retained during pre-training.

This thesis was originally motivated by studying deep representations for patient trajectories and their ability to transfer between healthcare databases. However, the choice of a representation loops back to the choice of a model, and hence to the measure of performance for relevant downstream tasks. For other domains such as NLP, consensual downstream tasks such as text classification have been established to evaluate the performance of representations. Aggregated in benchmarks such as GLUE (Wang et al., 2018), these tasks fueled model developments for years. Recently, Raji et al., 2021 pointed out the risk of misalignment between machine learning benchmarks and research claims about a task or real world objectives –what they call construct validity. This criticism is also pertinent for RWD since interests in using these data are fragmented –as developed in the next section. Moreover, healthcare interest in pure predictive problems is unclear –see Section 4.1. Finally, the circulation of data or representations in healthcare are almost non-existent due to privacy requirements. These specific features of RWD call into question the appropriateness of constructing universal representations in healthcare.

1.3.3 One choice of perspective: Recent statistical learning for EHRs

In this thesis, we focus on EHRs, even if some of our results could be relevant for insurance claims as well. The complexity of using EHRs in statistical frameworks is linked with their potential to address complex questions in public health: they register almost all aspect of the patients care trajectories in the hospital. Therefore, the dimensionality of EHRs data is high: diagnoses, laboratory measurements and procedures are logged with medical terminologies having ten of thousands of codes; unstructured text data is by nature high dimensional. Finally, the temporal dimension of EHRs poses a challenge for statistical models, usually assuming a static vector of features for each unit. For such high dimensional data where the number of (selected) features could easily reach the hundreds, linear models are not strictly more interpretable than deep neural network (Lipton, 2018). Therefore, it is tempting to turn towards machine learning techniques to leverage the full potential of EHRs. Are flexible statistical models a valid tool to answer today's questions in healthcare? If we are concerned with identifying which levers to pull to improve healthcare, we might not need to specify in our statistical models all mechanisms involved in the care process and its interaction with the complex EHR measurement system. One can hope that sufficient amount of data would allow flexible algorithm to detect useful regularities in the data and learn from them (Halevy et al., 2009).

1.4 Important questions in public health

1.4.1 The promises of RWD: what pressing needs to use health data

The recent increase in RWD collection is attracting the attention of many players. The optimism on the promises of healthcare data is shared by epidemiologists (Mooney et al., 2015; Hernán; Robins, 2016), artificial intelligence in medicine researchers (Schwartz et al., 1987; Yu et al., 2018), clinical researchers (Schwalbe; Wahl, 2020; Dzau, 2023), the industry (Pfizer, 2019; IQVIA, 2023) and government bodies (McGinnis et al., 2013; FDA, 2018; EMA, 2023). New scientific journals such as the *NEJM AI*[1] (Beam et al., 2023) at the intersection between algorithmic advances and clinical practices demonstrate the sustained interest in leveraging RWD for improving healthcare.

Drawing a full picture of these expectations would be over-ambitious and partial. But it is relevant to disentangle them in order to delimitate clearly the scope of questions that motivate my work. As discussed in Section 1.1, these motivations stem from a machine learning formation and public health concerns.

Primary uses of health data focus on patient care EHRs are primarily used to record patients' health state and cares (Safran et al., 2007; Datalink, 2022). This information might be shared within the healthcare team, but always with the goal to cary for the patient whose data is collected. Direct benefits to the patients are provided thanks to more detailed and searchable information than paper-based medical records. These benefits are amplified in modern care systems where patient trajectories are fragmented between numerous healthcare professionals.

From automating the care workflow to personalized medicine The automation of tedious tasks in healthcare is part of the original agenda of artificial intelligence in medicine

[1] https://ai.nejm.org/

(Schwartz et al., 1987). The goal is to give more time to the physicians by accelerating and facilitating parts of their work that require poor analytical capacities.

Recent discourse and successes are heavily influenced by this line of though, with the aim to automate ever more specialized parts of care. Machines read medical images faster and more efficiently than most practitioners (Zhou et al., 2021a). Machine learning algorithms trained on structured data from EHR (Rajkomar et al., 2018b) or administrative databases (Beaulieu-Jones et al., 2021) outperform rule-based clinical scores in predicting patient's readmission, in-hospital mortality or future comorbidities (Li et al., 2020b). Recently, large language models (LLMs) leveraged clinical notes from several hospitals for length of stay prediction (Jiang et al., 2023). Hope is high that LLM models will soon be able to help practitioners during consultation (Lee et al., 2023). The trend towards personalized medicine is gradually moving away from the automation of well-understood but tedious tasks to tailored individual care, where mechanisms are less understood (Schork, 2015; Topol, 2019).

New data for better knowledge acquisition? RWD data also bring indirect benefits –secondary uses– by accelerating and improving knowledge production: on pathologies (Campbell et al., 2022), on the conditions of use of health products and technologies (Safran et al., 2007; Tuppin et al., 2017), on the measures of their safety (Wisniewski et al., 2003). They can also be used to assess the organizational impact of health products and technologies (HAS, 2020; HAS, 2021). These descriptive usages are closer to the main goal of epidemiology: *the study of the distribution and determinants of disease frequency* (MacMahon, Pugh, et al., 1970).

Health Technology Assessment (HTA) agencies in many countries have conducted extensive work to better support the generation and use of RWD (FDA, 2021a; HAS, 2021; Kent et al., 2022; Plamondon et al., 2022). Study programs have been launched by regulatory agencies: for example, the *DARWIN EU program* by the European Medicines Agency and the *Real World Evidence Program* by the Food and Drug Administration (FDA, 2018). However, recent surges in data collection also encouraged deviation from the standard interventional design by relaxing some of the methodological constraint of RCTs. HTA agencies witnessed a deterioration of evidence for new drugs and strongly advocate against observational studies for replacing new drugs evaluation (Wieseler et al., 2023; Vanier et al., 2023). This debate crystallizes tensions between the pharmaceutical industry and regulators since drug prices are mainly driven by drug efficacy –assessed in trials. There are still active debates to better understand what is the place of these data to develop new evidence on the effectiveness –def. in Section 1.2– of interventions (Richesson et al., 2013; Wang et al., 2023b).

Public health questions – Old and new The *Coming revolution In Medicine* is described by Rutstein, 1967 as:*1) modern medicine's skyrocketing costs; 2) the chaos of an information explosion involving both paperwork proliferation and large amounts of new knowledge that no single physician could hope to digest; 3) a geographic maldistribution of M[edical] D[octors]s; 4) increasing demands on the physician's time as increasing numbers of individuals began to demand quality medical care.* **Fifty years later, the same preoccupations are destabilizing healthcare systems in rich countries:** increasing costs (OECD, 2023), knowledge produced too quickly to be assimilable by a single person (McGinnis et al., 2013), geographic disparities and lack of trained physicians (Anguis et al., 2021; AAMC, 2021).

In this context, public health authorities are asked to better understand what cares are the most effective. With constrained medical resources, efforts should be focused on the most effective interventions and prevent unnecessary cares. However, measuring the appropriateness of a care is a delicate question (Canadian Medical Association, 2015).

Scientific associations and regulators issue medical guidelines built upon the scientific literature to recommend ideal care trajectories. However, these recommendations often focus on single-disease approaches, insufficiently covering today's population of patients: increasingly old and multimorbid (Skou et al., 2022). **Fewer than half of the clinical guidelines for the nine most common chronic conditions consider older patients with multiple comorbid chronic conditions** (Boyd et al., 2005; Parekh; Barton, 2010). In the United States, the Committee on Learning Healthcare System suggested to better use routine data to adapt evidence to real practice and accelerate its diffusion (McGinnis et al., 2013).

In line with this roadmap, this work is motivated by the opportunities offered by EHRs to evaluate the effectiveness of medical guidelines. Formally, a guideline can be expressed as an action (or a series of actions) to take given a patient risk profile. The action/characteristic reflects the practice of physicians: they see and listen to a patient, then take appropriate actions given its profile. A medical recommendation tries to guide this link between patient characteristics and intervention.

1.4.2 Prediction or causation?

Is statistical learning a useful tool to evaluate these guidelines? There is a growing interest in predictive models in healthcare. This trend is reflected by the exponential increase in the proportion of publications per year in Pubmed shown in Figure 1.4. Nonetheless, the prognosis literature does not specifically emphasize prediction as its primary objective.

The early Framingham study concludes that risk reduction is more important than identifying the strength of specific risk factors since this quantity is subject to slight changes in the risk model (Brand et al., 1976): *It further suggests that the strength of a particular risk factor may not be as important from the point of view of intervention as the ability to safely and conveniently achieve even a moderate risk reduction in a large number of persons.* In a foundational article on EHR, heart failure prediction is motivated by aggressive interventions (Wu et al., 2010): *heart failure could potentially lead to improved outcomes through aggressive intervention, such as treatment with angiotensin converting enzyme (ACE)-inhibitors or Angiotensin II receptor blockers (ARBs).* More recently, Beam; Kohane, 2018 devise a machine learning spectrum, making the distinction between algorithms requiring heavy human assumptions and flexible models. But the goal of these algorithms is almost always to serve *decision-making*, not prediction. It is clearly described by Steyerberg, 2009 for diagnosis: *If we do a diagnostic test, we may detect an underlying disease. But some diseases are not treatable, or the natural course might be very similar to what is achieved with treatment.* Modeling has always been judged necessary but it is only recently that pattern recognitions is pursued as an objective per-se, associated with an exponential disinterest for causality as shown in Figure 1.4. Patel et al., 2009 discussed the inappropriateness of the unsupervised and supervised approaches: They *tend to discover relatively simple relationships in data and have not yet demonstrated the ability to discover complex causal chains of relationships.* This make them misaligned with scientific and practical objectives which are to *formulate and test hypotheses about how the human organism "works" in health and illness.*

The auxiliary role of prediction in healthcare calls for replacing or complementing statistical learning with another methodological tool.

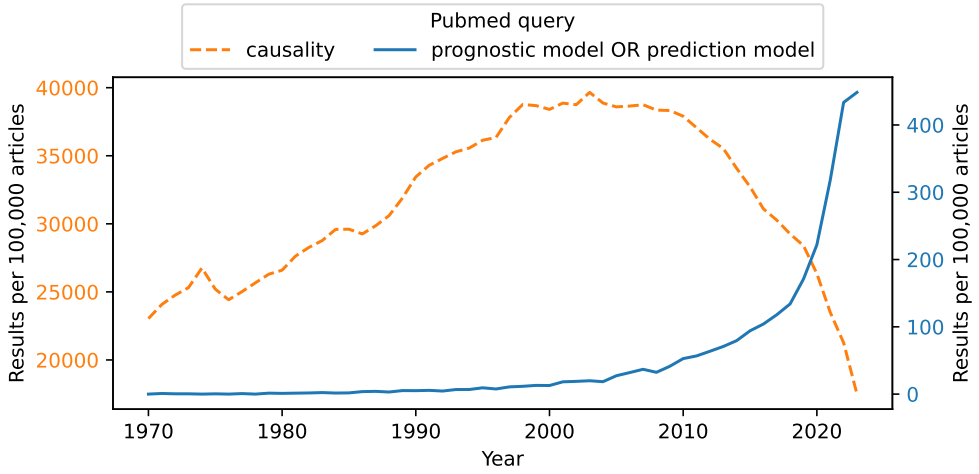


Fig. 1.4. Proportion of articles by year in Pubmed returned by queries on causality (orange) or predictive modeling (blue). The scale differs greatly but the symmetry of the trends is disturbing.

Background Information

The central concept of causality Causality is a central concept in epidemiology (Hill, 1965; Hernàn; Robins, 2020) and has been developed formally in statistics (Rubin, 1974), econometrics (Imbens; Wooldridge, 2009) and machine learning (Pearl; Mackenzie, 2018).

Causality departs from classical statistics by stressing the importance of interventions: the strong correlation between an outcome and a feature is no evidence of a causal link between them. This central point has been depicted as the ladder of causation (Pearl; Mackenzie, 2018): For a given set of observations, multiple consistent causal models exist, only one of which correctly reflects the reality. For concreteness, consider the didactic example from Veitch et al., 2022: *Consider three possible explanations for the association between ice cream and drowning. Perhaps eating ice cream does cause people to drown—due to stomach cramps or similar. Or, perhaps, drownings increase demand for ice cream—the survivors eat huge quantities of ice cream to handle their grief. Or, the association may be due (at least in part) to a common cause: warm weather makes people more likely to eat ice cream and more likely to go swimming (and, hence, to drown). Under all three scenarios, we can observe exactly the same data, but the implications for an ice cream ban are very different. Hence, answering questions about what will happen under an intervention requires us to incorporate some causal knowledge of the world — e.g., which of these scenarios is plausible?*

Causality also has connections with other important issues in healthcare such as fairness (Plecko; Bareinboim, 2022) or dataset shift, linked to the representativity of a dataset (Subbaswamy; Saria, 2020). We did not explore these aspects, but they motivate part of the work in Chapter 4.

A robust statistical framework – Neyman-Rubin I recall the framework of the potential outcomes, which enables statistical reasoning on causal treatment effects (Imbens; Rubin, 2015). I will progressively introduce supplementary concepts as needed in Chapters 4 and 5.

Given an outcome $Y \in \mathbb{R}$ (e.g., mortality risk or hospitalization length), function of a binary treatment $A \in \mathcal{A} = \{0, 1\}$ (e.g., a medical procedure, a drug administration), and baseline covariates $X \in \mathcal{X} \subset \mathbb{R}^d$, we observe the factual distribution, $O = (Y(A), X, A) \sim \mathcal{D} = \mathbb{P}(y, x, a)$. However, we want to model the existence of potential observations (unob-

served ie. counterfactual) that correspond to a different treatment. Thus we want quantities on the counterfactual distribution $O^* = (Y(1), Y(0), X, A) \sim \mathcal{D}^* = \mathbb{P}(y(1), y(0), x, a)$.

Popular quantities of interest –estimands– are: at the population level, the Average Treatment Effect

ATE $$\tau \stackrel{\text{def}}{=} \mathbb{E}_{Y(1),Y(0)\sim\mathcal{D}^*}[Y(1) - Y(0)];$$

at the individual level, to model heterogeneity, the Conditional Average Treatment Effect

CATE $$\tau(x) \stackrel{\text{def}}{=} \mathbb{E}_{Y(1),Y(0)\sim\mathcal{D}^*}[Y(1) - Y(0)|X = x].$$

We might be tempted to equate the mean of a potential outcome to the mean of the observed outcome conditionally on the treatment. However, in general: $\mathbb{E}[Y(a)] \neq \mathbb{E}[Y|A = a]$. This is because the right hand side reads as *the expected value of Y given* $A = a$, thus we restrict the outcome to the individuals that actually received the treatment. If this group has a different potential outcomes distribution than the other group, the equality does not hold as illustrated in Figure 1.2.

The statistical assumptions required to estimate these quantities from observational data are presented in Section 4.2.2. The common methods used for the estimations are detailed in Appendix D.4.1.

1.5 Contributions

First influenced by the growing successes of machine learning in predictive modeling, this work seeks to understand what models and framework might help evaluating the effectiveness of clinical guidelines using the time varying and high dimensional EHRs. What predictive models are useful ? Why prediction is not enough ? How flexible models can serve the true objective of healthcare: delivering appropriate treatment to each and every patient to improve her health (Canadian Medical Association, 2015)?

The contributions of each chapter, which are summarized below, have led to three articles as first-authors and one work in progress:

- Chapter 2 published in *PLOS Digital Health*,
- Chapter 3 is ongoing work,
- Chapter 4 is being finalized for submission,
- Chapter 5 submitted to *Artificial Intelligence in Medicine*.

Beyond these works, applied projects have also been conducted, leading to one other research work linked to invasive ICU treatment disparities, co-authored with Sara Mohammed, João Matos, Leo Anthony Celi and Tristan Struja. This work has been accepted to *Yale Journal of Biology and Medicine* (third-position author). On this project I helped on the design of the statistical analysis, more particularly on the model selection procedure with machine learning algorithms.

This work is a specific application separated from the thesis main questions. Therefore it is not detailed in this manuscript.

Chapter 2: Potential and challenges of Clinical Data Warehouse, a case study in France This chapter draws the first overview of the Clinical Data Warehouses (CDWs) in France. These technical and organizational infrastructures are emerging in the hospitals to collect and analyze the data produced in EHRs. This work is an attempt to better characterize the reality of data reuses in university hospitals. It documents key aspects of the collection and organization of routine care data into homogeneous databases: governance, transparency, types of data, data reuse main objectives, technical tools, documentation and data quality control processes.

We show that the emerging landscape of CDWs in France is highly heterogeneous and mostly focused on research or piloting. We highlight the necessity to create or perpetuate multidisciplinary warehouse teams capable of operating the CDW and supporting the various projects. The multi-level collaborations allow to mutualize resources and skills at the regional or national levels. We report poor data documentation and unequal adoption of internationally recognized common data models. Finally, we encourage to expand the scope of data beyond the hospital to better include city care. The qualitative aspect of this chapter contrasts with the general mathematical context of the thesis.

Chapter 3: Exploring a complexity gradient in representation and predictive models for EHRs Acknowledging the growing interest in predictive algorithms for EHR data, this chapter introduces two simple feature construction methods taking raw medical events as input features before feeding a predictive model. It benchmarks four predictive pipelines of increasing complexity on three predictive tasks: length of stay interpolation, next visit prognosis and cardiovascular adverse events prediction. It focuses on medium sized datasets where the population at risk (after inclusion and exclusion rules) ranges from 10,000 to 20,000 samples. In these setups, this work explores the complexity-performance tradeoff from simple baseline models to recent transformer-based neural networks.

We show that for these medium sample-size settings, simple baselines outperform transformer-based models both in predictive accuracy and computing resource efficiency. We note a performance decrease for prognosis with low case prevalences. To encourage more thorough study, we publish scikit-learn compatible implementations of our proposed medical event featurization pipelines.

Chapter 4: Prediction is not all we need: Causal thinking for decision making on EHRs This chapter exploits the causal framework to build clinically valuable models. It shows that predictions –even accurate as with machine learning, may not suffice to provide optimal healthcare for every patient. Anchored in causal thinking, it details key elements necessary to robustly estimate treatment effect from time-varying EHR data. We present a step-by-step framework to help build valid decision making from real-life patient records by emulating a randomized trial before individualizing decisions, e.g., with machine learning. We illustrate the various choices in studying the effect of albumin on sepsis mortality in the Medical Information Mart for Intensive Care database (MIMIC-IV). We study the impact of various choices at every step, from feature extraction to causal-estimator selection.

We find that subtle bias such as immortal time bias can change the conclusion of a study. However, we show that these errors can be captured by following a careful trial emulation design and by comparing different modeling hypotheses in a vibration analysis. We validate our estimation of average treatment effect with RCT gold-standards and inspect heterogeneous treatment effects in subpopulations. In a tutorial spirit, the code and the data are openly available.

Chapter 5: How to select predictive models for causal inference? This chapter built on the variability of results presented in chapter 4. Can we explain why some models yield better treatment effect estimation than others? Statistical learning theory establishes how to select models for prediction. This chapter shows that classic machine-learning model selection does not pick the best models for causal inference. We review more elaborated risks developed in the causal inference literature. These risks rely on the estimation of nuisances that allow for the identification of the causal effect. However, these causal risk have not been empirically evaluated in a wide variety of finite sample settings. Drawing from an extensive empirical study, this chapter study the performance of five causal risk to select an outcome model for treatment effect estimation.

Our results highlight that estimators for causal inference should be selected, validated, and tuned using different procedures and error measures than those classically used to assess prediction. Rather, selecting the best outcome model according to the R-risk leads to more valid causal estimates. Despite relying on the estimation of two nuisances, this risk outperform others risks. We also show theoretically that the R-risk is a reweighted version of the oracle unobserved risk between predicted and potential outcomes. This property lead to accurate estimation of treatment heterogeneity when treated and untreated population differ little, as in RCTs. To facilitate better model selection, we provide python code implementing our procedure.

1.6 Résumé extensif en Français

Pourquoi étudier l'inférence causale à partir des dossiers patients informatisés

***Des données qui cherchent une question [...] et posent des énigmes* (Cox, 2001)**

L'apprentissage automatique a connu de grands succès en traitement automatique du langage (TAL) et en analyse d'image, grâce à l'exploitation de grandes quantités de données de piètre qualité, faiblement labellisées. D'autres champs d'applications peuvent-ils bénéficier d'une telle approche ? Parce que *la pratique médicale et la recherche biomédicale, [sont] intrinsèquement des tâches de gestion de l'information* (Patel et al., 2009), de nombreux chercheurs prévoient un grand potentiel d'amélioration des soins grâce à l'apprentissage automatique appliqué à de nouvelles collections de données (Topol, 2019; Rajkomar et al., 2019).

Il existe actuellement une tension entre la collecte massive de données de soins de routine, telles que les données de remboursement ou les dossiers patients informatiques (DPI), un nouveau cadre statistique (l'apprentissage automatique (Breiman, 2001b)), et des questions analytiques urgentes en matière de santé publique. Pour comprendre l'importance et les défis de l'inférence causale à partir des DPI, examinons trois exemples concrets.

Etudier les statistiques pendant la révolution du traitement automatique du langage

Dans les années 2010, Halevy et al., 2009 a proposé de tirer parti des régularités présentes dans de grandes masses de données pour concevoir automatiquement des variables pertinentes pour de multiples tâches applicatives. Considérons un modèle destiné à classifié des images à partir d'une grande collection de pairs d'images et d'étiquettes. Au cours de l'entraînement, le modèle apprend progressivement des représentations internes d'images naturelles telles que des pieds, des visages ou des mains. Ces représentations intermédiaires peuvent être réutilisées –transférées– à des tâches applicatives ayant des objectifs différents, telles que prédire la gravité d'un traumatisme à partir d'une photo. Ce paradigme, appelé préapprentissage a connu un grand succès dans des domaines appliqués tels que la vision par ordinateur (Krizhevsky et al., 2012), puis le traitement du langage naturel (TAL) (Devlin et al., 2018) et la biologie structurelle (Jumper et al., 2021). Il est pertinent de savoir si le pré-entraînement pourrait être appliqué à d'autres domaines avec des données complexes, tels que le soin.

Les hôpitaux de Paris, un vaste dépôt de notes cliniques

En 2017, le nouvel entrepôt de données de santé des Hôpitaux de Paris (AP-HP), ambitionnait d'exploiter son vaste référentiel de notes cliniques à des fins de recherche. Pourtant, les données de soins de routine ne sont pas un matériau familier pour la recherche médicale. Pour les ingénieurs et les chercheurs en TAL, ces données diffèrent de celles de la recherche traditionnelle car elles requièrent de nouveaux outils, par exemple provenant du TAL afin de gérer la complexité et l'échelle inédites des processus de collecte de données. Pour les épidémiologistes et les médecins, la principale différence réside dans l'opposition entre données expérimentales et données observationnelles. Cette difference de nature apparaît comme plus importante pour comprendre comment bien mobiliser les données de routine pour améliorer le soin.

Les données de facturation : de nouvelles données pour la santé publique?

En 2018, la direction des statistiques du ministère français de la santé a extrait des milliards de données de remboursement depuis la plateforme d'exploitation de l'assurance maladie au sein d'un serveur dédié. Pour améliorer et accélérer l'analyse sur cette nouvelle plateforme, elle s'est appuyée sur les travaux de Bacry et al., 2020 qui tirent parti du calcul parallèle et sur de la documentation collaborative (HDH, 2023a). Cependant, la description des séquences de soins est restée complexe en raison du grand nombre d'événements médicaux différents utilisés dans ces données. Ce problème, connu sous le nom de *malédiction de la grande dimension* a été introduit en recherche opérationnelle (Bellman, 1957) et est bien connu en statistique (Breiman, 2001b) : Le nombre d'échantillons requis pour développer des algorithmes prédictifs croît généralement de manière exponentielle avec le nombre de dimensions. Les séquences d'événements de facturation peuvent être considérées comme des séquences de signes, tout comme le texte. Ainsi, des travaux en informatique médicale, s'inspirant des techniques de TAL, consiste à créer des représentations de faible dimension –embeddings– de concepts médicaux (Beam et al., 2019). Les relations entre les événements capturées par ces représentations sont étonnamment proches des associations connues, comme le montre la figure 1.1. Cependant, l'utilité avale de ces représentations reste incertaine. La santé publique est peu intéressée par les domaines de représentation de l'information ou même par des tâches de prédictions. Les services publics cherchent à comprendre l'hétérogénéité des consommations de soins : qu'est-ce qu'un soin approprié et comment le repérer à partir des données ? (Canadian Medical Association, 2015) ? Cette question est motivée par la nécessité croissante d'adapter le financement des soins à un système aux ressources limitées (McGinnis et al., 2013; Aubert et al., 2019).

Contributions

Initialement influencé par les succès croissants de l'apprentissage automatique pour la modélisation prédictive, ce travail cherche à comprendre quels modèles et quel cadre sont appropriés pour évaluer l'efficacité des recommendations de bonnes pratiques en santé à partir des données de vie réelle. Quels sont les modèles prédictifs utiles ? Pourquoi la prédiction n'est-elle pas suffisante ? Comment des modèles flexibles peuvent-ils contribuer aux véritables objectifs du soin : fournir un traitement approprié à chaque patient pour améliorer sa santé (Canadian Medical Association, 2015) ?

Les contributions de chaque chapitre –résumées ci-dessous, ont donné lieu à trois articles en tant que premier auteur et à un travail en cours :

- Le chapitre 2 est publié dans *PLOS Digital Health*,
- Le chapitre 3 est un travail en cours,
- Le chapitre 4 est en cours de finalisation pour soumission,
- Le chapitre 5 est soumis à *Artificial Intelligence in Medicine*.

Chapitre 2 : Opportunités et obstacles rencontrés par les entrepôts de données cliniques cliniques, une étude de cas en France Ce chapitre présente la première vue d'ensemble des entrepôts de données de santé hospitaliers (EHDS) en France. Ces infrastructures techniques et organisationnelles émergent dans les hôpitaux afin de collecter et analyser les données produites en routine. Ce travail tente de mieux caractériser la réalité des réutilisations de données dans les Centres Hospitaliers Universitaires. Il documente les

aspects clés de la collecte et de l'organisation des données de soins de routine dans des bases de données homogènes : gouvernance, transparence, types de données, objectifs principaux de la réutilisation des données, outils techniques, types d'analyse, documentation et processus de contrôle de la qualité des données.

A partir d'entretiens semi-dirigés, nous montrons que l'écosystème naissant des EDSH en France est très hétérogène et principalement axé sur la recherche ou le pilotage. Nous soulignons la nécessité de créer ou de pérenniser des équipes d'entrepôts pluridisciplinaires capables d'opérer et d'exploiter l'EDSH afin de soutenir les différents projets de données. Les collaborations à plusieurs échelles permettent de mutualiser les ressources et les compétences au niveau régional ou national. Nous constatons une faible documentation des données et une adoption inégale des modèles de données communs pourtant reconnus au niveau international. Enfin, nous encourageons une extension du champ des données au-delà de l'hôpital pour mieux inclure les soins de ville. L'aspect qualitatif de ce chapitre contraste avec le contexte général de la thèse, plus mathématique.

Chapitre 3 : Exploration d'un gradient de complexité pour les modèles prédictifs à partir de DPI Constatant l'intérêt croissant pour les algorithmes prédictifs à partir de données de DPI, ce chapitre introduit deux méthodes simples de construction de variables prenant les événements médicaux bruts en entrée avant d'alimenter un modèle prédictif. Il compare quatre pipelines prédictives de complexité croissante sur trois tâches médicales : classification de la durée du séjour, pronostic de la prochaine visite et prédiction d'événements cardiovasculaires indésirables. Ce travail se concentre sur des ensembles de données de taille moyenne où la population à risque (après les règles d'inclusion et d'exclusion) se situe entre 10 000 et 20 000 échantillons. Dans ces configurations, ce travail explore le compromis complexité-performance entre des modèles simples et des réseaux neuronaux récents à base d'architecture *transformer*.

Nous montrons que dans ces conditions de moyennes tailles d'échantillon, les modèles simples sont plus adaptés que les modèles à base de *transformer*, tant en termes de performance prédictive que d'efficacité en ressources de calcul. Nous constatons une diminution des performances pour les tâches pronostiques avec de faibles prévalences. Pour encourager l'étude plus approfondie de ces méthodes, nous publions les nouveaux modèles introduits avec une API scikit-learn.

Chapitre 4: La prédiction ne suffit pas: nécessité d'un cadre causal pour la prise de décision à partir des données de DPI Ce chapitre exploite le cadre causal pour concevoir des modèles d'aide à la décision utiles. Il montre que des prédictions –même précises comme avec l'apprentissage automatique, peuvent ne pas suffire à fournir des soins adaptés à chaque patient.

En tirant parti des principes de l'inférence causale, nous détaillons les éléments clés nécessaires pour estimer de manière robuste l'effet d'un traitement à partir de données de DPI variant dans le temps. Nous présentons des étapes détaillées permettant de développer des systèmes d'aide à la décision valides à partir des données de DPI grâce à l'émulation d'un essai randomisé. Nous illustrons ce guide par une étude de l'effet de l'albumine sur la mortalité due à la septicémie dans la base de données *Medical Information Mart for Intensive Care database* (MIMIC-IV). Nous étudions l'impact des différents choix d'analyses sur le résultat de l'étude, depuis l'extraction des caractéristiques des patients jusqu'à la sélection de l'estimateur causal. Nous constatons que des biais subtils, tels que le biais du temps immortel, peuvent modifier la conclusion d'une étude. Cependant, nous montrons que ces erreurs peuvent être capturées en émulant avec attention un essai randomisé hypothétique et

en comparant différentes choix de modélisation au sein d'une analyse de vibration. Nous validons notre estimateur de l'effet moyen du traitement à l'aide des résultats d'essais randomisés disponibles dans la littérature. Enfin, nous inspectons l'hétérogénéité de l'effet du traitement dans des sous-populations afin de guider le choix individuel de l'intervention. Dans un esprit didactique, le code et les données sont disponibles publiquement.

Chapitre 5 : Comment sélectionner des modèles prédictifs pour l'inférence causale ? Ce chapitre s'intéresse à la variabilité des résultats constatée dans le chapitre 4 pour différents choix d'estimateurs. Pouvons-nous expliquer pourquoi certains modèles permettent de mieux estimer l'effet du traitement que d'autres ? La théorie de l'apprentissage statistique établit comment sélectionner les modèles pour la prédiction. Ce chapitre montre que les méthodes de sélection utilisées en apprentissage automatique ne permettent pas de choisir les meilleurs modèles pour l'inférence causale. Nous passons en revue des risques plus élaborés présents dans la littérature d'inférence causale. Ces risques reposent sur l'estimation de nuisances qui permettent l'identification de l'effet causal. Cependant, ces risques causaux n'ont pas été évalués empiriquement pour une grande variété de contextes en échantillons finis. Ce chapitre étudie grâce à une étude empirique approfondie les performances de cinq risques causaux pour sélectionner un modèle d'estimation de l'effet de traitement.

Nos résultats montrent que les estimateurs pour l'inférence causale doivent être sélectionnés, validés et ajustés à l'aide de procédures et de mesures d'erreur différentes de celles utilisées classiquement en apprentissage statistique. La sélection du meilleur modèle à l'aide du risque R conduit à de meilleures estimations causales. Malgré le fait qu'il repose sur l'estimation de deux nuisances, ce risque est plus performant que les autres. Nous montrons également de manière théorique que le risque R est une version repondérée du risque non observé oracle entre les deux modèles d'outcomes potentiels. Cette propriété permet une estimation précise de l'hétérogénéité du traitement lorsque la population traitée et la population non traitée diffèrent peu, comme dans les essais randomisés. Pour faciliter la sélection des modèles, nous fournissons un code python mettant en oeuvre notre procédure.

Chapter 2

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

Souvent, entre [la construction des données et leur traitement ou leur interprétation], se dresse la "banque de données", qui fonctionne comme un sas de passage de l'une à l'autre. Or le monde de la "construction" est lui-même tendu entre deux façons de rendre compte de ses pratiques : la mesure, issue du langage des sciences de la nature, le codage conventionnel, inspiré, selon les cas, du droit, des sciences politiques, ou des sciences cognitives.
– Alain Desrosière, Entre réalisme métrologique et conventions d'équivalence : les ambiguïtés de la sociologie quantitative, 2001

Chapter's content

Despite increasing collection of routine care data, reusing it does not come free of charges. Attention must be paid to the entire life cycle of the data to create robust knowledge and develop innovation. In this chapter, we build upon the first overview of Clinical Data Warehouses (CDWs) in France to document key aspects of the collection and organization of routine care data into homogeneous databases: governance, transparency, types of data, data reuse main objectives, technical tools, documentation and data quality control processes. The landscape of CDWs in France dates from 2011 and accelerated in the late 2020, showing a progressive but still incomplete homogenization. National and European projects are emerging, supporting local initiatives in standardization, methodological work and tooling. From this sample of CDWs, we draw general recommendations aimed at consolidating the potential of routine care data to improve healthcare. Particular attention must be paid to the sustainability of the warehouse teams and to the multi-level governance. The transparency of the data transformation tools and studies must improve to allow successful multi-centric data reuses as well as innovations for the patient. The qualitative aspect of this chapter contrasts with the general mathematical context of the thesis. We have borrowed the methodology from the field of sociology on the advice of Professor Emmanuel Didier.

This chapter corresponds to the article entitled *Good practices for clinical data warehouse implementation: A case study in France* published to *PLOS Digital Health*,
Authors: Matthieu Doutreligne, Adeline Degremont, Pierre-Alain Jachiet, Antoine Lamer and Xavier Tannier.

Outline

2.1 Motivation and background: A changing world

2.1.1 Healthcare data collection is tightly linked with local organization

In practice, the possibility of mobilizing routinely collected data depends on their degree of concentration, in a gradient that goes from centralization in a single, homogenous Hospital Information Systems (HISs, 1.2.2) to fragmentation in a multitude of HIS with heterogeneous formats. The structure of the HIS reflects the governance structure. Thus, the ease of working with these data depends heavily on the organization of the healthcare actors.

Claims data are often centralized by national agencies In South Korea, the government agency responsible for healthcare system performance and quality (HIRA) is connected to the HIS of all healthcare stakeholders. HIRA data consists of national insurance claims (Kyoung; Kim, 2022). England has a centralized health care system under the National Health Service (NHS). Despite, not having detailed clinical data, this allowed the NHS to merge claims data with detailed data from two large urban medicine databases, corresponding to the two major software publishers (OpenSAFELY, 2023). This data is currently accessed through Opensafely, a first platform focused on Covid-19 research (OpenSAFELY, 2022). In the United States, even if scattered between different insurance providers, claims are pooled into large databases such as Medicare, Medicaid or IBM MarketScan. Lastly, in Germany, the distinct federal claims have been centralized only very recently (Kreis et al., 2016).

Clinical data are mostly distributed among many entities Despite different interoperability choices, large institutional clinical data-sharing networks begin to emerge. South Korea very recently launched an initiative to build a national wide data network focused on intensive care. United States is building Chorus4ai, an analysis platform pooling data from 14 university hospitals (CHoRUS, 2023). To unlock the potential of clinical data, the German Medical Informatics Initiative (Gehring; Eulenfeld, 2018) created four consortia in 2018. They aim at developing technical and organizational solutions to improve the consistency of clinical data.

Israel stands out as one of the rare countries that pooled together both claims and clinical data at a large scale: half of the population depends on one single healthcare provider and insurer (Clalit, 2023).

The case of France In France, the national insurer collects all hospital activity and city care claims into a unique reimbursement database (Tuppin et al., 2017). However, clinical

Background Information

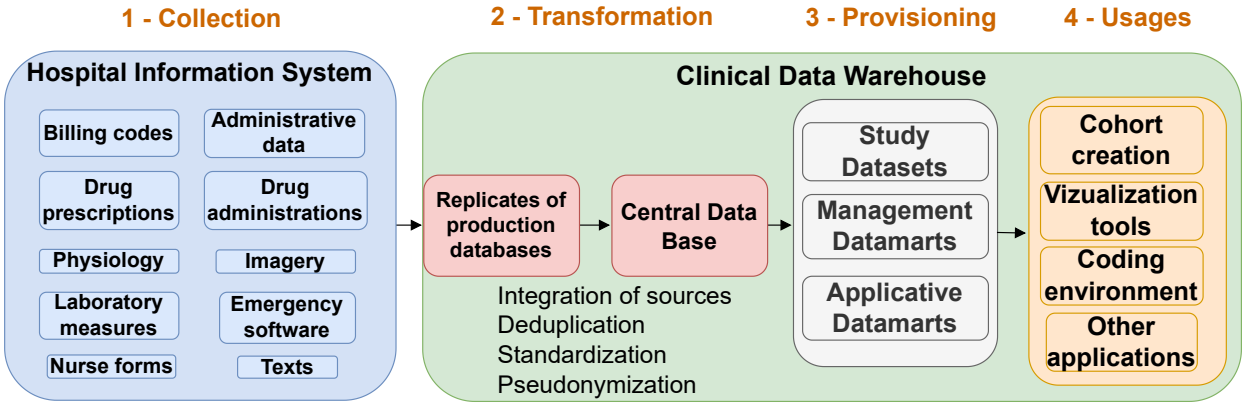


Fig. 2.1. Clinical Data Warehouse: Four steps of data flow from the Hospital Information System: 1) collection, 2) transformations and 3) provisioning.

Background Information

data is historically scattered at each care site in numerous HISs.

2.1.2 An infrastructure fo healthcare data : The Clinical Data Warehouses

Clinical Data Warehouse (CDW) Dedicated infrastructures are needed to pool data from one or more medical information systems –whatever the organizational framework– to homogeneous formats, for management, research or care reuses (Chute et al., 2010; Pavlenko et al., 2020). Fig 2.1 illustrates for a CDW, the four phases of data flow from the various sources that make up the HIS.

1. **Collection** and copying of original sources.
2. **Transformation**: Integration and harmonization
   - Integration of sources into a unique database.
   - Deduplication of identifiers.
   - Standardization: A unique data model, independent of the software models harmonizes the different sources in a common schema, possibly with common nomenclatures.
   - Pseudonymization: Removal of directly identifying elements.
3. **Provision** of sub-population data sets and transformed datamarts for primary and secondary reuse.
4. **Usages** thanks to dedicated applications and tools accessing the datamarts and data sets.

Multiplication of CDW in France For about ten years, several hospitals developed CDWs from electronic medical records (Cuggia et al., 2011; Jannot et al., 2017; Garcelon et al., 2017; Wack, 2017; Daniel et al., 2018; Malafaye et al., 2018; Artemova et al., 2019; Lelong et al., 2019; Conan et al., 2021; Lamer et al., 2022). This work has accelerated recently, with the growing development of dedicated infrastructures at the regional and

national levels. Regional cooperation networks are being set up –such as the Ouest Data Hub (Hugo, 2022). In July 2022, the Ministry of Health opened a 50 million euros call for projects to set up and strengthen a network of hospital Clinical Data Warehouses (CDWs) coordinated with the national platform, the Health Data Hub by 2025.

Poor understanding of CDW scope Despite these few examples, the precise scope of CDWs is still poorly understood: How are they created ? What data do they process ? How common are they ? What studies do they run ? Acknowledging the key importance of better structuring healthcare data, we create the first overview of CDWs in France. We build upon this landscape, to draw general recommendations aiming at consolidating the potential of routine care data reuse.

2.2 Speaking to the data collectors: Interviews of French University Hospitals

Based on an overview of university hospital CDWs in France, this study make general recommendations for properly leveraging the potential of CDWs to improve healthcare. It focuses on: governance, transparency, types of data, data reuse, technical tools, documentation and data quality control processes.

Interviews were conducted from March to November 2022 with 32 French regional and university hospitals, both with existing and prospective CDWs.

2.2.1 Interviews and study coverage

Semi-structured interviews We conducted semi-structured interviews on the following themes: the initiation and construction of the CDWs; the current status of the project and the studies carried out; opportunities and obstacles; quality criteria for observational research. Appendix B.1 lists all interviewed people with their team title.

We designed an interview form, sent to participants in advance. We used it as a support to conduct 90 minutes interviews recorded for reference (the complete form is available in Appendix B.2).

Based on these interviews, we collected structured information on both the characteristics of the actors, and those of the data warehouses. We completed them based on the notes taken during the interviews, the recordings, and by asking the participants for additional information. Detailed tables are available on `https://gitlab.has-sante.fr/has-sante/public/rapport_edsh/`.

2.2.2 A classification of observational studies

In addition to the interviews, we reviewed the study reporting portals, which we found for 8 out of 14 operational CDWs. We developed a classification of studies, based on the typology of retrospective studies described by the OHDSI research network (Schuemie, 2021). We enriched this typology by comparing it with the collected studies resulting in the six following categories. Studies were classified according to this nomenclature based on their title and description.

- **Outcome frequency**: Incidence or prevalence estimation for a medically well-defined target population.

- **Population characterization**: Characterization of a specific set of covariates. Feasibility and pre-screening studies belong to this category (Pasco et al., 2019).
- **Risk factors**: Identification of covariates most associated with a well-defined clinical target (disease course, care event). These studies look at association study without quantifying the causal effect of the factors on the outcome of interest.
- **Treatment Effect**: Evaluation of the effect of a well-defined intervention on a specific outcome target. These studies intend to show a causal link between these two variables (Hernán, 2021).
- **Development of diagnostic and prognostic algorithms**: Improve or automate a diagnostic or prognostic process, based on clinical data from a given patient. This can take the form of a risk, a preventive score, or the implementation of a diagnostic assistance system. These studies are part of the individualized medicine approach, with the goal of inferring relevant information at the level of individual patient's files.
- **Medical informatics**: Methodological or tool oriented. These studies aim to improve the understanding and capacity for action of researchers and clinicians. They include the evaluation of a decision support tool, the extraction of information from unstructured data, or automatic phenotyping methods.

2.3 Observations from a rapidly evolving and heterogeneous ecosystem

Fig 2.2 summarizes the development state of progress of CDWs in France. Out of 32 regional and university hospitals in France, 14 have a CDW in production, 5 are experimenting, 5 have a prospective CDW project, 8 did not have any CDW project at the time of writing. The results are described for all projects that are at least in the prospective stage minus the three that we were unable to interview after multiple reminders (Orléans, Metz and Caen), resulting in a denominator of 21 university hospitals.

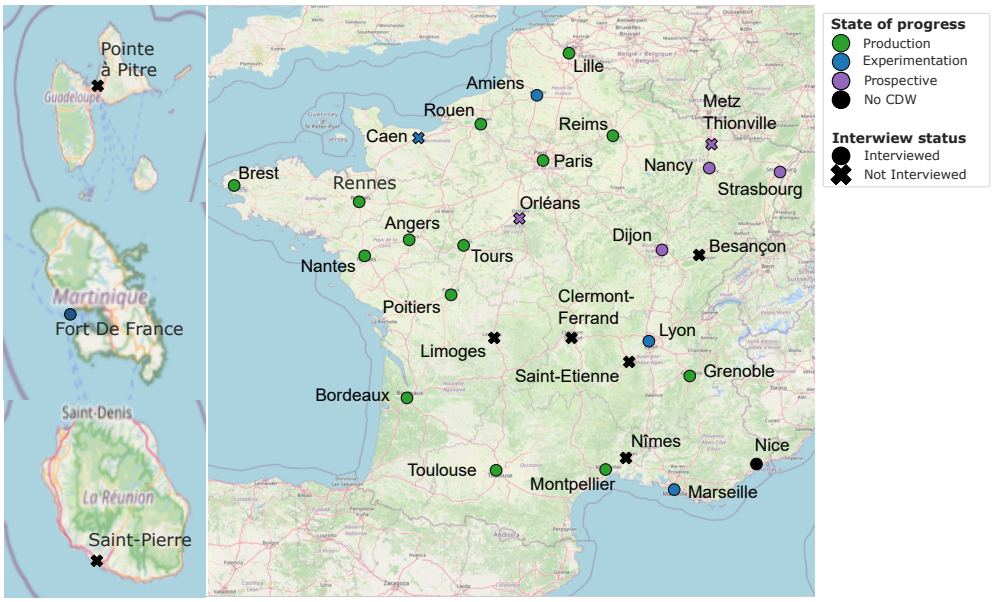


Fig. 2.2. Repartition of CDWs in France.
The base layer map comes from `https://wiki.openstreetmap.org/wiki/Mapnik`.

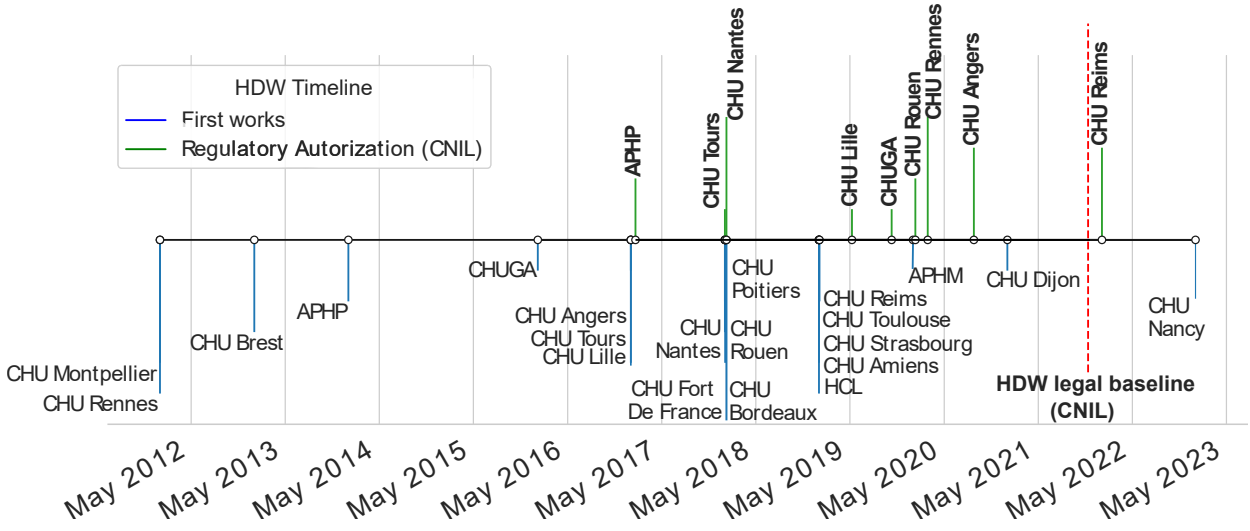


Fig. 2.3. The French CDWs implementations date back to the first academic works in data reuse in early 2010s and accelerated recently.

2.3.1 Governance: CDWs are federating multiple teams in the hospital

Initiation and actors Fig 2.3 shows the history of the implementation of CDWs. A distinction must be made between the first works –in blue–, which systematically precede the regulatory authorization –in green– from the French Commission on Information Technology and Liberties (CNIL).

The CDWs have so far been initiated by one or two people from the hospital world with an academic background in bioinformatics, medical informatics or statistics. The sustainability of the CDW is accompanied by the construction of a cooperative environment between different actors: Medical Information Department (MID), Information Systems Department (IT), Clinical Research Department (CRD), clinical users, and the support of the management or the Institutional Medical Committee. It is also accompanied by the creation of a team, or entity, dedicated to the maintenance and implementation of the CDW. More recent initiatives, such as those of the HCL (Hospitals of the city of Lyon) or the *Grand-Est* region, are distinguished by an initial, institutional and high-level support.

The CDW has a federating potential for the different business departments of the hospital with the active participation of the CRD, the IT Department and the MID. Although there is always an operational CDW team, the human resources allocated to it vary greatly: from half a full-time equivalent to 80 people for the AP-HP, with a median of 6.0 people. The team systematically includes a coordinating physician. It is multidisciplinary with skills in public health, medical informatics, informatics (web service, database, network, infrastructure), data engineering and statistics.

Historically, the first CDWs were based on in-house solution development. More recently, private actors are offering their services for the implementation and implementation of CDWs (15/21). These services range from technical expertise in order to build up the data flows and data cleaning up to the delivery of a platform integrating the different stages of data processing.

2.3.2 Management of studies

Before starting, projects are systematically analyzed by a scientific and ethical committee. A local submission and follow-up platform is often mentioned (12/21), but its functional

scope is not well defined. It ranges from simple authorization of the project to the automatic provision of data into a Trusted Research Environment (TRE) (Goldacre et al., 2022). The processes for starting a new project on the CDW are always communicated internally but rarely documented publicly (8/21).

2.3.3 Uneven transparency of ongoing studies

Ongoing studies in CDWs are unevenly referenced publicly on hospital websites. Some institutions have comprehensive study portals, while others list only a dozen studies on their public site while mentioning several hundreds ongoing projects during interviews. In total, we found 8 of these portals out of 14 CDWs in production. Uses other than ongoing scientific studies are very rarely documented. The publication of the list of ongoing studies is very heterogeneous and fragmented between several sources: `clinicaltrials.gov`, the mandatory project portal of the Health Data Hub (HDH, 2023b) or the website of the hospital data warehouse. [1]

2.3.4 Triple usage of data: Research, management, clinic

Strong dependance to the Hospital Information System CDW data reflect the HIS used on a daily basis by hospital staff. Stakeholders point out that the quality of CDW data and the amount of work required for rapid and efficient reuse are highly dependent on the source HIS. The possibility of accessing data from an HIS in a structured and standardized format greatly simplifies its integration into the CDW and then its reuse.

Categories of Data Although the software landscape is varied across the country, the main functionalities of HIS are the same. We can therefore conduct an analysis of the content of the CDWs, according to the main categories of common data present in the HIS.

The common base for all CDWs is constituted by data from the Patient Administrative Management software (patient identification, hospital movements) and the billing codes. Then, data flows are progressively developed from the various softwares that make up the HIS. The goal is to build a homogeneous data schema, linking the sources together, controlled by the CDW team. The prioritization of sources is done through thematic projects, which feed the CDW construction process. These projects improve the understanding of the sources involved, by confronting the CDW team with the quality issues present in the data.

Table 2.1 presents the different ratio of data categories integrated in French CDWs. Structured biology and texts are almost always integrated (20/21 and 20/21). The texts contain a large amount of information. They constitute unstructured data and are therefore more difficult to use than structured tables. Other integrated sources are the hospital drug circuit (prescriptions and administration, 16/21), Intensive Care Unit (ICU, 2/21) or nurse forms (4/21). Imaging is rarely integrated (4/21), notably for reasons of volume. Genomic data are well identified, but never integrated, even though they are sometimes considered important and included in the CDW work program.

[1] The full collection of ongoing study is available on this url: `https://gitlab.has-sante.fr/has-sante/public/rapport_edsh/-/blob/master/data/cycle_eds/cycle_eds_etudes.csv?ref_type=heads`

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 %

Table 2.1. Type of data integrated into the French CDWs.

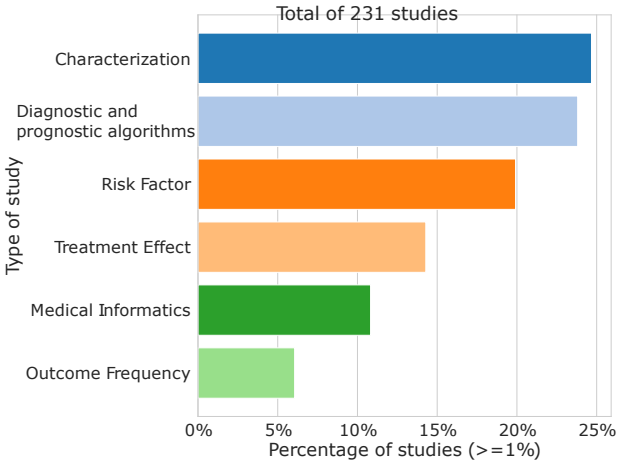


Fig. 2.4. Percentage of studies by objective.

Today, CDWs are built predominantly for scientific research (21/21) The studies are mainly observational (non-interventional). Fig 2.4 presents the distribution of the six categories defined in 2.2.2 for 231 studies collected on the study portals of nine hospitals. The studies focus first on population characterization (25 %), followed by the development of diagnostic and prognostic algorithms (24 %), the study of risk factors (19 %) and the treatment effect evaluations (15 %).

The CDWs are used extensively for internal projects such as student theses (at least in 9/21) and serve as an infrastructure for single-service research: their great interest being the de-siloing of different information systems. For most of the institutions interviewed, there is still a lack of resources and maturity of methods and tools for conducting inter-institutional research (such as in the *Grand-Ouest* region of France) or via European calls for projects (EHDEN). These two research networks are made possible by supra-local governance and a common data schema, respectively eHop (Madec et al., 2019) and OMOP (Hripcsak et al., 2015b). The Paris hospitals, thanks to their regional coverage and the choice of OMOP, are also well advanced in multi-centric research. At the same time, the *Grand-Est* region is building a network of CDW based on the model of the *Grand-Ouest* region, also using eHop.

Data reuse – CDW are used for monitoring and management (16/21) The CDW have sometimes been initiated to improve and optimize billing coding (4/21). The clinical texts gathered in the same database are queried using keywords to facilitate the structuring of information. The data are then aggregated into indicators, some of which are reported at the national level. These types of indicators also inform the administrative management of the institution. Finally, closer to the clinic, some actors state that the CDW could also be used to provide regular and appropriate feedback to healthcare professionals on their practices. This feedback would help to increase the involvement and interest of healthcare professionals in CDW projects. The CDW is sometimes of interest for health monitoring (e.g., during Covid-19) or pharmacovigilance (13/21).

Data reuse – Strong interest for CDW in the context of care (13/21) Some CDWs develop specific applications that provide new functionalities compared to care software. Search engines can be used to query all the hospital's data gathered in the CDW, without data compartmentalization between different softwares. Dedicated interfaces can then offer a unified view of the history of a patient's data, with inter-specialty transversality, which is

particularly valuable in internal medicine. These cross-disciplinary search tools also enable healthcare professionals to conduct rapid searches in all the texts, for example to find similar patients (Garcelon et al., 2017). Uses for prevention, automation of repetitive tasks and care coordination are also highlighted. Concrete examples are the automatic sorting of hospital prescriptions by order of complexity, or the setting up of specialized channels for primary or secondary prevention.

2.3.5 A multi-layered technical architecture

The technical architecture of modern CDWs has several layers:

- Data processing: connection and export of source data, diverse transformation (cleaning, aggregation, filtering, standardization).
- Data storage: database engines, file storage (on file servers or object storage), indexing engines to optimize certain queries.
- Data exposure: raw data, APIs, dashboards, development and analysis environments, specific web applications.

Supplementary cross-functional components ensure the efficient and secure operation of the platform: identity and authorization management, activity logging, automated administration of servers and applications.

The analysis environment (Jupyterhub or RStudio datalabs) is a key component of the platform, as it allows data to be processed within the CDW infrastructure. A few CDWs had such operational datalab at the time of our study (6/21) and almost all of them have decided to provide it to researchers. Currently, clinical research teams are still often working on data extractions, in less secure environments.

2.3.6 Rare data quality checks and multiple standard formats

Quality tools Systematic data quality monitoring processes are being built in some CDWs. Often (8/21), scripts are run at regular intervals to detect technical anomalies in data flows. Rare data quality investigation tools, in the form of dashboards, are beginning to be developed internally (3/21). Theoretical reflections are underway on the possibility of automating data consistency checks, for example, demographic or temporal. Some facilities randomly pull records from the EHR to compare them with the information in the CDW.

Standard format No single standard data model stands out as being used by all CDWs. All are aware of the existence of the OMOP (research standard) (Hripcsak et al., 2015b) and HL7 FHIR (communication standard) models (Braunstein, 2019). Several CDWs consider the OMOP model to be a central part of the warehouse, particularly for research purposes (9/21). This tendency has been encouraged by the European call for projects EHDEN, launched by the OHDSI research consortium, the originator of this data model. In the *Grand-Ouest* region of France, the CDWs use the eHop warehouse software. The latter uses a common data model also named eHop. This model will be extended with the future warehouse network of the *Grand Est* region also choosing this solution. Including this grouping and the other establishments that have chosen eHop, this model includes 12 establishments out of the 32 university hospitals. This allows eHop adopters to launch ambitious interregional projects. However, eHop does not define a standard nomenclature to be used in its model and is not aligned with emerging international standards.

Documentation Half of the CDWs have put in place documentation accessible within the organization on data flows, the meaning and proper use of qualified data (10/21 mentioned). This documentation is used by the team that develops and maintains the warehouse. It is also used by users to understand the transformations performed on the data. However, it is never publicly available. No schema of the data once it has been transformed and prepared for analysis is published.

2.4 Recommendations: How to consolidate EHRs and expand usages

We give the first overview of the CDWs in university hospitals of France with 32 hospitals reviewed. The implementation of CDW dates from 2011 and accelerated in the late 2020. Today, 24 of the university hospitals have an ongoing CDW project. From this case study, some general considerations can be drawn, that should be valuable to all healthcare system implementing CDWs on a national scale.

2.4.1 Governance: CDWs are infrastructures

Multidisciplinary teams As the CDW becomes an essential component of data management in the hospital, the creation of an autonomous internal team dedicated to data architecture, process automation and data documentation should be encouraged (Goldacre et al., 2022). This multidisciplinary team should develop an excellent knowledge of the data collection process and potential reuses in order to qualify the different flows coming from the source IS, standardize them towards a homogenous schema and harmonize the semantics. It should have a sound knowledge of public health, as well as the technical and statistical skills to develop high-quality software that facilitates data reuse.

Lack of sustainable funding The resources specific to the warehouse are rare and often taken from other budgets, or from project-based credits. While this is natural for an initial prototyping phase, it does not seem adapted to the perennial and transversal nature of the tool. As a research infrastructure of growing importance, it must have the financial and organizational means to plan for the long term.

The governance of the CDW has three layers – Within the university hospital, interregional, and national/international The first level allow to ensure the quality of data integration as well as the pertinence of data reuse by clinicians themselves. The interregional level is well adapted for resources mutualization and collaboration. Finally, the national and international levels assure coordination, encourage consensus for committing choices such as metadata or interoperability, and provide financial, technical and regulatory support.

2.4.2 Transparency: Keep the bar high

For better registration of observational studies Health Technology Assessment agencies advocate for public registration of comparative observational study protocols before conducting the analysis (Berger et al., 2017; FDA, 2021a; HAS, 2021). They often refer to `clinicaltrials.gov` as potential but not ideal registration portal for observational

studies. The research community advocates for public registrations of all observational studies (Rushton, 2011; PLOS Medicine Editors, 2014). More recently, it emphasizes the need for more easy data access and the publication of study code (Pavlenko et al., 2020; Kohane et al., 2021; NIH, 2023). We embrace these recommendations and we point to the unfortunate duplication of these study reporting systems in France. One source could be favored at the national level and the second one automatically fed from the reference source, by agreeing on common metadata.

The patient's perspective There is currently no way to know if a specific patient personal data is included for a specific project. Better patient information about the reuse of their data is needed to build trust over the long term. A strict minimum is the establishment and update of the declarative portals of ongoing studies at each institution.

2.4.3 New data, new challenges

Shift the focus to data engineering When using CDW, the analyst has not defined the data collection process and is generally unaware of the context in which the information is logged. This new dimension of medical research requires a much greater development of data science skills to change the focus from the implementation of the statistical design to the data engineering process. Data reuse requires more effort to prepare the data and document the transformations performed.

Poor adoption of common standards The more heterogeneous a HIS system is, the poorest quality will have a CDW built above it. There is a need for increasing interoperability, to help EHR vendors interfacing the different hospital softwares, thus facilitating CDW development. One step in this direction would be the open source publication of HIR data schema and vocabularies. At the analysis level, international recommendations insist on the need for common data formats (Zhang et al., 2022; Kohane et al., 2021). However, there is still a lack of adoption of research standards from hospital CDWs to conduct robust studies across multiple sites. Building open-source tools on top of these standards such as those of OHDSI (Schuemie, 2021) could foster their adoption. Finally, in many clinical domains, sufficient sample size is hard to obtain without international data sharing collaborations. Thus, more incitation is needed to maintain and update the terminology mappings between local nomenclatures and international standards.

Lack of translational researches Many ongoing studies concern the development of diagnostic and prognostic algorithms whose goal is to save time for healthcare professionals. These are often research projects, not yet integrated into routine care. The analysis of study portals and the interviews revealed that data reuse oriented towards primary care is still rare and rarely supported by appropriate funding. The translation from research to clinical practice takes time and need to be supported on the long run to yield substantial results.

2.4.4 Technical architecture: Towards more harmonization and open source ?

Tools, methods and data formats of CDW lack harmonization due the presence of many actors. The strong technical innovation in the field led to the emergence of many heterogeneous solutions. As suggested by the recent report on the use of data for research in the UK

(Goldacre et al., 2022), it would be wise to focus on a small number of model technical platforms.

These platforms should favor open-source solutions to assure transparency by default, foster collaboration and consensus and avoid technological lock-in of the hospitals.

2.4.5 Data quality and documentation: more incentives needed

Focus on data quality Quality is not sufficiently considered as a relevant scientific topic itself. However, it is the backbone of all research done within an CDW. In order to improve the quality of the data with respect to research uses, it is necessary to conduct continuous studies dedicated to this topic (Zhang et al., 2022; Kohane et al., 2021; Shang et al., 2018; Looten et al., 2019). These studies should contribute to a reflection on methodologies and standard tools for data quality, such as those developed by the OHDSI research network (Schuemie, 2021).

Open-source is key to improve quality Finally, there is a need for open-source publication of research code to ensure quality retrospective research (Shang et al., 2018; Seastedt et al., 2022). Recent research in data analysis has shown that innumerable biases can lurk in training data sets (Gebru et al., 2021; Mehrabi et al., 2021). Open publication of data schemas is considered an indispensable prerequisite for all data science and artificial intelligence uses (Gebru et al., 2021). Inspired by dataset cards (Gebru et al., 2021) and dataset publication guides, it would be interesting to define a standard CDW card documenting the main data flows.

2.5 Conclusion

Limitations The interviews were conducted in a semi-structured manner within a limited time frame. As a result, some topics were covered more quickly and only those explicitly mentioned by the participants could be recorded. The uneven existence of study portals introduces a bias in the recording of the types of studies conducted on CDW. Those with a transparency portal already have more maturity in use cases.

For clarity, our results are focused on the perimeter of university hospitals. We have not covered the exhaustive health care landscape in France. CDW initiatives also exist in primary care, in smaller hospital groups and in private companies.

The French CDW ecosystem is beginning to take shape It benefits from an acceleration thanks to national funding, the multiplication of industrial players specializing in health data and the beginning of a supra-national reflection on the European Health Data Space (EC, 2022). However, some points require special attention to ensure that the potential of the CDW translates into patient benefits.

The priority is the creation and perpetuation of multidisciplinary warehouse teams This team should be capable of operating the CDW and supporting the various projects. A combination of public health, data engineering, data stewardship, statistics and IT competences is a prerequisite for the success of the CDW. The team should be the privileged point of contact for data exploitation issues and should collaborate closely with the existing hospital departments.

The constitution of a multi-level collaboration network is another priority The local level is essential to structure the data and understand its possible uses. Interregional, national and international coordination would make it possible to create thematic working groups, in order to stimulate a dynamic of cooperation and mutualization.

A common data model should be encouraged It should specify metadata allowing to map the integrated data, in order to qualify the uses to be developed today from the CDWs. More broadly, open-source documentation of data flows and transformations performed for quality enhancement would require more incentives to unleash the potential for innovation for all health data users.

Expanding the scope of the data beyond the purely hospital domain Many risk factors and patient follow-up data are missing from the CDWs, but are crucial for understanding pathologies. Combining city data and hospital data would provide a complete view of patient care.

Chapter 3

Exploring a complexity gradient in representation and predictive models for EHRs

> *What must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations?*
>
> *– Vladimir Vapnik, The nature of Statistical Learning Theory, 2000*

Chapter's content

As chapter 2 shows, Electronic Health Records (EHRs) contain multiple categories of data sparking interest in the development of predictive algorithms. Text and image put aside, this data can be represented as time-stamped medical codes with a high number of categories and biological measurements. Current state-of-the-art predictive models for EHRs build on increasingly elaborated pipelines –for instance using the transformer architecture– to handle the complexity of these data. Given the operational difficulties to transfer and adapt these models on local care environments, we explore a complexity-benefit tradeoff by comparing them to simple aggregation of events. We use three predictive tasks involving time-varying structured EHRs and increasingly clinically relevant problems. We introduce a simple aggregation of static embeddings –transferred from national claims and publicly available–, showing that it outperforms transformer-based models on simple tasks with medium sample sizes. We highlight the sample and computing resource efficiency of these models. Finally, clinically relevant problems generally present a strong class imbalance, with low outcome prevalence. This makes frugal models particularly attractive because of their capacity to learn from few examples. Despite being attractive for large sample sizes –over the million, complex models, as with transformers may be less adpated to typical clinical settings than lightweight data-processing pipelines using tree-based models.

This chapter presents ongoing work. A communication was accepted for the Simpa2023 day on patient similarity.

work done with Judith Abecassis, Julie Adjerad, Theo Jolivet, Jean-Baptiste Julla and Gaël Varoquaux.

Outline

3.1 The modern quest for medical oracles

The increasing availability of routine care data and advances in machine learning are raising hope for predictive modeling in healthcare. EHRs are promising because of the richness of their data and their integration in routine care (Raghupathi; Raghupathi, 2014). However, predictive algorithms trying to model this ever richer data sources gain in complexity. There is currently a lack of of guidance on the performance-complexity tradeoff in predictive models (Hond et al., 2022). We explore a complexity gradient of predictive pipelines on three tasks ranging from general populations of patients to more specific clinical cases.

3.1.1 Focus on predictive models for planning or risk scores

In biostatistics and clinical medicine, prognosis models have been motivated by *risk stratification*: the Framingham risk score for coronary heart disease was an early attempt to characterize behavior changes that could lead to decreased risk (Brand et al., 1976). Clinical *risk scores* are either focused on the short term with alarm models (Tang et al., 2007; Rothman et al., 2013; Wong et al., 2021) or on the long term with screening models. In parallel, artificial intelligence in medicine looks for prognosis model as part of larger *decision-making* systems (Szolovits, 1982). Finally, in complex healthcare organizations, accurate individual predictions help to use efficiently constrained medical resource by informing *care planning* (Topol, 2019). Appendix C.2.1 details these four type of predictive objectives for healthcare.

Following the trend for evaluating current state-of-the-art models (Wornow et al., 2023), we focus on purely predictive models: risk scores or planning.

3.1.2 Predictive pipelines fueling medical predictions are increasingly complex

Original complexity of the data EHRs contain time varying variables with a high cardinality. Time-consuming work from medical and computer experts is required to clean and transform these sources into data tables suited for statistical analysis (Bacry et al., 2020; Hripcsak et al., 2015a). More specifically, data preprocessing usually requires mapping non standard terminologies, create derived variables by aggregating measures over specific time periods. The choices of baseline covariates should be driven by medical experts which are often not familiar with the complexity of the full EHR data processing pipeline. We propose to leverage models that take raw structured data as input to avoid as much as possible these preprocessing steps.

From simple hand-crafted linear models to elaborated weakly-supervised tranformers Predictions on EHRs originally used linear models on few carefully selected static variables for tens of thousand patients coming from a single center (Goldstein et al., 2017). To overcome the high cardinality of EHRs medical vocabularies, a series of deep learning methods have been developed (Shickel et al., 2017). Time varying features were first modelled thanks to recurrent neural network (Lipton et al., 2016), then with large transformer-based architectures (Li et al., 2020b). Appendix C.2.2 further details this evolution of predictive model complexity.

3.1.3 The illusion of large populations

Well-defined clinical questions concern small numbers of cases Diseases are rare, often with less than 5% of prevalence (CNAM, 2023), even lower if considering incidence.

For example, consider the flowchart for the prediction of cardiovascular complications –a rather common clinical condition–, we first extracted a cohort of 2,101,819 patients. After running all inclusion criteria –shown in Figure C.3, we are left with only 4,360 cases among a population at risk of 165,930 patients. Currently, large models are trained on larger number of cases (see a review of number of cases for three major transformer-based models in Appendix C.3). Privacy rights enforced by the General Data Protection Regulation (GDPR) make it difficult to access large repositories of healthcare data. We should thus develop and evaluate prognosis models on small samples, accessible in many hospitals.

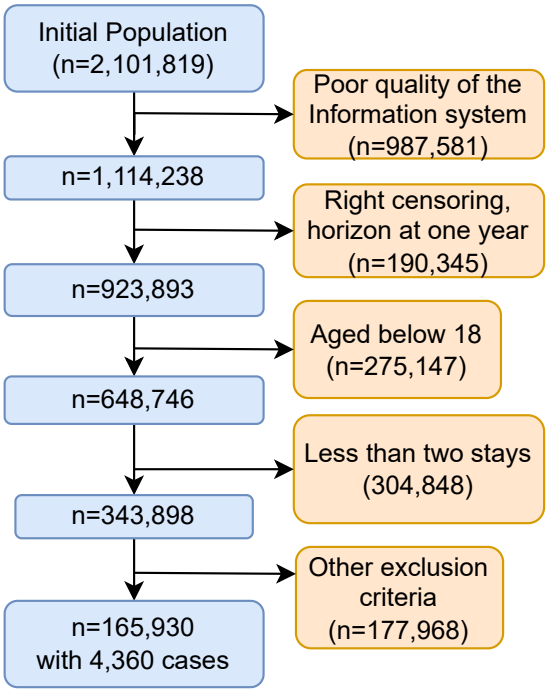


Fig. 3.1. Simplified selection flowchart for major cardiovascular events prediction. Most of the exclusion occurs because of poor quality of the information system, pediatric patients, right censoring or patients having less than two stays in the database. Appendix C.4c provides the complete flowchart.

Transferring predictive model is no silver bullet A potential solution to local small sample sizes is to transfer models pretrained on large populations. However, privacy requirements also prevent such transfer between institutions [1]. Even if publicly available, major dataset shifts might break effective transfer of predictive pipelines (Finlayson et al., 2021): heterogeneity in coding practices (Juven, 2013), socio-economic status (Gianfrancesco et al., 2018), inconsistency of practices in different organizations (Agniel et al., 2018). Futoma et al., 2020 even argues that it is not possible to generalize machine learning models in healthcare: Different population, information systems and healthcare practices may require tailored tools (Rose, 2018). As an example of failed generalization, Wong et al., 2021

[1] We asked for CEHR-BERT to be published, but so far without success: https://github.com/cumc-dbmi/cehr-bert/issues/2

externally evaluated a sepsis prediction model deployed in hundered of US caresites. The model transfer performance dropped from 0.76 to 0.83 area under the receiver operating curve (ROC AUC) declared by the manufacturer to 0.63.

3.1.4 Current barriers to predictive models usefulness

Credibility of external validity requires simple model deployments Some predictive risks are used in daily clinical practice: For example, the Glasgow Coma Scale (Teasdale; Jennett, 1974) or the APACHE III score (Knaus et al., 1991), though these are very simple scores built from expertise rather than machine learning. But only a small part of the published models are successfully deployed into clinical practice (Wyatt; Altman, 1995; Kelly et al., 2019). Poor adoptions are due to lack of evidence for clinical credibility, accuracy, generality or clinical effectiveness. Improving the strength of this evidence requires repeatedly testing these systems on different prospective populations with intuitive metrics for physicians (Kelly et al., 2019; Varoquaux; Colliot, 2022; Wornow et al., 2023). Hence deployment simplicity is a key factor for model adoption.

Simple pipelines facilitate model deployments Hospital softwares already run on complex information systems. We should aim for predictive models that can run on pre-existing hardware such as the successful QRISK for cardiovascular risk prediction (Hippisley-Cox et al., 2017). Relying on such commodity hardware facilitates model training, calibration and deployment, eventually leading to better clinical adoption and integration into healthcare processes. To benefit the largest number of patients accross the world, predictive models should be an appropriate technology: it should be *easily and economically utilized from readily available resources by local communities to meet their needs* (Pearce, 2012). However, recent predictive pipelines strongly focus on performance gains, at the cost of increasing architecture sizes requiring ever greater computing and technical resources –as detailed in Appendix C.4. There is a need to balance good predictive performance with model frugality.

3.1.5 Objective and outline of the paper

Contributions We study the performance of increasingly complex models on three operational and clinical tasks. We explore the performance trade-offs between small models that can be run on few samples and large models requiring more samples to become effective. We propose a new model based on the transfer of static medical embeddings trained from large claims databases. In the technically and legally constrained environment of healthcare data, what predictive pipelines are the most efficient to yield good practical utility?

Outline Section 3.2 details four predictive pipelines of increasing complexity. Subsection 3.3.1 defines three clinical tasks covering different benchmarking and clinical usefulness. Subsection 3.3.1 details our evaluation pipeline. Finally subsection 3.3.2 presents our results and section 3.4 discusses their implications.

Main findings With small to moderate sample sizes -up to 20,000 patients-, simple models have better performance than transformer-based models on general tasks such as length of stay interpolation or prognosis. Among the simpler model based on row counts or static embeddings, no featurization choice largely outperforms others in any single task but random forest estimators are always more performant than linear estimators. Static embeddings

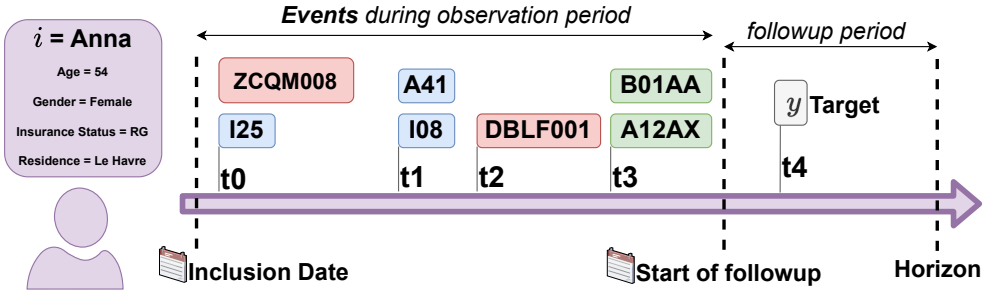


Fig. 3.2. Event time representation of EHRs data for predictive tasks. Events are observed between the inclusion and start of followup dates. Target is considered only between the start of followup and the horizon to avoid right censoring. The event types are distinguished by their color: billing diagnoses in blue, billing procedures in red and drug administrations in green.

models are the most compute and memory efficient solution. We quantify the detrimental effects on prognosis performance of low target prevalences –linked to the number of cases.

3.2 From basic to complex: four increasingly sophisticated predictive pipelines

3.2.1 A simple information-preserving data format: sequence of events

We model the patient healthcare trajectory as a sequence of events (Beam et al., 2019; Bacry et al., 2020; Chazard et al., 2022). Each event is described by a triplet: a person identifier i, a datetime t and a medical code c, $e = (i, t, c)$ as shown in Figure 3.2. For each patient i, the complete trajectory is the ordered collection of its T_i events $S_i = \{e_k\}_{k=1..T_i}$. S is the collection of sequences for all N patients: $S = \{S_i\}_{i=1..N}$.This simple format integrates together a wide variety of time varying features without imposing data transformation or feature selection –often brittle to modelization choices. However, it calls for adequate aggregation methods to reduce the long event table into a patient-wise features table of shape $N \times d$ matrix for d that can be passed over to statistical estimators. Here, we explore four different choices of aggregation of increasing complexity, followed by common statistical estimators. We call these aggregation methods *featurizers* and note them $g(\cdot, \lambda_g)$, where λ_g are the parameters of the featurizers.

3.2.2 Demographic features: g_{demo}

One simple featurization choice ignores these sequence of events and only focus on a few demographic features available in the patient record: age, gender, admission origin, discharge destination, admission date. These demographics are added as new columns for all other subsequent featurizers.

3.2.3 Decayed counting of event features: g_{count}

We process the raw events by computing a sparse count matrix C of shape $(N, n_{vocabulary})$ where each row collapses the patient history by counting the number of times a concept is present in the patient history as illustrated in Figure 3.3. To better take into account the temporal dimension, we also computed a decayed count of the events in the patient history,

noted C_δ.

For an event triplet (i, t, c), we compute the time delta between the event and the followup time.

$$\Delta t = |t - T_0|$$

Then, we decay the count matrix with an exponential of half life δ.

$$C_\delta[i, c] += e^{-\frac{\Delta t}{\delta}}$$

Finally, we obtain the patient features by concatenating each delayed count matrix; e.g., with decay 0 and decay δ.

$$g_{count}(S, (0, \delta)) = [C_0, C_\delta]$$

Several decays can be selected and concatenated together. The decays are considered hyperparameters which are selected by cross-validation.

3.2.4 Static embeddings of event features: $g_{emb-local}$ or $g_{\text{emb-SNDS}}$

The SVD-PPMI algorithm (introduced by Beam; Kohane, 2018, detailed in Appendix C.5.3) performs a dimension reduction on the cooccurrence matrix between medical concepts. It creates neighboring vector representations for codes that coocur frequently together and thus induces some sense of clinical relation. Building upon this algorithm, we consider two sequence representation techniques. These embeddings are static by opposition to transformer-based embeddings that adapts to the context.

Static embeddings locally trained: $g_{emb-local}$ We apply the SVD-PPMI algorithm to the training cohort yielding static embeddings Φ_{local}. Aggregation at the stay level is done using the same count matrices as for 3.2.3 –potentially with decayed counts: $g_{emb-local}(S, (0, \delta)) = [C_0 \cdot \Phi_{local}, C_\delta \cdot \Phi_{local}]$.

Transfer trained static embeddings Instead of retraining the embeddings on the train cohort, we rely on static embeddings Φ_{SNDS} pre-computed on French national claims data from 3,112,565 patients (Doutreligne et al., 2021). The aggregation is performed similarly as for decayed counting and local embeddings: $g_{\text{emb-SNDS}}((S, (0, \delta))) = [C_0 \cdot \Phi_{SNDS}, C_\delta \cdot \Phi_{SNDS}]$.

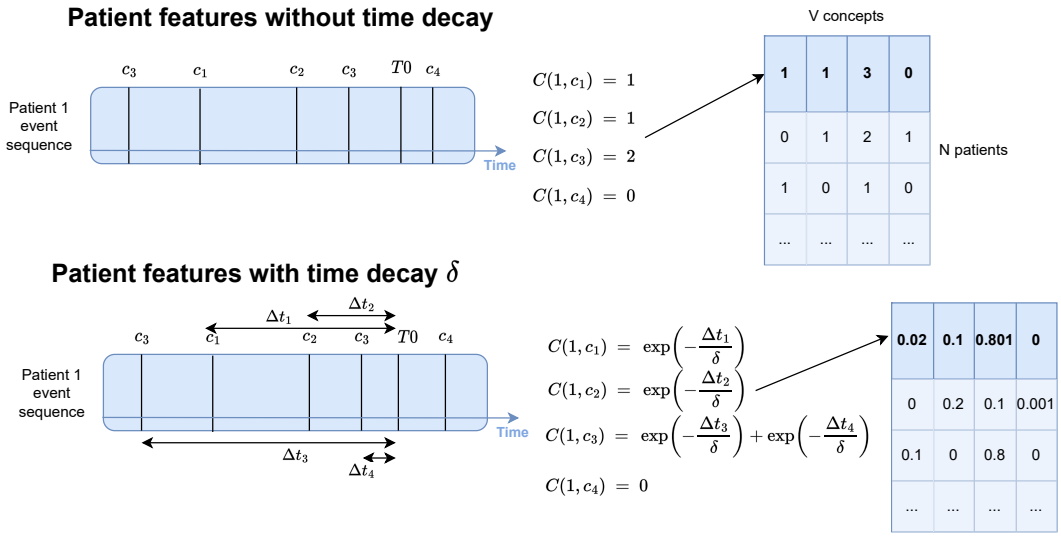


Fig. 3.3. Illustration of the decayed counting procedure.

3.2.5 Transformer based: g_{cbert}

Relying on the transformer architecture (Vaswani et al., 2017), recent trajectory modeling algorithms showed promising results for various EHR prediction tasks (Li et al., 2020b; Rasmy et al., 2021; Pang et al., 2021). We benchmarked CEHR-BERT (Pang et al., 2021), one of these recent transformer-based models particularly adapted to our data format.

Transformer models are trained in two steps: a) First, pre-train the model on an auxiliary task using a large database. In our case, this task is a Masked Language Model (MLM): the network tries to predict the medical concept of randomly masked events in the sequences. The CEHR-BERT implementation also tries to predict the type of the next visit. b) Then fine-tune the model on the task of interest. We use the train set for pretraining and finetuning. Details on this architecture are given in Appendix C.5.4.

3.2.6 Final step estimator

These choices of featurizers are not task-specific. To obtain probabilities of occurrence for the target, the analyst needs to choose and train an estimator f and its corresponding parameters λ_f. The predictions are given by the chain of the featurizer and the estimator: $\hat{y}(s_i) = f(g(s_i, \lambda_g), \lambda_f)$.

For CEHR-BERT, the predictor is the final layer of the neural network. It is trained together with the sequence representations during the fine-tuning step. For the other featurizers, we create a scikit-learn pipeline [2] where a featurizer is followed by an estimator: either a penalized logistic regression or a random forest. The hyperparameters of the featurizers and estimators can be cross-validated together. We experimented with gradient boosting trees, without observing significant performance gains.

3.3 Empirical Study – Benchmarking three operational and clinical tasks

3.3.1 Experiments to explore the performance-complexity trade off

Heathcare database – Greater Paris Hospitals We use data extracted from the data warehouse of the Greater Paris Hospitals (AP-HP), hosting routine care data from 38 hospitals in the Paris area. Details on this database is given in Appendix C.6.1. We include every medical event among drug administrations, ICD10 diagnosis and procedure billing codes.

Prediction tasks framing Here, we detail our framework to precisely define each predictive task, taking inspiration from OHDSI, 2021; Tomašev et al., 2021. First, we define a *study period* during which data acquisition was sufficiently stable. Then, we define and select the *cohort* detailed for each task. Inclusion criteria are detailed in the flowcharts of Figure C.4. We define the task with: an *index visit*, an *observation period* during which events are fed to the predictive model, an *horizon* after the end of the observation period defining the followup, and a target event potentially occurring during the followup.

We study three tasks of increasing clinical value and implementation complexity: Long Length Of Stay interpolation (LOS), prognosis the prediction of the next stay ICD10 chapters,

[2] https://scikit-learn.org/stable/modules/compose.html#combining-estimators

and prediction at one year of incident Major Adverse Cardiovascular Events (MACE). Table 3.1 summarizes key characteristics of the cohorts and task definitions and Appendix C.6.2 details each task.

	Long LOS	**Prognosis**	**MACE**
Task	Binary classification	Multi-Label binary classification	Binary classification
Index Visit	First inpatient visit	Random Non-Final Visit	Random Visit
Observation Period	Index visit	Full trajectory before end of index visit	Full Trajectory before end of index visit
Horizon	End of index visit	End of next visit	12 Months
Median Age	56.4	61.5	60.0
Female	54.6%	53.3%	53.7%
Cohort Size	27,053	10,786	165,948
Prevalence	23.1%	From 1.3 to 55.9%	2.6 %
Number of cases	6,249	From 139 to 6,029	4,315
Description	Long stay classification (longer than 7 days)	Next stay prognosis: ICD10 chapter classification	MACE prognosis at one year

Table 3.1. Tasks definitions and cohort characteristics.

Evaluation procedure – Exploring sample efficiency We split each cohort described in Table 3.1 into a train and a test set following a temporal split based on each patient inclusion date with a 0.8/0.2 ratio (split detailed in Appendix C.6.3). Each patient only appears in one of the two sets. To study the sample efficiency of the different pipelines, we further restrict the effective train set size to increasing ratio of its full size. Figure 3.4 summarizes the procedure. We report Area Under the Receiver Operating Characteristic Curve (ROC AUC) or Area Under the Precision Recall Curve results (AUPRC) which is more adapted to highly imbalanced tasks.

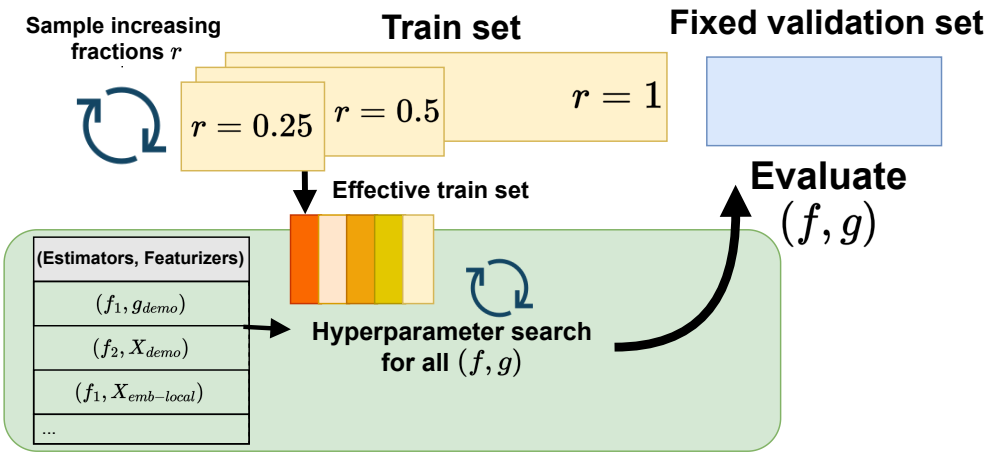


Fig. 3.4. Evaluation procedure: For each effective train set of size r, each pair of featurizer and estimator are cross-validated together to obtain the best parameters, then evaluated on a fixed validation set.

Appendix C.8 details an alternative geographic split validation procedure for LOS and prognosis tasks, exploring the validity of our results when testing the models on patient from other hospitals.

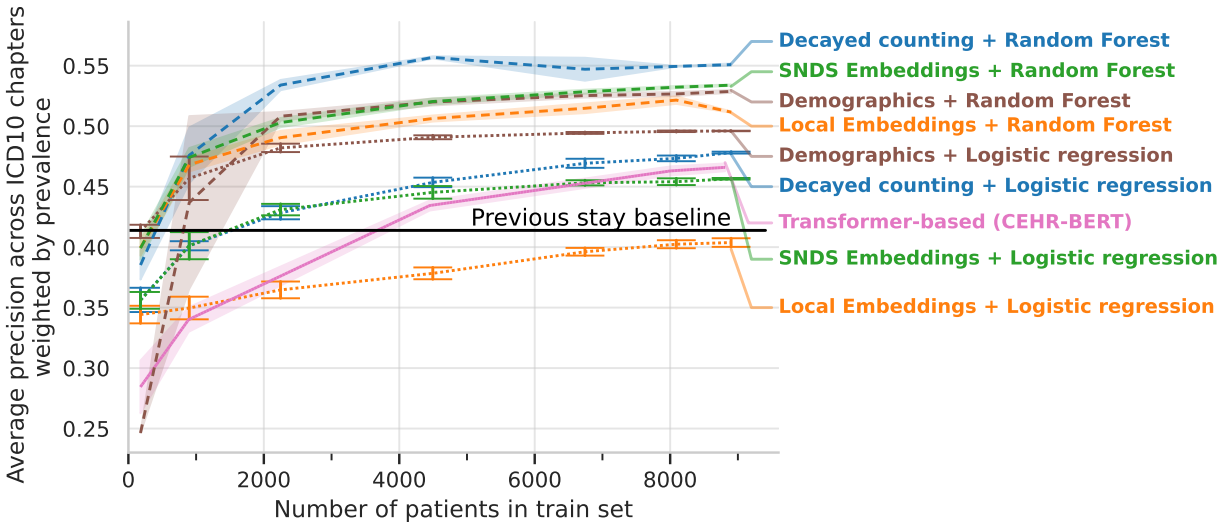


Fig. 3.5. Prognosis AUPRC, weighted by prevalences over 21 ICD10 chapters with more than 1% prevalence. The performance is averaged over 5 folds. The shaded area represents the standard deviation. The horizontal black line displays the naive baseline that predicts the previous stay codes for the target stay. Random forest have better performance. Count encoder outperforms other featurizers, suggesting the importance of low count events that are smoothed out in embedding methods. We report AUPRC rather than ROC AUC, since it takes better into account the difference in prevalence between chapters. Appendix C.7.2 details AUPRC for each chapter and displays averaged ROC AUC.

3.3.2 Results – Tree-based models on event counts, a simple but efficient performer

Decayed counting followed by random forest outperforms elaborate embedding models Both for the prognosis (shown if Figure 3.5) and the LOS task (shown in Figure 3.6), decayed counting with random forest outperforms other pipelines. For these two tasks, the transformer model is far from the best performance, certainly because of too small samples.

For the prognosis task, we added to all featurizers the previous index stay ICD10 chapters as supplementary features. This intuitive baseline is reported as the horizontal black line. The good performance of the logistic regression with demographic features only indicates that a simple linear combination of the index stay chapters is a strong baseline.

The challenge of low prevalence Figure 3.7 shows the AUPRC performances on the prognosis task plotted against the prevalence for a random forest estimator. Low prevalences translate into small number of cases, which decreases overall performance and increase variances among chapters with less than 5% prevalences. Appendix C.7.2 shows similar results for ROC AUC and linear estimators.

Static embeddings reduce computational costs Figure 3.8 shows the training time of the different pipelines averaged over the 21 chapters of the prognosis task, highlighting the efficiency of the static embedding methods. There is a 10-times speed up of local embeddings over the transformer-based model.

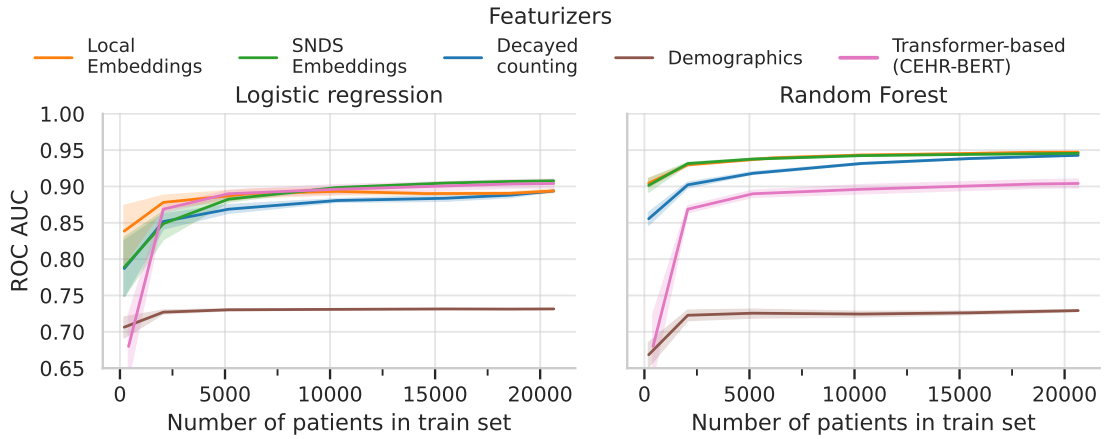


Fig. 3.6. Length of stay ROC AUC for different featurizers and estimators. The performance is averaged over 5 folds. The shaded area represents the standard deviation. The task performance seems to saturate at 95% ROC AUC for random forest and all featurizers but the demographics and CEHR-BERT, suggesting that the Bayes error rate is reached. However, for lower sample regimes –below 12,500 patients, we see a clear benefit of static embeddings over decayed counting of events (both for logistic regression and random forest). Appendix C.7.1 details AUPRC.

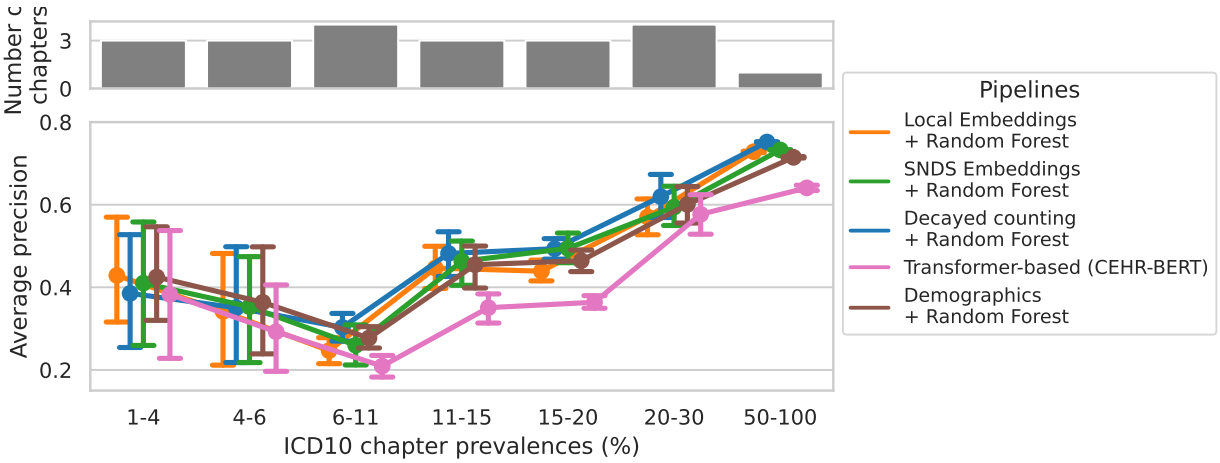


Fig. 3.7. Higher target prevalences yield better AUPRC. The different chapters are binned in prevalence bins. The estimator used for this plot is random forest trained on the full effective train set of size 8560. Appendix C.7.2 details the ROC AUC curves.

3.4 Conclusion

Training state-of-the-art predictive models from EHRs from scratch requires important computing resources, and access to very large cohorts. We explored a performance-complexity trade-off by studying different types of predictive algorithms, from simple baseline to large transformer-based pipelines.

Tree-based models outperform other pipelines For simple predictive tasks such as LOS interpolation or prognosis at the ICD10 level and small sample sizes, simple pipelines based either on decayed countings or static embeddings followed by random forest are sufficient to reach good performance. Transformer-based models struggle with these low-sample regimes. In low computing resources environments, static embedding pipelines are twice as fast as decayed counting and more than five times faster than transformers.

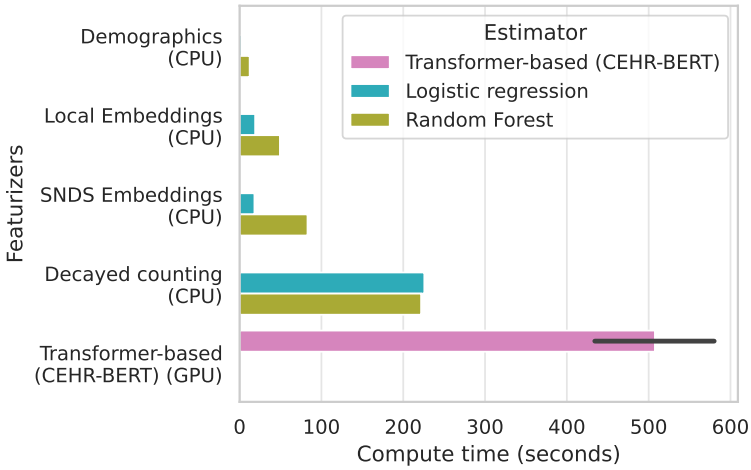


Fig. 3.8. Static embeddings are quicker to train. The training times are reported for the full train set for the prognosis task averaged on the 21 ICD10 chapters. For all featurizers with the exception of CEHR-BERT, jobs have been submitted with 10 CPU cores and 36GB memory on a slurm cluster. For CEHR-BERT, the calculation took place on the same cluster with a Nvidia T4 GPU with 16Gb of memory.

MACE is challenging due to low prevalence MACE is more complicated to evaluate because of the low prevalence of the outcome. This forced us to use a larger cohort to collect enough cases. Both simple and complex models require larger computing resources than those we can access. However, the type of these resources are different. Static embeddings can reach satisfying performance on big samples with large memory consumption (e.g., 100GB of RAM). Transformer architectures require large GPUs to be trained even for small sample sizes. Because of constraints for these two types of resources on the AP-HP computing cluster, we are currently struggling to evaluate the MACE task on the full cohort.

Comparing to bigger or pretrained models The high number of laboratory measurements –only available for inpatients– could improve the performance for all tasks. However, the addition of these high-frequency features exceeds the computing resources currently at our disposal. Our benchmark also lacks a pre-trained model on large structured data and the interesting avenue of repurposing large language model for predictive tasks directly from clinical notes. These two approaches require large GPU resources, making a poor use of the commodity hardware already present in hospitals. Healthcare financial resources are already scarce in rich countries. It is still an open question to assess if the prevention benefits brought by large scale models from EHRs outweigh the cost of their deployment.

Chapter 4

Prediction is not all we need: Causal thinking for decision making on Electronic Health Records

> *All scientific work is incomplete - whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time.*
>
> *–Austin Bradford Hill, The Environment and Disease: Association or Causation?, 1965*

Chapter's content

While Chapter 3 focused on predictive models for patient outcomes, this chapter details the importance of causality to build clinically-valuable models and introduces a framework to facilitate such endeavor. It shows that predictions –even accurate as with machine learning, may not suffice to provide optimal healthcare for every patient. Indeed, prediction can be driven by shortcuts in the data, such as racial biases. Causal thinking is needed for data-driven decisions. Here, we give an introduction to the key elements, focusing on routinely-collected data, Electronic Health Records (EHRs) and claims data. Using such data to assess the value of an intervention requires care: temporal dependencies and existing practices easily confound the causal effect. We present a step-by-step framework to help build *valid* decision making from real-life patient records by emulating a randomized trial before individualizing decisions, *eg* with machine learning. Our framework highlights the most important pitfalls and considerations in analyzing EHRs or claims data to draw causal conclusions. We illustrate the various choices in studying the effect of albumin on sepsis mortality in the Medical Information Mart for Intensive Care database (MIMIC-IV). We study the impact of various choices at every step, from feature extraction to causal-estimator selection. In a tutorial spirit, the code and the data are openly available.

This chapter corresponds to the article entitled *Step-by-step causal analysis of Electronic Health Records to ground decision making* submitted to *npj Digital Medicine*,

Authors: Matthieu Doutreligne, Tristan Struja, Judith Abecassis, Claire Morgand, Leo Anthony Celi and Gaël Varoquaux.

Outline

4.1 Motivation : Healthcare is concerned with decision making, not mere prediction

Medicine increasingly relies on data with the promise of better clinical decision-making Machine learning is central to this endeavor. On medical images, it achieves human-level performance to diagnose various conditions (Aggarwal et al., 2021; Esteva et al., 2021; Liu et al., 2019). Using Electronic Health Records (EHRs) or administrative data, it outperforms traditional rule-based clinical scores to predict a patient's readmission risk, mortality, or future comorbidities (Rajkomar et al., 2018b; Li et al., 2020b; Beaulieu-Jones et al., 2021). And yet, there is growing evidence that machine-learning models may not benefit patients equally. They reproduce and amplify biases in the data (Rajkomar et al., 2018a), such as gender or racial biases (Singh et al., 2022; Gichoya et al., 2022; Röösli et al., 2022), or marginalization of under-served populations (Seyyed-Kalantari et al., 2021). The models typically encode these biases by capturing shortcuts: stereotypical features in the data or unequal sampling (Geirhos et al., 2020; Winkler et al., 2019; DeGrave et al., 2021). For instance, an excellent predictive model of mortality in the Intense Care Unit (ICU) might be of poor clinical value if it uses information available only too late. These shortcuts are at odds with healthcare's ultimate goal: appropriate care for optimal health outcome for each and every patient (Canadian Medical Association, 2015; Ghassemi et al., 2020). Making the right decisions requires more than accurate predictions.

Causal thinking is a key ingredient to ground data-driven decision making (Prosperi et al., 2020) Indeed, decision-making logic cannot rely purely on learning from the data, which itself results from a history of prior decisions (Plecko; Bareinboim, 2022). Rather, reasoning about a putative intervention requires comparing the potential outcomes with and without the intervention, the difference between these being the causal effect. In medicine, causal effects are typically measured by Randomized Controlled Trials (RCTs). Yet, RCTs may not suffice for individualized decision making: They may suffer from selection biases (Travers et al., 2007; Averitt et al., 2020), failure to recruit disadvantaged groups, and become outdated by evolving clinical practice. Their limited sample size seldom allows to explore treatment heterogeneity across subgroups. Rather, routinely-collected data naturally probes real-world practice and displays much less sampling bias. It provides a unique opportunity to assess benefit-risk trade-offs associated with a decision (Desai et al., 2021), with sufficient data to capture heterogeneity (Rekkas et al., 2023). Estimating causal effects from this data is challenging however, as the intervention is far from being given at random, and, as a result, treated and untreated patients cannot be easily compared. Without dedicated efforts, machine-learning models simply pick up these difference and are not usable for decision

making. Rather dedicated statistical techniques are needed to emulate a "target trial" from *observational* data – without controlled interventions.

Different time resolutions EHRs and claims are two prominent sources of real-life healthcare data with different time resolutions. EHRs are particularly suited to guide clinical decisions, as they are rich in high-resolution and time-varying features, including vital signs, laboratory tests, medication dosages, etc. Claims, on the other hand, inform best on medico-economic questions or chronic conditions as they cover in-patient and out-patient care during extended time periods. But there are many pitfalls to sound and valid causal inferences (Hernan et al., 2019; Schneeweiss; Patorno, 2021). Data with temporal dependencies, as EHRs and claims, are particularly tricky, as it is easy to induce time-related biases (Suissa, 2008; Wang et al., 2023b).

Objectives and structure of the chapter Here we summarize the main considerations to derive valid decision-making evidence from EHRs and claims data. Many guidelines on causal inference from observational data have been written in various fields such as epidemiology (Hernàn; Robins, 2020; Schneeweiss; Patorno, 2021; Zeng et al., 2022), statistics (Belloni et al., 2014; Chernozhukov et al., 2018b), machine learning (Shalit; Sontag, 2016; Sharma, 2018; Moraffah et al., 2021) or econometrics (Imbens; Wooldridge, 2009). Time-varying features of EHR data, however, raise particular challenges that call for an adapted framework. We focus on single interventions: only one prescription during the study period, e.g., a patient either receives mechanical ventilation or not during admission to an intensive care unit compared to, e.g., blood transfusion which may be given repeatedly. Section 4.2 details our proposed step-by-step analytic framework on EHR data. Section 4.3 instantiates the framework by emulating a trial on the effect of albumin on sepsis using the Medical Information Mart for Intensive Care database (MIMIC-IV) database (Johnson et al., 2020). Section 4.4 discusses our results and its implications on sound decision making. These sections focus on being accessible, appendices and online Python code[1] expand more technical details, keeping a didactic flavor.

4.2 Step-by-step framework for robust decision making from EHR data

The need for a causal framework, even with machine learning Data analysis without causal framing risks building shortcuts. As an example of such failure, we trained a predictive model for 28-day mortality in patients with sepsis within the ICU. We fit the model using clinical measures available during the first 24 hours after admission. To simulate using this model to decide whether or not to administrate resuscitation fluids, we evaluate its performance on unseen patients first on the same measures as the ones used in training, and then using only the measures available before this treatment, as would be done in a decision making context. The performance drops markedly: from 0.80 with all the measures available during the first 24 hours after admission to 0.75 using only the measures available before the treatment (unit: Area Under the Curve of the Receiving Operator Characteristic, ROC AUC). The model has captured shortcuts: good prediction based on the wrong features of the data, useless for decision making. On the opposite, a model trained on pre-treatment measures achieves 0.79 in the decision-making setting (further details in appendix D.1). This

[1] https://github.com/soda-inria/causal_ehr_mimic/

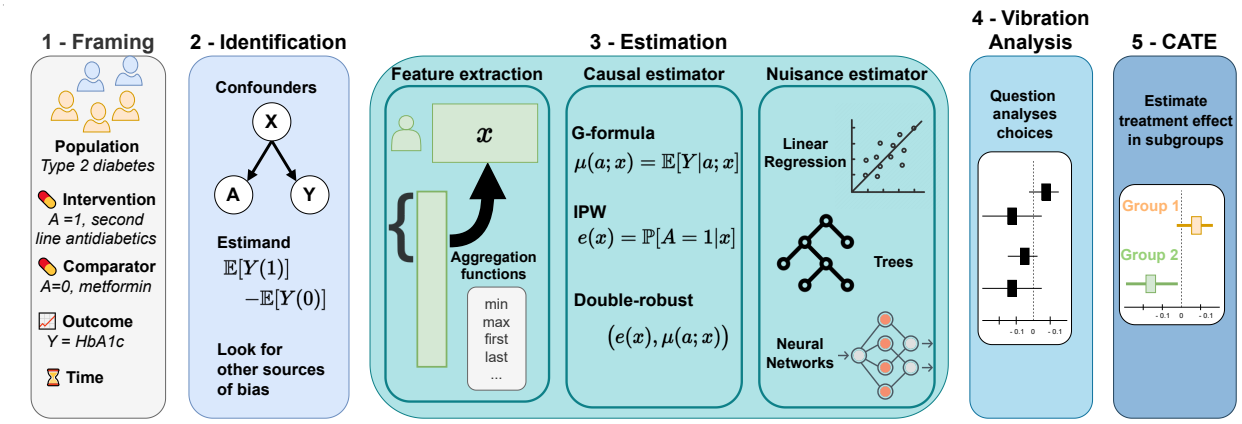


Fig. 4.1. Step-by-step analytic framework – The complete inference pipeline confronts the analyst with many choices, some guided by domain knowledge, others by data insights. Making those choices explicit is necessary to ensure robustness and reproducibility.

illustrates the importance of accounting for the putative interventions even for predictive models.

Whether a data analysis uses machine learning or not, many pitfalls threaten its value for decision making. To avoid these traps, we outline in this section a simple step-by-step analytic framework, illustrated in Figure 4.1. We first study the medical question as a target trial (Hernan, 2021), the common evidence for decisions. This enables assessing the validity of the analysis before probing heterogeneity –predictions on sub-groups– for individualized decision.

4.2.1 Step 1: study design – Frame the question to avoid biases

PICO(T) format Grounding decisions on evidence needs well-framed questions, defined by their PICO components: Population, Intervention, Control, and Outcome (Richardson et al., 1995). To concord with a (hypothetical) target randomized clinical trial, an analysis must emulate all these components (Hernán; Robins, 2016; Wang et al., 2023b), *eg* via *potential outcome* statistical framework (Hernàn; Robins, 2020) –Table 4.1 and Figure 4.2. EHRs and Claims need an additional time component: PICOT (Riva et al., 2012).

Without dedicated care, defining those PICO(T) components from EHRs can pick up bias: non-causal associations between treatment and outcomes. We detail two common sources of bias in the Population and Time components: selection bias and immortal time bias, respectively.

PICO component	Description	Notation	Example
Population	What is the target population of interest?	$X \sim \mathbb{P}(X)$, the covariate distribution	Patients with sepsis in the ICU
Intervention	What is the treatment?	$A \sim \mathbb{P}(A = 1) = p_A$, the probability to be treated	Crystalloids and albumin combination
Control	What is the clinically relevant comparator?	$1 - A \sim 1 - p_A$	Crystalloids only
Outcome	What are the outcomes ?	$Y(1), Y(0) \sim \mathbb{P}(Y(1), Y(0))$, the potential outcomes distribution	28-day mortality
Time	Is the start of follow-up aligned with intervention assignment?	N/A	Intervention administered within the first 24 hours of admission

Table 4.1. PICO(T) components help to clearly define the medical question of interest.

Selection Bias In EHRs, outcomes and treatments are often not directly available and need to be inferred from indirect events. These signals could be missing not-at random, sometimes correlated with the treatment allocation (Weiskopf et al., 2023). For example, not all billing codes are equally well filled in, as billing is strongly associated with case-severity and cost. Consider comparing the effect on mortality of fluid resuscitation with albumin to that of crystalloids. As albumin is much more costly, patients who have received this treatment are much more likely to have a sepsis billing code, independent of the seriousness of their condition. On the contrary, for patients treated with crystalloids, only the most severe cases will have a billing code. Naively comparing patients on crystalloid treatment with less sick patients on albumin treatment would overestimate the effect of albumin.

Immortal time bias Another common bias comes from timing: improper alignment of the inclusion defining event and the intervention time (Suissa, 2008; Hernan et al., 2016; Wang et al., 2022). Figure 4.3 illustrates this Immortal time bias –related to survivor bias (Lee; Nunan, 2020). It occurs when the follow-up period, i.e. cohort entry, starts before the intervention, e.g., prescription for a second-line treatment. In this case, the treated group will be biased towards patients still alive at the time of assignment and thus overestimating the effect size. Other common temporal biases are lead time bias (Oke et al., 2021; Fu et al., 2021), right censorship (Hernan et al., 2016), and attrition bias (Bankhead C, 2017).

Good practices include explicitly stating the cohort inclusion event (OHDSI, 2021, Chapter 10:Defining Cohorts) and defining an appropriate grace period between starting time and the intervention assignment (Hernan et al., 2016). At this step, a population timeline can help (e.g., Figure 4.5).

4.2.2 Step 2: identification – List necessary information to answer the causal question

The identification step builds a causal model to answer the research question (Figure 4.6). Indeed, the analysis must compensate for differences between treated and non-treated that are not due to the intervention (Pearl; Mackenzie, 2018, chapter 1, Hernàn; Robins, 2020, chapter 1).

Background Information

Causal Assumptions Not every question can be answered from a given dataset: valid causal inference requires assumptions. We assume the following four assumptions, referred as strong ignorability and necessary to assure identifiability of the causal estimands with

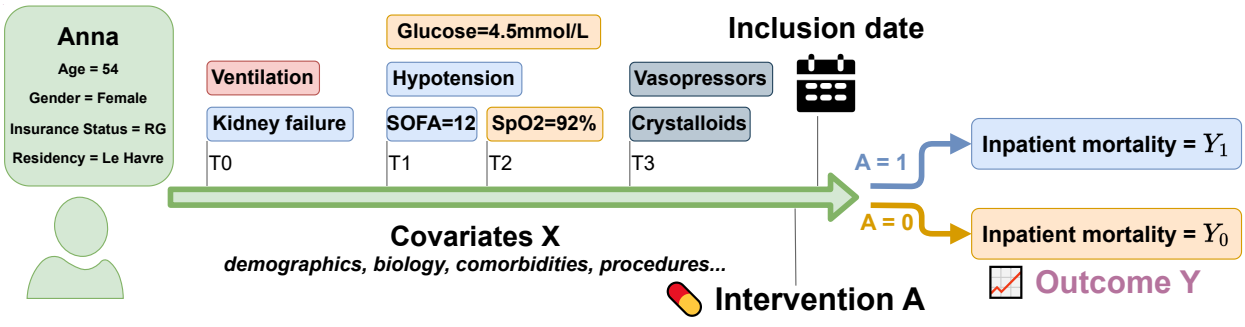


Fig. 4.2. Study design – The first step of the analysis consists in identifying a valid treatment effect question from patient healthcare trajectories and defining a target trial emulating a RCT using the PICO(T) framework.

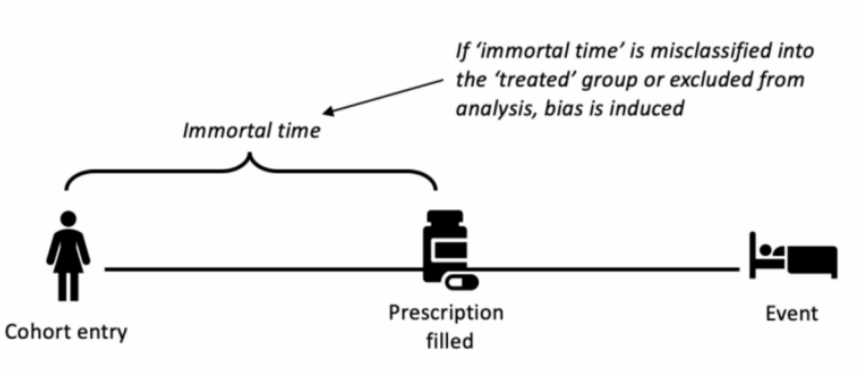


Fig. 4.3. Poor experimental design can introduce Immortal time bias, which leads to a treated group with falsely longer longevity (Lee; Nunan, 2020).

observational data (Rubin, 2005). See Naimi; Whitcomb, 2023 for a concise introduction for epidemiologists:

Assumption 1 (Unconfoundedness)

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

This condition –also called ignorability– is equivalent to the conditional independence on $e(X)$ (Rosenbaum; Rubin, 1983): $\{Y(0), Y(1)\} \perp\!\!\!\perp A|e(X)$.

Assumption 2 (Overlap, also known as Positivity))

$$\eta < e(x) < 1 - \eta \quad \forall x \in \mathcal{X} \text{ and some } \eta > 0$$

The treatment is not perfectly predictable. Or with different words, every patient has a chance to be treated and not to be treated. For a given set of covariates, we need examples of both to recover the ATE.

As noted by D'Amour et al., 2021, the choice of covariates X can be viewed as a trade-off between these two central assumptions. A bigger covariates set generally reinforces the ignorability assumption. On the contrary, overlap can be weakened by large $\mathcal{X}$ because of the potential inclusion of instruments: variables only linked to the treatment which could lead to arbitrarily small propensity scores.

Assumption 3 (Consistency) *The observed outcome is the potential outcome of the assigned treatment:*

$$Y = A\,Y(1) + (1 - A)\,Y(0)$$

Here, we assume that the intervention A has been well defined. This assumption focuses on the design of the experiment. It clearly states the link between the observed outcome and the potential outcomes through the intervention (Hernán; Robins, 2020).

Assumption 4 (Generalization) *The training data on which we build the estimator and the test data on which we make the estimation are drawn from the same distribution $\mathcal{D}^*$, also known as the "no covariate shift" assumption (Jesson et al., 2020).*

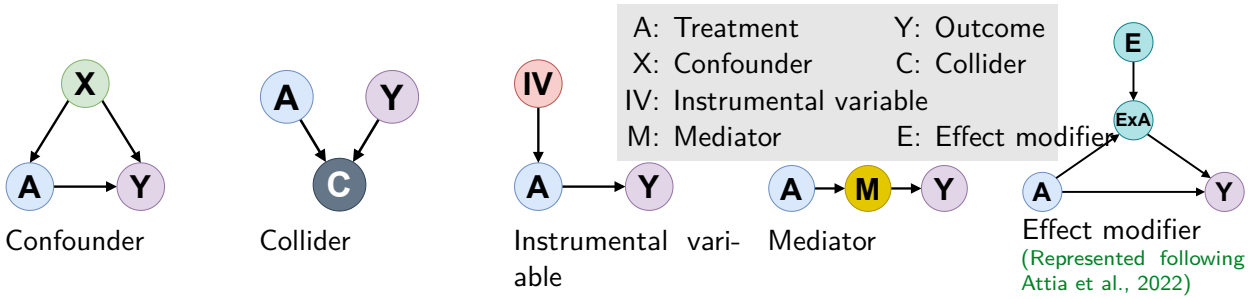


Fig. 4.4. The five categories of causal variables needed for our framework.

Background Information

Categorizing covariates Potential predictors –covariates– should be categorized depending on their causal relations with the intervention and the outcome (Figure 4.4): *confounders* are common causes of the intervention and the outcome; *colliders* are caused by both the intervention and the outcome; *instrumental variables* are a cause of the intervention but not the outcome, *mediators* are caused by the intervention and is a cause of the outcome. Finally, *effect modifiers* interact with the treatment, and thus modulate the treatment effect in subpopulations (Attia et al., 2022).

Background Information

To capture a valid causal effect, the analysis should only include confounders and possible treatment-effect modifiers to study the resulting heterogeneity. Regressing the outcome on instrumental and post-treatment variables (colliders and mediators) will lead to biased causal estimates (VanderWeele, 2019). Drawing causal Directed Acyclic Graphs (DAGs) (Greenland et al., 1999), *eg* with a webtool such as DAGitty (Textor et al., 2011), helps capturing the relevant variables from domain expertise.

Background Information

Estimand or effect measure The *estimand* is the final statistical quantity estimated from the data. Depending on the question, different estimands are better suited to contrast the two potential outcomes E[Y(1)] and E[Y(0)] (Imbens, 2004; Colnet et al., 2023). For continuous outcomes, risk difference is a natural estimand, while for binary outcomes (e.g., events) the choice of estimand depends on the scale of the study. Whereas the risk difference is very informative at the population level, e.g., for medico-economic decision making, the risk ratio and the hazard ratio are more informative to reason on sub-populations such as individuals or sub-groups (Colnet et al., 2023).

4.2.3 Step 3: Estimation – Compute the causal effect of interest

Confounder aggregation Some confounders are captured via measures collected over multiple time points. These need to be aggregated at the patient level. Simple forms of aggregation include taking the first or last value before a time point, or an aggregate such as mean or median over time. More elaborate choices may rely on hourly aggregations of information such as vital signs. These provide more detailed information on the health evolution, thus reducing confounding bias between rapidly deteriorating and stable patients. However, it also increases the number of confounders, resulting in a larger covariate space, hence increasing the estimate's variance and endangering the positivity assumption. The choices should be guided by expert knowledge. If multiple choices appear reasonable, one should compare them in a vibration analysis (see Section 4.2.4). Indeed, aggregation may impact results, as Sofrygin et al., 2019 show, revealing that some choices of averaging time

scale lead to inconclusive links between HbA1c levels and survival in diabetes.

Beyond measures and clinical codes, unstructured clinical text may capture confounding or prognostic information (Horng et al., 2017; Jiang et al., 2023) which can be added in the causal model (Zeng et al., 2022).

Causal estimators or statistical modeling A given estimand can be estimated through different methods. One can model the outcome with regression models (also known as G-formula, Robins; Greenland, 1986) and use it as a predictive counterfactual model for all possible treatments for a given patient. Alternatively, one can model the propensity of being treated use it for matching or Inverse Propensity Weighting (IPW) (Austin; Stuart, 2015). Finally, doubly robust methods model both the outcome and the treatment, benefiting from the convergence of both models (Wager, 2020b). Various doubly robust models have emerged: Augmented Inverse Propensity Score (AIPW) (Robins et al., 1994), Double Robust Machine Learning (Chernozhukov et al., 2018b), or Targeted Maximum Likelihood Estimation (TMLE) (Schuler; Rose, 2017) to name a few. We detail their statistical properties in Appendix D.4.1, giving some hints on when to choose one method over the others.

Estimation models of outcome and treatment The causal estimators use models of the outcome or the treatment –called nuisances as they are not the main inference targets in our causal effect estimation problem. Which statistical model is best suited is an additional choice and there is currently no clear best practice (Wendling et al., 2018b; Dorie et al., 2019). The trade-off lies between simple models risking misspecification of the nuisance parameters versus flexible models risking to overfit the data at small sample sizes. Stacking models of different complexity in a super-learner is a good solution to navigate the trade-off (Van der Laan et al., 2007; Doutreligne; Varoquaux, 2023).

4.2.4 Step 4: Vibration analysis – Assess the robustness of the hypotheses

Some choices in the pipeline may not be clear cut. Several options should then be explored, to derive conceptual error bars going beyond a single statistical model. This process is sometimes called robustness analysis (Neumayer; Plümper, 2017) or sensitivity analysis (Thabane et al., 2013; Hernàn; Robins, 2020; FDA, 2021b). However, in epidemiology, sensitivity analysis refers to quantifying the bias from unobserved confounders (Schneeweiss, 2006). Following Patel et al., 2015, we use the term vibration analysis to describe the sensitivity of the results to all analytic choices. The vibration analysis can identify analytic choices that deserve extra scrutiny. It complements a comparison to previous studies –ideally RCTs– to establish the validity of the pipeline.

4.2.5 Step 5: Treatment heterogeneity – Compute treatment effects on subpopulations

Once the causal design and corresponding estimators are established, they can be used to explore the variation of treatment effects among subgroups. Measures of the heterogeneity of a treatment nourish decisions tailored to a patient's characteristics. A causally-grounded model, *eg* using machine learning, can be used to predict the effect of the treatment from all the covariates –confounders and effect modifiers– for an individual: the *Individual Treatment Effect* (ITE Lu et al., 2018). Studying heterogeneity only along specific covariates, or a

given patient stratification, is related to the *Conditional Average Treatment Effect* (CATE) (Robertson et al., 2021). Practically, CATEs can be estimated by regressing the individual predictions given by the causal estimator against the sources of heterogeneity (details in D.11.3).

4.3 Application: evidence from MIMIC-IV on which resuscitation fluid to use

We now use the above framework to extract evidence-based decision rules for resuscitation. Ensuring optimal organ perfusion in patients with septic shock requires resuscitation by reestablishing circulatory volume with intravenous fluids. While crystalloids are readily available, inexpensive and safe, a large fraction of the administered volume is not retained in the vasculature. Colloids offer the theoretical benefit of retaining more volume in the circulation, but might be more costly and have adverse effects (Annane et al., 2013). The scientific community long debated which fluid benefits patients most (Mandel; Palevsky, 2023).

Emulated trial: Effect of albumin in combination with crystalloids compared to crystalloids alone on 28-day mortality in patients with sepsis We illustrate the impact of the different analytical steps to conclude on the effect of albumin in combination with crystalloids compared to crystalloids alone on 28-day mortality in patients with sepsis using MIMIC-IV (Johnson et al., 2020). This question is clinically relevant and multiple published RCTs can validate the average treatment effect. Appendix D.3 provides further examples of potential target trials.

Evidence from the literature Meta-analyses from multiple pivotal RCTs found no effect of adding albumin to crystalloids (Li et al., 2020a) on 28-day and 90-day mortality. Further, an observational study in MIMIC-IV (Zhou et al., 2021b) found no significant benefit of albumin on 90-day mortality for severe sepsis patients. Given this previous evidence, we thus expect no average effect of albumin on mortality in sepsis patients. However, studies –RCT (Caironi et al., 2014) and observational (Li et al., 2020a)– have found that septic-shock patients do benefit from albumin.

4.3.1 Study design: effect of crystalloids on mortality in sepsis

- Population: Patients with sepsis within the ICU stay according to the sepsis-3 definition. Other inclusion criteria: sufficient follow-up of at least 24 hours, and age over 18 years described in table 4.2.
- Intervention: Treatment with a combination of crystalloids and albumin during the first 24 hours of an ICU stay.
- Control: Treatment with crystalloids only in the first 24 hours of an ICU stay.
- Outcome: 28-day mortality.
- Time: Follow-up begins after the first administration of crystalloids. Thus, we potentially introduce a small immortal time bias by allowing a time gap between follow-up and the start of the albumin treatment –shown in Figure 4.5. Because we are only considering the first 24 hours of an ICU stay, we hypothesize that this gap is insufficient to affect our results. We test this hypothesis in the vibration analysis step 4.3.4.

	Missing	Overall	Cristalloids only	Cristalloids + Albumin	P-Value
n		18421	14862	3559	
Female, n (%)		7653 (41.5)	6322 (42.5)	1331 (37.4)	
White, n (%)		12366 (67.1)	9808 (66.0)	2558 (71.9)	
Emergency admission, n (%)		9605 (52.1)	8512 (57.3)	1093 (30.7)	
admission_age, mean (SD)	0	66.3 (16.2)	66.1 (16.8)	67.3 (13.1)	<0.001
SOFA, mean (SD)	0	6.0 (3.5)	5.7 (3.4)	6.9 (3.6)	<0.001
lactate, mean (SD)	4616	3.0 (2.5)	2.8 (2.4)	3.7 (2.6)	<0.001

Table 4.2. Characteristics of the trial population measured on the first 24 hours of ICU stay. Appendix D.5 describes all confounders used in the analysis.

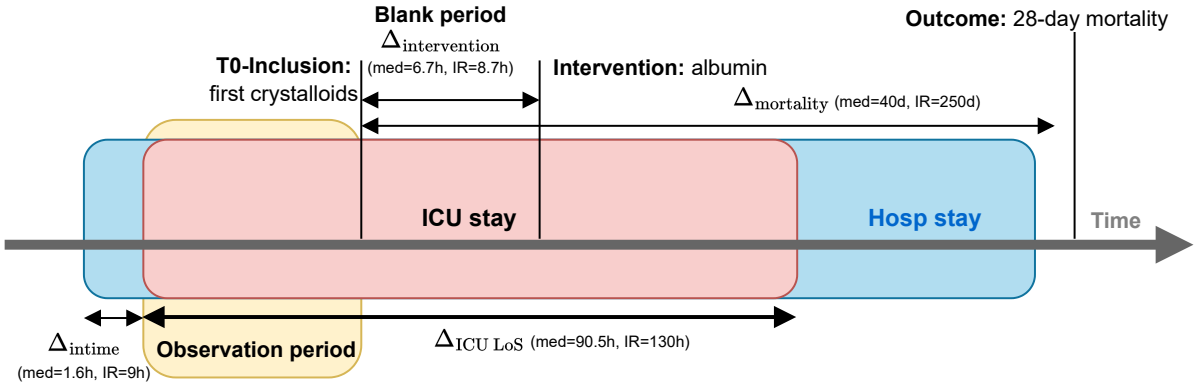


Fig. 4.5. Defining the inclusion event, the starting time T0 for follow-up, the intervention's assignment time and the observation window for confounders is crucial to avoid time and selection biases. In our study, the gap between the intervention and the inclusion is small compared to the occurrence of the outcome to limit immortal time bias: 6.7 hours vs 40 days for mortality.

In MIMIC-IV, these inclusion criteria yield 18,121 patients with 3,559 patients treated with a combination of crystalloids and albumin (Appendix D.6 details the selection flowchart).

4.3.2 Identification: listing confounders

We enrich the confounders selection procedure described by Zhou et al., 2021b with expert knowledge, creating the causal DAG shown in Figure 4.6. Gray confounders are not controlled for, since they are not available in the data. However, resulting confounding biases are captured by proxies such as comorbidity scores (SOFA or SAPS II) or other variables (e.g., race, gender, age, weight). Appendix D.7 details confounders summary statistics for treated and controls.

4.3.3 Estimation

Confounder aggregation We tested multiple aggregations such as the last value before the start of the follow-up period, the first observed value, and both the first and last values as separated features.

Causal estimators We implemented multiple estimation strategies, including Inverse Propensity Weighting (IPW), outcome modeling (G-formula) with T-Learner, Augmented Inverse Propensity Weighting (AIPW) and Double Machine Learning (DML). We used the

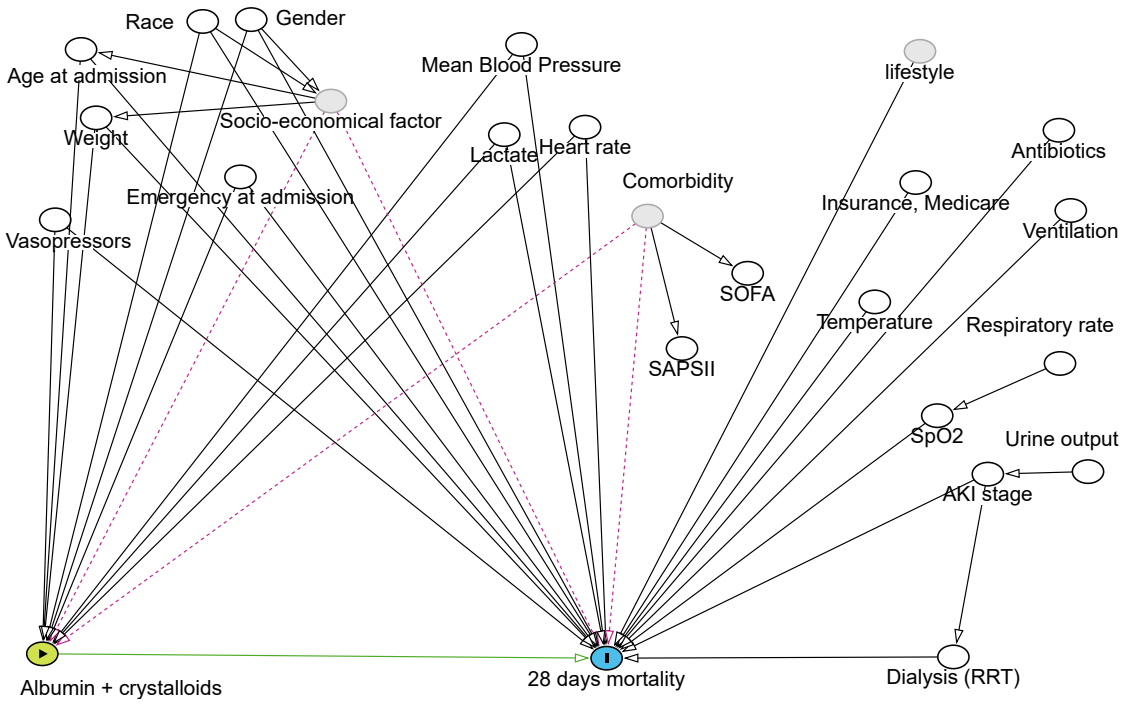


Fig. 4.6. Causal graph for the Albumin vs crystalloids emulated trial – The green arrow indicates the effect studied. Black arrows show causal links known to medical expertise. Dotted red arrows highlight confounders not directly observed. For readability, we draw only the most important edges from an expert point of view. All white nodes corresponds to variables included in our study.

python packages dowhy (Sharma, 2018) for IPW implementation and EconML (Battocchi et al., 2019) for all other estimation strategies. Confidence intervals were estimated by bootstrap (50 repetitions). Appendices D.4.1 and D.4.3 detail the estimators and the available Python implementations.

Outcome and treatment estimators To model the outcome and treatment, we used two common but different estimators: random forests and ridge logistic regression implemented with scikit-learn (Pedregosa et al., 2011). We chose the hyperparameters with a random search procedure (detailed in Appendix D.4.4). While logistic regression handles predictors in a linear fashion, random forests should have the benefit of modeling non-linear relations as well.

4.3.4 Vibration analysis: Understanding variance or sources of systematic errors in our study

Varying estimation choices – Confounders aggregation, causal and nuisance estimators Figure 4.7 shows varying confidence intervals (CI) depending on the method. Doubly-robust methods provide the narrowest CIs, whereas the outcome-regression methods have the largest CI. The estimates of the forest models are closer to the consensus across prior studies (no effect) than the estimates from the logistic regression indicating a better fit of the non-linear relationships in the data. We only report the first and last pre-treatment feature aggregation strategies, since detailed analysis showed little differences for other choices of feature aggregation (see Appendix D.8). Confronting this analysis with the prior published evidence of little-to-no effect, it seems reasonable to select the models using random forests for nuisance. Out of these, theory suggests to trust more double machine learning or doubly

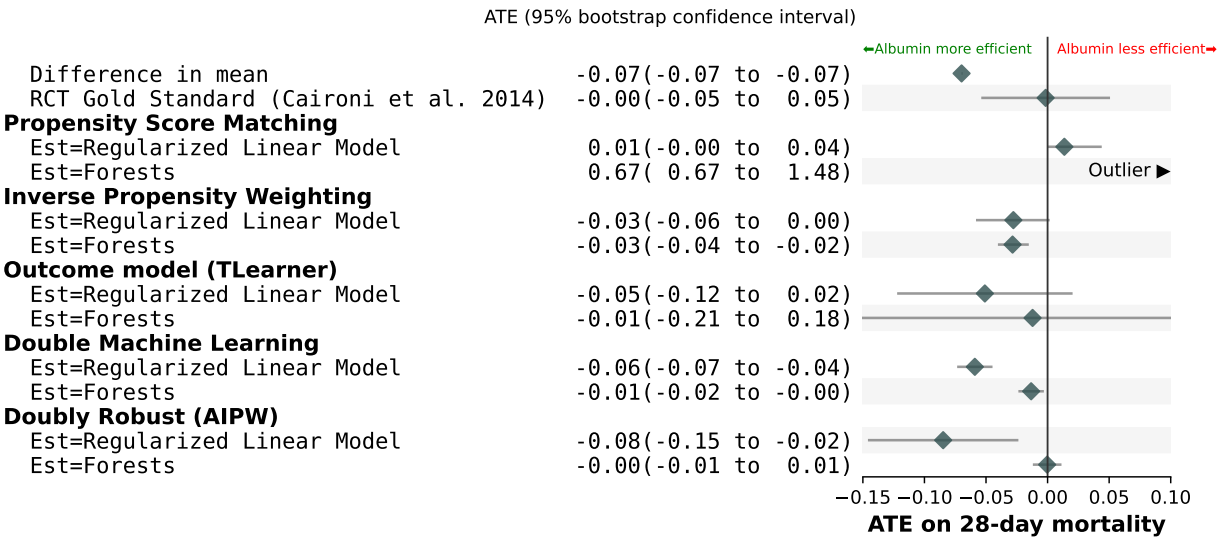


Fig. 4.7. Forest plot for the vibration analysis – Different estimators give different results, sometimes even outside of each-other's bootstrap confidence intervals. Score matching yields unconvincingly high estimates, inconsistent with the published RCT. With other causal approaches, using linear estimators for nuisances suggest a reduced mortality risk for albumin, while using forests for nuisance models points to no effect, which is consistent with the RCT gold standard. The diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 50 bootstrap repetitions.

robust approaches.

Study design – Illustration of immortal time bias To illustrate the risk of immortal-time bias, we varied the eligibility period by allowing patients to receive the treatment or the control in a shorter or longer time window than 24 hours. As explained in Subsection 4.2.1, a large eligibility period means that patients in the study are more likely to be treated if they survived till the intervention and hence the study is biased to overestimate the beneficial effect of the intervention. Figure 4.8 shows that larger eligibility periods change the direction of the estimate and lead to Albumin seeming markedly more efficient. Should the analyst not have in mind the mechanism of immortal time bias, this vibration analysis ought to raise an alarm and hopefully lead to correct the study design.

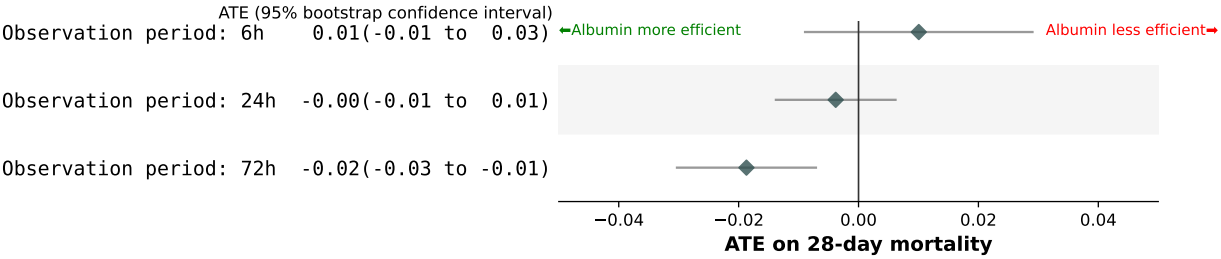


Fig. 4.8. Detecting immortal time bias – Increasing the observation period increases the temporal blank period between inclusion and treatment initialization, associating thus patients surviving longer with treatment: Immortal Time Bias. A longer observation period (72h) artificially favors the efficacy 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 D.9. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 30 bootstrap repetitions.

4.3.5 Treatment heterogeneity: Which treatment for a given sub-population?

We now study treatment heterogeneity using the pipeline validated by confronting the vibration analysis to the literature: a study design avoiding immortal time bias, and the double machine learning model using forest for nuisances and a linear model for the final heterogeneity regression. We explore heterogeneity along four binary patient characteristics, displayed on Figure 4.9. We find that albumin is beneficial with patient with septic shock before fluid administration, consistent with the Caironi et al., 2014 RCT. It is also beneficial for older patients (age >=60) and males, consistent with (Zhou et al., 2021b), as well as white patients.

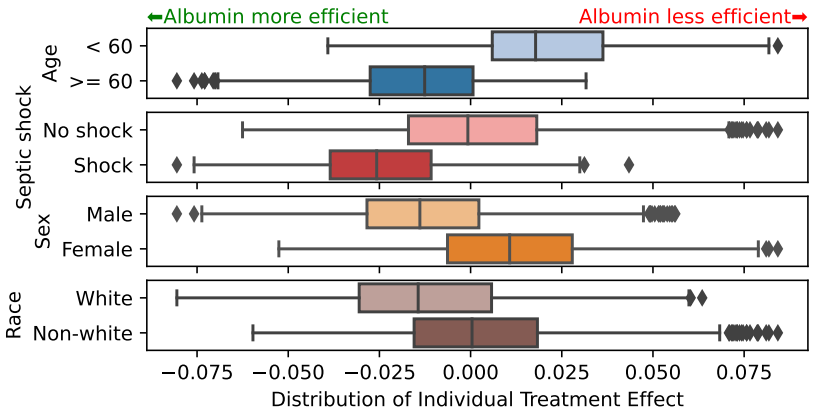


Fig. 4.9. The subgroup distributions of Individual Treatment effects showed better treatment efficacy for patients older than 60 years, septic shock, and to a lower extent males. The final estimator is ridge regression. The boxes contain the 25^{th} and 75^{th} percentiles of the CATE distributions with the median indicated by the vertical line. The whiskers extend to 1.5 times the inter-quartile range of the distribution.

4.4 Discussion and conclusion

A didactic causal framework for decision-making from EHR Our analytic framework strives to streamline extracting valid decision-making rules from EHR data. Decision-making is tied to a choice: to treat or not to treat, for a given intervention. A major pitfall, source of numerous shortcuts of machine-learning systems, is to extract non-causal associations between the intervention and the outcome. Our framework is designed to avoid these pitfalls by starting with rigorous causal analysis, in the form of a target trial, to validate study design and analytic choices before more elaborate analysis, potentially using machine-learning for individual predictions. We argue that in the absence of a precise framing including treatment allocation, automated decision making is brittle. It is all too easy, for instance, to build a predictive system on post-treatment data, rendering it unreliable for decision making. EHR data come with particular challenges: information may be available indirectly, *e.g.*, via billing codes, the time-wise dimension requires aggregations (Sub-section 4.2.3). These challenges can create subtle causal biases (Subsection 4.2.1). To ensure that our framework addresses all aspects of EHR analysis and to expose it in a didactic way, we detailed a complete analysis of a publicly-available EHR dataset, supported by open code.

A well-framed target trial can be validated Assessing the validity of an analysis is challenging even for experts (Ioannidis, 2005; Breznau et al., 2022). Our framework recommends using a well-specific target trial to establish a valid pipeline because it helps confronting the resulting average treatment effect to other evidence (Hernán; Robins, 2016;

Wang et al., 2023b). Our resuscitation-fluid analysis matches well published findings: Pooling evidence from high-quality RCTs, no effect of albumin in severe sepsis was demonstrated for both 28-day mortality (odds ratio (OR) 0.93, 95% CI 0.80-1.08) and 90-day mortality (OR 0.88, 95% CI 0.76-1.01) (Xu et al., 2014). This consistency validates our study design and analytic choices. Varying analytic choices and confronting them to prior studies can reveal loopholes in the analysis, as we demonstrated with immortal time bias: extending the time between ICU admission and intervention to 72 hours, we observed an inflation of effect size consistent with such bias. Looping back to reference RCTs reveals that these include patients within 8 to 24 hours of ICU admission (SAFE Study Investigators, 2011; Annane et al., 2013; Caironi et al., 2014).

Decision-making from EHRs Once the causal analysis has been validated, it can be used for decision making. A sub-population analysis (as in Figure 4.9) can distill rules on which groups of patients should receive the treatment. Ideally, dedicated RCTs can be run with inclusion criteria matching these sub-groups. However, the cost and the ethical concerns of running RCTs limit the number of sub-groups that can be explored. In addition, the sub-group view risks oversimplifying, as opposed to patient-specific effect estimates to support more individualized clinical decision making (Kent et al., 2018). For this, predictive modeling shines. Causally-grounded machine learning can give good counterfactual prediction (Prosperi et al., 2020; Hernan et al., 2019; Richens et al., 2020), if it predicts well the treated and untreated outcomes as shown in 5.1. Even without focusing on a specific intervention, anchoring machine learning on causal mechanisms gives models that are more robust to distributional shift (Schölkopf et al., 2021), safer for clinical use (Richens et al., 2020), and more fair (Plecko; Bareinboim, 2022). Capturing individualized effects via machine-learning models does require many diverse individuals. EHRs and claims data are well suited for these models, as they easily cover more individuals than a typical clinical study.

EHRs and RCTs, complementary sources of evidence But EHRs cannot inform on trade-offs that have not been explored in the data. No matter how sophisticated, causal inference cannot conclude if there is no data to support an apple-to-apple comparison between treated and non-treated individuals. For example, treatment allocation is known to be influenced by race- and gender-concordance between the patient and the care provider. Yet, if the EHR data does not contain this information, it cannot nourish evidence-based decisions on such matter. EHRs and RCTs complement each other: a dedicated study, with a randomized intervention, as an RCT, can be crafted to answer a given question on a given population. But RCTs cannot address all the subpopulations, local practices, healthcare systems (Rothwell, 2006; Travers et al., 2007; Kennedy-Martin et al., 2015). Our framework suggest integrating the evidence from RCTs designed with matching PICO formulation to ensure the validity of the analysis and to use the EHR to explore heterogeneity.

Conclusion Without causal thinking machine learning does not suffice for optimal clinical decision making for each and every patient. It will replicate non-causal associations such as shortcuts improper for decision making. As models can pick up information such as race implicitly from the data (Adam et al., 2022), they risk propagating biases when building AI models which can further reinforce health disparities. This problem is acknowledged by the major tech companies which are deploying causal inference tooling to mitigate biases (Google, 2023; Microsoft, 2023; PwC, 2023). On the medical side, causal modeling can create actionable decision-making systems that reduce inequities (Mitra et al., 2022; Ehrmann et al.,

2023). However, as we have seen, subtle errors can make an intervention seemingly more –or less– beneficial to patients. No sophisticated data-processing tool can safeguard against invalid study design or modeling choices. The goal of our step-by-step analytic framework is to help the data analyst work around these loopholes, building models that avoid shortcuts and extract the best decision-making evidence. Applied to study the addition of albumin to crystalloids to resuscitate sepsis patients, it shows that this addition is not beneficial in general, but that it does improve survival on specific individuals, such as patients undergoing sceptic shock.

Chapter 5

How to select predictive models for causal inference?

Quand à moi, après une longue existence, je ne crois toujours point au "hasard", mais plutôt à une loi des coïncidences dont nous ne connaissons pas le méchanisme.
– Amadou Hampâté Bâ, Mémoires (1994)

Chapter's content

In the previous chapters, we showed the strong interest in predictive models for healthcare, bridging to increasingly complex machine learning algorithms. We also pointed out that even when giving likely outcomes, they are not immediately transposable to decision making –choosing whether to treat or not to treat. Such reasoning on the effect of an intervention is a causal-inference task. We demonstrated that causal thinking was necessary to avoid introducing biases in the study design or during confounders selection. But, even with a robust causal framework such as in Chapter 4, the practitioner is left to choose among the plethora of predictive models available for health data (some detailed in Appendix C.2.2). In a given situation, which of these models yield the most valid causal estimates? Here, we highlight that classic machine-learning model selection does not pick the best models for causal inference. Indeed, causal model selection should control both outcomes for each individual, treated or not treated, whereas only one outcome is observed. Theoretically, simple risks used in machine learning do not control causal effects when treated and non-treated population differ too much. More elaborate risks use "nuisances" re-weighting to approximate the causal error on the observed data. But does estimating these nuisances add noise to model selection? Drawing from an extensive empirical study, we outline an efficient causal model-selection procedure. To select the best predictive model to guide decisions: use the so-called R-risk, use flexible estimators to compute the nuisance models on the train set, and split out 10% of the data to compute risks.

This chapter corresponds to the article entitled *How to select predictive models for decision making or causal inference?* submitted to *Artificial Intelligence in Medicine*,

Authors: Matthieu Doutreligne, Gaël Varoquaux.

Outline

5.1 Motivation: causal predictive models cannot rely on the Machine Learning toolbox

5.1.1 Extending prediction to prescription needs causality

Progress in machine learning brings predictive models to new health data (Beam; Kohane, 2018; Rajkomar et al., 2019) Automated analysis of medical images is increasingly accurate, *eg* for brain images, (Khojaste-Sarakhsi et al., 2022; Zhang; Sejdić, 2019) or mammography (Yala et al., 2019; Shen et al., 2019; Nassif et al., 2022). New prognostic models leverage routinely-collected patient records (Mooney; Pejaver, 2018): predicting heart failure from claims (Desai et al., 2020), suicide attempts from questionnaires (Simon et al., 2018)... Clinical notes contain much prognostic information but require text modeling (Horng et al., 2017; Wang et al., 2020; Spasic, Nenadic, et al., 2020; Jiang et al., 2023). Data may be difficult to control and model, but the accuracy of the prediction can be verified on left-out data (Altman et al., 2009; Poldrack et al., 2020; Varoquaux; Colliot, 2022). Given a model predicting a health outcome, precision medicine would like it to guide decisions: will an individual benefit from an intervention such as surgery (Fontana et al., 2019)? Contrasting predictions with and without the treatment gives an answer, but statistical validity requires causal inference (Snowden et al., 2011; Blakely et al., 2020).

Causal-inference bridges to predictive modeling via the rich statistical literature on *outcome models* This estimation approach is also known as G-computation, G-formula (Robins; Greenland, 1986), Q-model (Snowden et al., 2011), conditional mean regression (Wendling et al., 2018a). A central challenge of inference of treatment effects is that of confounding: spurious associations between treatment allocation and baseline health, *eg* only prescribing a drug to mild cases (Hernán; Robins, 2020; VanderWeele, 2019). Controlled allocation of treatment, as in Randomized Controlled Trials (RCTs), alleviate this concern. Yet most machine-learning models are trained on *observational* data, close to real-word practice (Black, 1996; Hernán, 2021) but challenging for causal inference. Causal inference has been central to epidemiology, typically with methods that model treatment assignment (Austin; Stuart, 2015; Grose et al., 2020), based on propensity scores (Rosenbaum; Rubin, 1983). Recent empirical results (Wendling et al., 2018a; Dorie et al., 2019) show benefits of outcome modeling to estimate average treatment effects. Maybe a greater benefit is that these methods naturally go beyond average effects, estimating individualized or conditional average treatment effects (CATE), central to precision medicine. For this purpose, such methods are also invaluable on randomized trials (Su et al., 2018; Lamont et al., 2018; Hoogland et al., 2021).

Explosion of outcome modeling or machine learning methods Many deep-learning methods have been developed for medical image analysis (Shen et al., 2017; Monshi et al., 2020). Even outcome-modeling methods specifically designed for causal inference are numerous: Bayesian Additive Regression Trees (Hill, 2011), Targeted Maximum Likelihood Estimation (Laan; Rose, 2011; Schuler; Rose, 2017), causal boosting (Powers et al., 2018), causal multivariate adaptive regression splines (Powers et al., 2018), random forests (Wager; Athey, 2018; Athey et al., 2019), Meta-learners (Künzel et al., 2019), R-learners (Nie; Wager, 2017), Doubly robust estimation (Chernozhukov et al., 2018a)... The wide variety of methods raises the problem of selecting between different estimators based on the data at hand. Indeed, estimates of treatment effects can vary markedly across different predictive models. For instance, Figure 5.1 shows large variations obtained across different outcome estimators on semi-synthetic datasets (Dorie et al., 2019). Flexible models such as random forests are doing well in most settings except when treated and untreated populations differ noticeably, in which case a linear model (ridge) is to be preferred. However random forests with different hyper-parameters (max depth= 2) yield poor estimates. A simple rule of thumb such as preferring flexible models does not work in general; model selection is needed.

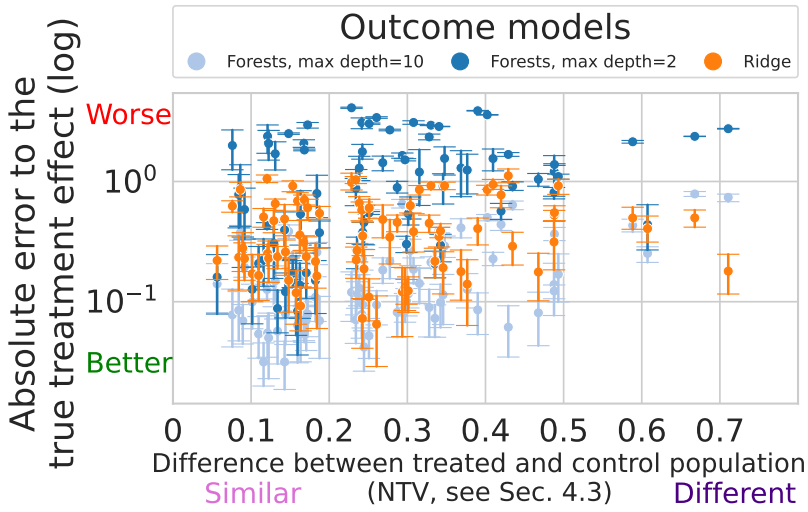


Fig. 5.1. Different outcome models lead to different estimation errors on the Average Treatment Effects, on 77 classic simulations with known true causal effect (Dorie et al., 2019). The different models are ridge regression and random forests with different hyper-parameters (details E.1). The different configurations are plotted as a function of increasing difference between treated and untreated population –see sous-section 5.4.3. There is no systematic best performer; data-driven model selection is important.

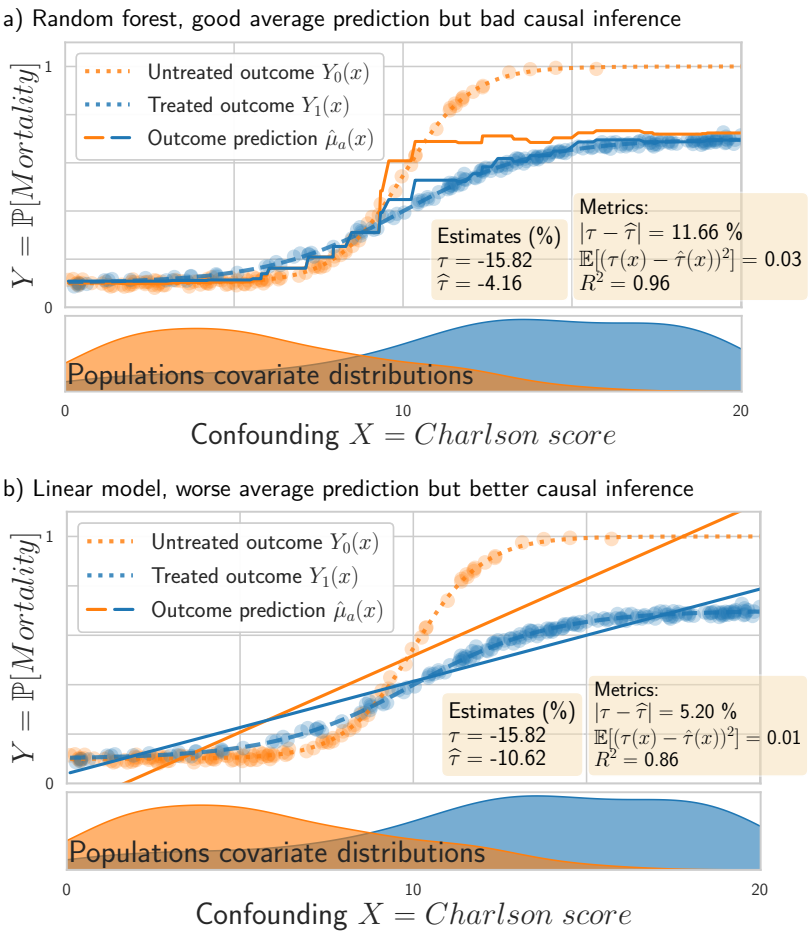
Standard practices to select models in predictive settings rely on the error on the outcome (Poldrack et al., 2020; Varoquaux; Colliot, 2022). However, as we will see, these practices may not pick the best models for causal inference, as they can be misled by inhomogeneities due to treatment allocation. Given complex, potentially noisy, data, which model is to be most trusted to yield valid causal estimates? As no single learner performs best on all data sets, there is a pressing need for clear guidelines to select outcome models for causal inference.

Objectives and structure of the chapter In this chapter, we study *model selection procedures* in practical settings: *finite samples* settings and without *well-specification* assumption. Asymptotic causal-inference theory calls for complex risks, but a practical question is whether model-selection procedures, that rely on data split, can estimate these risks reliably enough. Indeed, they come with more quantities to estimate, which may bring additional variance, leading to worse model selection.

Fig. 5.2. Illustration
a) a random-forest estimator with high performance for standard prediction (high R^2) but that yields poor causal estimates (large error between true effect τ and estimated $\hat{\tau}$), b) a linear estimator with smaller prediction performance leading to better causal estimation.

Selecting the estimator with the smallest error to the individual treatment effect $\mathbb{E}[(\tau(x) - \hat{\tau}(x))^2]$ –the τ-risk, def. 1 – would lead to the best causal estimates; however computing this error is not feasible: it requires access to unknown quantities: $\tau(x)$.

While the random forest fits the data better than the linear model, it gives worse causal inference because its error is inhomogeneous between treated and untreated. The R^2 score does not capture this inhomogeneity.



We first illustrate the problem of causal model selection and briefly review prior art. Then, Section 5.2 sets causal model selection in the *potential outcome* framework and details the causal risks and model-selection procedure. Section 5.3 gives theoretical results. Section 5.4 details a thorough empirical study, covering many different settings. Finally, Section 5.5 discusses the findings. Results outline how to best select outcome models for causal inference with an adapted cross-validation to estimate the so-called $R-$risk. This risk compensates for systematic differences between treated and non-treated individuals using two *nuisance* models, themselves estimated from data and thus imperfect; yet these imperfections do not undermine the $R-$risk.

5.1.2 Illustration: the best predictor may not estimate best causal effects

Using a predictor to reason on causal effects relies on contrasting the prediction of the outcome for a given individual with and without the treatment –as detailed in section 5.2. Given various predictors of the outcome, which one should we use? Standard predictive modeling or machine-learning practice selects the predictor that minimizes the expected error. However, this predictor may not be the best model to reason about causal effects of an intervention, as we illustrate below.

Figure 5.2 gives a toy example: the probability Y of an undesirable outcome (*eg* death), a binary treatment $A \in \{0, 1\}$, and a covariate $X \in \mathbb{R}$ summarizing the patient health status (e.g., the Charlson index (Charlson et al., 1987)). We simulate a treatment beneficial (decreases Y) for patients with high Charlson scores (bad health status) but with little effect for patients in good condition (low Charlson scores).

Figure 5.2a shows a random forest predictor with a counter-intuitive behavior: it predicts well on average the outcome (as measured by a regression R^2 score) but perform poorly to estimate causal quantities: the average treatment effect τ (as visible via the error $|\tau - \hat{\tau}|$) or the conditional average treatment effect (the error $\mathbb{E}[(\tau(x) - \hat{\tau}(x))^2]$, called CATE). On the contrary, Figure 5.2b shows a linear model with smaller R^2 score but better causal inference.

The problem is that causal estimation requires controlling an error on both treated and non-treated outcome for the same individual: the observed outcome, and the non-observed *counterfactual* one. The linear model is misspecified –the outcome functions are not linear–, leading to poor R^2; but it interpolates better to regions where there are few untreated individuals –high Charlson score– and thus gives better causal estimates. Conversely, the random forest puts weaker assumptions on the data, thus has higher R^2 score but is biased by the treated population in the poor-overlap region, leading to bad causal estimates.

This toy example illustrates that the classic minimum Mean Square Error criterion is not suited to choosing a model among candidate estimators for causal inference.

5.1.3 Prior work: model selection for outcome modeling (g-computation)

A natural way to select a predictive model for causal inference would be an error measure between a causal quantity such as the CATE and models' estimate. But such error is not a "feasible" risk: it cannot be computed solely from observed data and requires oracle knowledge.

Simulation studies of causal model selection Using eight simulations setups from Powers et al., 2018, where the oracle CATE is known, Schuler et al. (2018) compare four causal risks, concluding that for CATE estimation the best model-selection risk is the so-called R-risk (Nie; Wager, 2017) –def. 6, below. Their empirical results are clear for randomized treatment allocation but less convincing for observational settings where both simple Mean Squared Error –MSE, μ-risk(f) def. 2– and reweighted MSE –μ-risk$_{IPW}$ def. 3– appear to perform better than R-risk on half of the simulations. Another work (Alaa; Schaar, 2019) studied empirically both MSE and reweighted MSE risks on the semi-synthetic ACIC 2016 datasets (Dorie et al., 2019), but did not include the R-risk. We complete these prior empirical work by studying a wider variety of data generative processes and varying the influence of overlap, an important parameter of the data generation process which makes a given causal metric appropriate (D'Amour et al., 2021). We also study how to best adapt cross-validation procedures to causal metrics which themselves come with models to estimate.

Theoretical studies of causal model selection Several theoretical works have proposed causal model selection procedures that are *consistent*: select the best model in a family given asymptotically large data. These works rely on introducing a CATE estimator in the testing procedure: matching (Rolling; Yang, 2014), an IPW estimate (Gutierrez; Gerardy, 2016), a doubly robust estimator (Saito; Yasui, 2020), or debiasing the error with influence functions (Alaa; Schaar, 2019). However, for theoretical guarantees to hold, the test-set correction needs to converge to the oracle: it needs to be flexible enough –well-posed– and asymptotic data. From a practical perspective, meeting such requirements implies having a good CATE estimate, thus having solved the original problem of causal model selection.

Statistical guarantees on causal estimation procedures Much work in causal inference has focused on procedures that guarantee asymptotically consistent estimators, such as

Targeted Machine Learning Estimation (TMLE) (Laan; Rose, 2011; Schuler; Rose, 2017) or Double Machine Learning (Chernozhukov et al., 2018a). Here also, theories require asymptotic regimes and models to be *well-specified*.

By contrast, Johansson et al. (2022) studies causal estimation without assuming that estimators are well specified. They derive an upper bound on the oracle error to the CATE (τ-risk) that involves the error on the outcome and the similarity of the distributions of treated and control patients. However, they use this upper bound for model optimization, and do not give insights on model selection. In addition, for hyperparameter selection, they rely on a plugin estimate of the τ-risk built with counterfactual nearest neighbors, which has been shown ineffective (Schuler et al., 2018).

5.2 Formal setting: causal inference and model selection

5.2.1 The Neyman-Rubin Potential Outcomes framework

Settings We consider the Potential Outcomes framework introduced in 1.4.1.

Causal assumptions Assumptions are necessary for causal estimands to be identifiability in observational settings (Rubin, 2005). We assume the usual strong ignorability assumptions: *1) unconfoundedness* $\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$, *2) strong overlap* ie. every patient has a strictly positive probability to receive each treatment, *3) consistency*, and *4) generalization* (introduced in 4.2.2).

Estimating treatment effects with outcome models Should we know the two expected outcomes for a given X, we could compute the difference between them, which gives the causal effect of the treatment. These two expected outcomes can be computed from the observed data: the consistency 3 and unconfoundedness 1 assumptions imply the equality of two different expectations:

$$\mathbb{E}_{Y(a)\sim\mathcal{D}^\star}[Y(a)|X=x] = \mathbb{E}_{Y\sim\mathcal{D}}[Y|X=x, A=a] \tag{5.1}$$

On the left, the expectation is taken on the counterfactual unobserved distribution. On the right, the expectation is taken on the factual observed distribution conditionally on the treatment. This equality is referred as the g-formula identification (Robins, 1986). For the rest of the paper, the expectations will always be taken on the factual observed distribution $\mathcal{D}$. This identification leads to outcome based estimators (ie. g-computation estimators (Snowden et al., 2011)), targeting the ATE τ with outcome modeling:

$$\tau = \mathbb{E}_{Y\sim\mathcal{D}^\star}[Y(1)-Y(0)|X=x] = \mathbb{E}_{Y\sim\mathcal{D}}[Y|A=1] - \mathbb{E}_{Y\sim\mathcal{D}}[Y|A=0] \tag{5.2}$$

This equation builds on two quantities: the conditional expectancy of the outcome given the covariates and either treatment or no treatment, called *response function*:

Response function $$\mu_a(x) \stackrel{\text{def}}{=} \mathbb{E}_{Y\sim\mathcal{D}}[Y|X=x, A=a]$$

Given a sample of data and the oracle response functions μ_0, μ_1, the finite sum version of equation 5.2 leads to an estimator of the ATE written:

$$\hat{\tau} = \frac{1}{n}\left(\sum_{i=1}^{n} \mu_1(x_i) - \mu_0(x_i)\right) \tag{5.3}$$

Table 5.1. Review of causal risks — The R-risk* is called τ-risk$_R$ in Schuler et al. (2018).

Risk	Equation	Reference
$mse(\tau(X), \tau_f(X)) = \tau$-risk	$\mathbb{E}_{X\sim p(X)}[(\tau(X) - \hat{\tau}_f(X))^2]$	Eq. 1 (Hill, 2011)
$mse(Y, f(X)) = \mu$-risk	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}[(Y - f(X;A))^2]$	Def. 2 (Schuler et al., 2018)
μ-risk$^*_{IPW}$	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(\frac{A}{e(X)} + \frac{1-A}{1-e(X)}\right)(Y - f(X;A))^2\right]$	Def. 3 (Laan et al., 2003)
τ-risk$^\star_{IPW}$	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(Y\left(\frac{A}{e(X)} - \frac{1-A}{1-e(X)}\right) - \hat{\tau}_f(X)\right)^2\right]$	Def. 4 (Wager; Athey, 2018)
U-risk*	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(\frac{Y-m(X)}{A-e(X)} - \hat{\tau}_f(X)\right)^2\right]$	Def. 5 (Nie; Wager, 2017)
R-risk*	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left((Y - m(X)) - (A - e(X))\hat{\tau}_f(X)\right)^2\right]$	Def. 6 (Nie; Wager, 2017)

This estimator is an oracle **finite sum estimator** by opposition to the population expression of τ, $\mathbb{E}[\mu_1(x_i) - \mu_0(x_i)]$, which involves an expectation taken on the full distribution $\mathcal{D}$, which is observable but requires infinite data. For each estimator ℓ taking an expectation over $\mathcal{D}$, we use the symbol $\hat{\ell}$ to note its finite sum version.

Similarly to the ATE, at the individual level, the CATE:

$$\tau(x) = \mu_1(x) - \mu_0(x) \tag{5.4}$$

Background Information

Robinson decomposition The *R-decomposition* of the outcome model plays an important role, (Robinson, 1988): introducing two quantities, the conditional mean outcome and the probability to be treated (known as propensity score (Rosenbaum; Rubin, 1983)):

$$\text{Conditional mean outcome} \qquad m(x) \overset{\text{def}}{=} \mathbb{E}_{Y\sim\mathcal{D}}[Y|X = x] \tag{5.5}$$

$$\text{Propensity score} \qquad e(x) \overset{\text{def}}{=} \mathbb{P}[A = 1|X = x] \tag{5.6}$$

the outcome can be written

$$\text{R-decomposition} \quad y(a) = m(x) + \Big(a - e(x)\Big)\tau(x) + \varepsilon(x;a) \quad \text{with} \quad \mathbb{E}[\varepsilon(X;A)|X,A] = 0 \tag{5.7}$$

m and e are often called *nuisances* (Chernozhukov et al., 2018a); they are unknown. ε is residual noise of mean zero.

5.2.2 Model-selection risks, oracle and feasible

Causal model selection We formalize model selection for causal estimation. Thanks to the g-formula identification (equation 5.1), a given outcome model $f : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ –learned from data or built from domain knowledge– induces feasible estimates of the ATE and CATE (eqs 5.3 and 5.4), $\hat{\tau}_f$ and $\hat{\tau}_f(x)$. Let $\mathcal{F} = \{f : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}\}$ be a family of such estimators. Our goal is to select the best candidate in this family for the observed dataset O using a risk ℓ:

$$f^*_\ell = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\, \ell(f, O) \tag{5.8}$$

We now detail possible risks ℓ, risks useful for causal model selection, and how to compute them.

The τ-risk: an oracle error risk As we would like to target the CATE, the following evaluation risk is natural:

Definition 1 (τ-risk(f)) *also called PEHE (Schulam; Saria, 2017; Hill, 2011):*

$$\tau\text{-}risk(f) = \mathbb{E}_{X\sim p(X)}[(\tau(X) - \hat{\tau}_f(X))^2]$$

Given observed data from $p(X)$, the expectation is computed with a finite sum, as in eq. 5.3, to give an estimated value $\widehat{\tau\text{-risk}}(f)$. However this risk is not feasible as the oracles $\tau(x)$ are not accessible with the observed data $(Y, X, A) \sim \mathcal{D}$.

Feasible error risks *Feasible* risks are based on the prediction error of the outcome model and *observable* quantities.

All expectations below are on observed distribution: $(Y, X, A) \sim \mathcal{D}$.

Definition 2 (Factual μ-risk) *(Shalit et al., 2017) This is the usual Mean Squared Error on the target y. It is what is typically meant by "generalization error" in supervised learning:*

$$\mu\text{-}risk(f) = \mathbb{E}\left[(Y - f(X; A))^2\right]$$

We now detail risks that use the nuisances e –propensity score, def 5.6– and m –conditional mean outcome, def 5.5. We give the definitions as *semi-oracles*, function of the true unknown nuisances, but later instantiate them with estimated nuisances, noted $\left(\check{e}, \check{m}\right)$. Semi-oracles risks are superscripted with the * symbol.

Definition 3 (μ-risk$^\star_{IPW}$) *(Laan et al., 2003) Let the inverse propensity weighting function $w(x, a) = \frac{a}{e(x)} + \frac{1-a}{1-e(x)}$, we define the semi-oracle Inverse Propensity Weighting risk,*

$$\mu\text{-}risk^\star_{IPW}(f) = \mathbb{E}\left[\left(\frac{A}{e(X)} + \frac{1-A}{1-e(X)}\right)(Y - f(X; A))^2\right]$$

Definition 4 (τ-risk$^\star_{IPW}$) *(Wager; Athey, 2018) The CATE $\tau(x)$ can be estimated with a regression against inverse propensity weighted outcomes (Athey; Imbens, 2016; Gutierrez; Gerardy, 2016; Wager; Athey, 2018), the τ-risk$_{IPW}$.*

$$\tau\text{-}risk^\star_{IPW}(f) = \mathbb{E}\left[\left(Y\frac{A - e(X)}{e(X)(1 - e(X))} - \tau_f(X)\right)^2\right]$$

Definition 5 (U-risk*) *(Künzel et al., 2019; Nie; Wager, 2017) Based on the Robinson decomposition –eq. 5.7, the U-learner uses the $A - e(X)$ term in the denominator. The derived risk is:*

$$U\text{-}risk^\star(f) = \mathbb{E}\left[\left(\frac{Y - m(X)}{A - e(X)} - \tau_f(X)\right)^2\right]$$

Note that extreme propensity weights in the denominator term might inflate errors in the numerator due to imperfect estimation of the mean outcome m.

Definition 6 (R-risk*) *(Nie; Wager, 2017; Schuler et al., 2018) The R-risk also uses two nuisances m and e:*

$$R\text{-}risk^\star(f) = \mathbb{E}\left[\Big((Y - m(X)) - (A - e(X))\,\tau_f(X)\Big)^2\right]$$

It is also based on the Robinson decomposition –eq. 5.7.

These risks are summarized in Table 5.1.

5.2.3 Estimation and model selection procedure

Causal model selection (as in equation 5.8) may involve estimating various quantities from the observed data: the outcome model f, its induced risk as introduce in the previous section, and possibly nuisances required by the risk. Given a dataset with N samples, we split out a train and a test sets $(\mathcal{T}, \mathcal{S})$. We fit each candidate estimator $f \in \mathcal{F}$ on $\mathcal{T}$. We also fit the nuisance models $(\check{e}, \check{m})$ on the train set $\mathcal{T}$, setting hyperparameters by a nested cross-validation before fitting the nuisance estimators with these parameters on the full train set. Causal quantities are then computed by applying the fitted candidates estimators $f \in \mathcal{F}$ on the test set $\mathcal{S}$. Finally, we compute the model-selection metrics for each candidate model on the test set. This procedure is described in Algorithm 1 and Figure 5.3.

As extreme inverse propensity weights induce high variance, clipping can be useful for numerical stability (Swaminathan; Joachims, 2015; Ionides, 2008).

Algorithm 1 Model selection procedure

Given train and test sets $(\mathcal{T}, \mathcal{S}) \sim \mathcal{D}$, a candidate estimators f, a causal metrics ℓ:

1. Prefit: Learn estimators for unknown nuisance quantities $(\check{e}, \check{m})$ on the training set $\mathcal{T}$
2. Fit: learn $\hat{f}(\cdot, a)$ on $\mathcal{T}$
3. Model selection:
   $\forall x \in \mathcal{S}$ predict $\left(\hat{f}(x,1), \hat{f}(x,0)\right)$ and evaluate the estimator storing the metric value:
   $\ell(f, \mathcal{S})$ – possibly function of $\check{e}$ and $\check{m}$

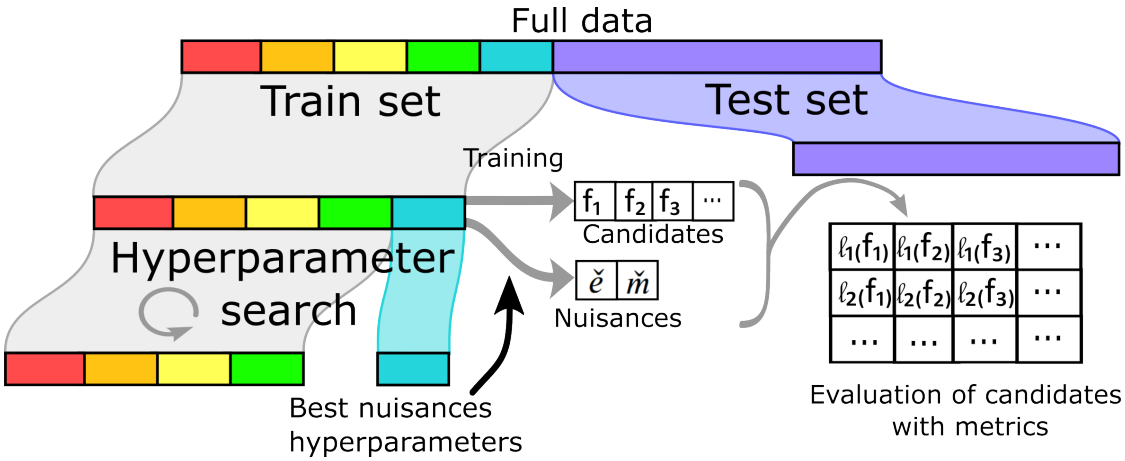


Fig. 5.3. Estimation procedure for causal model selection.

5.3 Theory: Links between feasible and oracle risks

We now relate two feasible risks, μ-risk$_{IPW}$ and the R-risk to the oracle τ-risk. Both results make explicit the role of overlap for the performance of causal risks.

These bounds depend on a specific form of residual that we now define: for each potential outcome, $a \in \{0; 1\}$, the variance conditionally on x is (Shalit et al., 2017):

$$\sigma_y^2(x; a) \stackrel{\text{def}}{=} \int_y (y - \mu_a(x))^2 \, p(y \mid x = x; A = a) \, dy$$

Integrating over the population, we get the Bayes squared error: $\sigma_B^2(a) = \int_{\mathcal{X}} \sigma_y^2(x;a)p(x)dx$ and its propensity weighted version: $\tilde{\sigma}_B^2(a) = \int_{\mathcal{X}} \sigma_y^2(x;a)\, p(x;a)\, dx$. In case of a purely deterministic link between the covariates, the treatment, and the outcome, these residual terms are null.

5.3.1 Upper bound of τ-risk with μ-risk$_{IPW}$

Proposition 1 (Upper bound with μ-risk$_{IPW}$) *(Johansson et al., 2022) Given an outcome model f, let a weighting function $w(x;a) = \frac{a}{e(x)} + \frac{1-a}{1-e(x)}$ as the Inverse Propensity Weight. Then, under overlap (assumption 2), we have:*

$$\tau\text{-risk}(f) \leq \; 2\,\mu\text{-risk}_{IPW}(w,f) - 2\left(\sigma_B^2(1) + \sigma_B^2(0)\right)$$

This result has been derived in previous work (Johansson et al., 2022). It links μ-risk$_{IPW}$ to the squared residuals of each population. For completeness, we provide the proof in E.2.

The upper-bound comes from the triangular inequality applied to the residuals of both populations. The two quantities are equal when the absolute residuals on treated and untreated populations are equal on the whole covariate space: $\forall x \in \mathcal{X}, |\mu_1(x) - f(x,1)| = |\mu_0(x) - f(x,0)|$. The main difference between the oracle τ-risk and the reweighted mean squared error, μ-risk$_{IPW}$, comes from heterogeneous residuals between populations. This bound shows that minimizing the μ-risk$_{IPW}$ helps to minimize the τ-risk, which leads to interesting optimization procedures (Johansson et al., 2022). However, there is no guarantee that this bound is tight, which makes it fragile for model selection.

Assuming strict overlap (probability of all individuals being treated or not bounded away from 0 and 1 by η, 4.2.2), the above bound simplifies into a looser one involving the usual mean squared error: $\tau\text{-risk}(f) \leq \frac{2}{\eta}\,\mu\text{-risk}(f) - 2\left(\sigma_B^2(1) + \sigma_B^2(0)\right)$. For weak overlap (propensity scores not bounded far from 0 or 1), this bound is very loose (as shown in Figure 5.2) and is not appropriate to discriminate between models with close performance.

5.3.2 Reformulation of the R-risk as reweighted τ-risk

We now derive a novel rewriting of the R-risk, making explicit its link with the oracle τ-risk.

Proposition 2 (R-risk as reweighted τ-risk) *Given an outcome model f, its R-risk appears as weighted version of its τ-risk (Proof in E.2.2):*

$$R\text{-risk}^*(f) = \int_x e(x)\Big(1-e(x)\Big)\Big(\tau(x) - \tau_f(x)\Big)^2 p(x)dx \qquad + \tilde{\sigma}_B^2(1) + \tilde{\sigma}_B^2(0) \tag{5.9}$$

The R-risk targets the oracle at the cost of an overlap re-weighting and the addition of the reweighted Bayes residuals, which are independent of f. In good overlap regions the weights $e(x)\Big(1-e(x)\Big)$ are close to $\frac{1}{4}$, hence the R-risk is close to the desired gold-standard τ-risk. On the contrary, for units with extreme overlap violation, these weights go down to zero with the propensity score.

5.3.3 Interesting special cases

Randomization special case If the treatment is randomized as in RCTs, $p(A=1 \mid X = x) = p(A=1) = p_A$, thus μ-risk$_{IPW}$ takes a simpler form:

$$\mu\text{-risk}_{IPW} = \mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(\frac{A}{p_A} + \frac{1-A}{1-p_A}\right)(Y - f(X;A))^2\right]$$

However, we still can have large differences between τ-risk and μ-risk$_{IPW}$ coming from heterogeneous errors between populations as noted in Section 5.3.1 and shown experimentally in Schuler et al. (2018) and our results below.

Concerning the R-risk, replacing $e(x)$ by its randomized value p_A in Proposition 2 yields the oracle τ-risk up to multiplicative and additive constants:

$$R\text{-risk} = p_A\,(1-p_A)\,\tau\text{-risk} \;+\; (1-p_A)\,\sigma_B^2(0) \;+\; p_A\sigma_B^2(1)$$

Thus, selecting estimators with R-risk* in randomized setting controls the τ-risk. This explains the strong performance of R-risk in randomized setups (Schuler et al., 2018) and is a strong argument to use it to estimate heterogeneity in RCTs.

Oracle Bayes predictor If we have access to the oracle Bayes predictor for the outcome ie. $f(x,a) = \mu(x,a)$, then all risks are equivalent up to the residual variance:

$$\tau\text{-risk}(\mu) = \mathbb{E}_{X\sim p(X)}[(\tau(X) - \tau_\mu(X))^2] = 0 \tag{5.10}$$

$$\mu\text{-risk}(\mu) = \mathbb{E}_{(Y,X,A)\sim p(Y;X;A)}[\Big(Y - \mu_A(X)\Big)^2] \tag{5.11}$$

$$= \int_{\mathcal{X},\mathcal{A}} \varepsilon(x,a)^2 p(a \mid x)\, p(x)\, dx\, da \leq \sigma_B^2(0) + \sigma_B^2(1) \tag{5.12}$$

$$\mu\text{-risk}_{IPW}(\mu) = \sigma_B^2(0) + \sigma_B^2(1) \quad \text{from Lemma 1} \tag{5.13}$$

$$R\text{-risk}(\mu) = \tilde{\sigma}_B^2(0) + \tilde{\sigma}_B^2(1) \leq \sigma_B^2(0) + \sigma_B^2(1) \quad \text{from Proposition 2} \tag{5.14}$$

Thus, differences between causal risks only matter in finite sample regimes. Universally consistent learners converge to the Bayes risk in asymptotic regimes, making all model selection risks equivalent. In practice however, choices must be made in non-asymptotic regimes.

5.4 Empirical Study

We evaluate the following causal metrics, oracle and feasible versions, presented in Table 5.1: $\widehat{\mu\text{-risk}}^*_{IPW}$, $\widehat{R\text{-risk}}^*$, $\widehat{U\text{-risk}}^*$, $\widehat{\tau\text{-risk}_{IPW}}^*$, $\widehat{\mu\text{-risk}}$, $\widehat{\mu\text{-risk}}_{IPW}$, $\widehat{R\text{-risk}}$, $\widehat{U\text{-risk}}$, $\widehat{\tau\text{-risk}_{IPW}}$. We benchmark the metrics in a variety of settings: many different simulated data generation processes and three semi-simulated datasets [1].

5.4.1 Caussim: Extensive simulation settings

Data Generation We use simulated data, on which the ground-truth causal effect is known. Going beyond prior empirical studies of causal model selection (Schuler et al., 2018; Alaa; Schaar, 2019), we use many generative processes, to reach more general conclusions (as discussed in E.12).

We generate the response functions using random bases. Basis extension methods are common in biostatistics, *eg* functional regression with splines (Howe et al., 2011; Perperoglou et al., 2019). By allowing the function to vary at specific knots, they give flexible non-linear models. We use random approximation of Radial Basis Function (RBF) kernels (Rahimi;

[1] Scripts for the simulations and the selection procedure are available at https://github.com/soda-inria/caussim.

Recht, 2008) to generate the response functions. RBF use the same process as polynomial splines but replace polynomial by Gaussian kernels. Unlike polynomials, Gaussian kernels have decreasing influences in the input space. This avoids unrealistic divergences of the response surfaces at the ends of the feature space.

The number of basis functions –*ie. knots*–, controls the complexity of the ground-truth response surfaces and treatment. We first use this process to draw the non-treated response surface μ_0 and the causal effect τ. We then draw the observations from a mixture two Gaussians, for the treated and non treated. We vary the separation between the two Gaussians to control the overlap between treated and non-treated populations, an important parameter for causal inference (related to η in section 5.3.1). Finally, we generate observed outcomes adding Gaussian noise yielding a dataset as plotted in Figure 5.4. We generate 1 000 of such datasets, with uniformly random overlap parameters. Details in E.4.1.

Family of candidate estimators We test model selection on a family of candidate estimators that approximate imperfectly the data-generating process. To build such an estimator, we first use a RBF expansion similar to the one used for data generation. We choose two random knots and apply a transformation of the raw data features with a Gaussian kernel. This step is referred as the featurization. Then, we fit a linear regression on these transformed features. We consider two ways of combining these steps for outcome mode; we use common nomenclature (Künzel et al., 2019; Shen et al., 2023) to refer to these different meta-learners that differ on how they model, jointly or not, the treated and the non treated:

- SLearner: A single learner for both population, taking the treatment as a supplementary covariate.
- SftLearner: A single set of basis functions is sampled at random for both populations, leading to a given feature space used to model both the treat and the non treated, then two separate different regressors are fitted on this shared representation.
- TLearner: Two completely different learners for each population, hence separate feature representations and regressors.

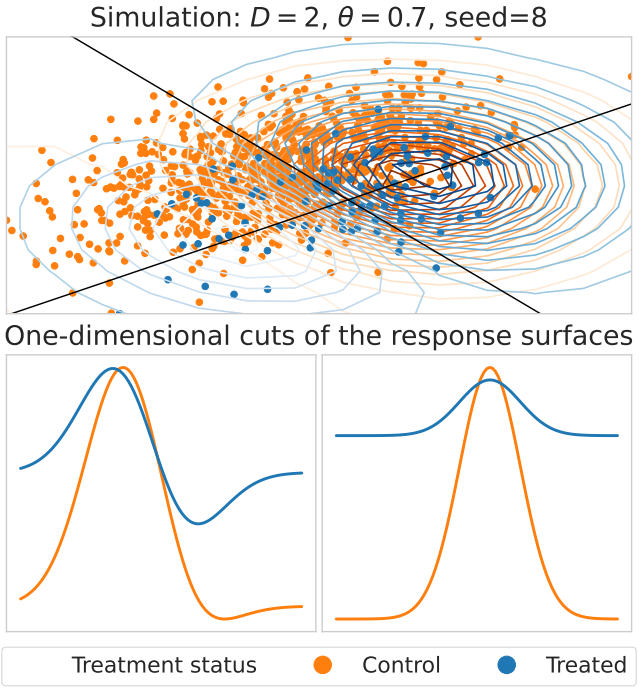


Fig. 5.4. Example of the simulation setup in the input space with two knots –*ie.*basis functions. The top panel shows the observations in feature space, while the bottom panel displays the two response surfaces on a 1D cut along the black lines drawn on the top panel.

We do not include more elaborated meta-learners such as R-learner (Nie; Wager, 2017) or X-learner (Künzel et al., 2019). Our goal is not to have the best possible learner but to have a variety of sub-optimal learners to compare the different causal metrics. For the same reason, we did not include more powerful outcome models such as random forests or boosting trees.

For the regression step, we fit a Ridge regression on the transformed features with 6 different choices of the regularization parameter $\lambda \in [10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}]$, coupled with a TLearner or a SftLearner. We sample 10 different random basis for learning and featurization yielding a family $\mathcal{F}$ of 120 candidate estimators.

5.4.2 Semi-simulated datasets

Datasets We also use classic benchmarks of the causal-inference literature, semi-simulated data adding a known synthetic causal effect to real –non synthetic– covariate:

ACIC 2016 (Dorie et al., 2019): The dataset is based on the Collaborative Perinatal Project (Niswander; Stroke, 1972), a RCT studying infants' developmental disorders. The initial intervention was a child's birth weight ($A = 1$ if weight $< 2.5kg$), and outcome was the child's IQ after a follow-up period. The study contained $N = 4\,802$ data points with $D = 55$ features (5 binary, 27 count data, and 23 continuous). They simulated 77 different setups varying parameters for treatment and response models, overlap, and interactions between treatment and covariates [2]. We used 10 different seeds for each setup, totaling 770 dataset instances.

ACIC 2018 (Shimoni et al., 2018): Starting from data from the Linked Births and Infant Deaths Database (LBIDD) (MacDorman; Atkinson, 1998) with $D = 177$ covariates, treatment and outcome models are simulated with complex models to reflect different scenarios. The data do not provide the true propensity scores, so we evaluate only feasible metrics, which do not require this nuisance parameter. We used all 432 datasets[3] of size $N = 5\,000$.

Twins (Louizos et al., 2017): It is an augmentation of real data on twin births and mortality rates (Almond et al., 2005). There are $N = 11\,984$ samples (pairs of twins), and $D = 50$ covariates[4], The outcome is the mortality and the treatment is the weight of the heavier twin at birth. This is a "true" counterfactual dataset (Curth et al., 2021) in the sense that we have both potential outcomes with each twin. They simulate the treatment with a sigmoid model based on GESTAT10 (number of gestation weeks before birth) and x the 45 other covariates:

$$\mathbf{t}_i \mid \mathbf{x}_i, \mathbf{z}_i \sim \text{Bern}\left(\sigma\left(w_o^\top \mathbf{x} + w_h(\mathbf{z}/10 - 0.1)\right)\right) \quad (5.15)$$
$$\text{with } w_o \sim \mathcal{N}(0, 0.1 \cdot I),\ w_h \sim \mathcal{N}(5, 0.1)$$

We add a non-constant slope in the sigmoid to control the overlap between treated and control populations. We sampled uniformly 1 000 different overlap parameters between 0 and 2.5, totaling 1 000 dataset instances. Unlike the previous datasets, only the overlap varies for these instances. The response surfaces are set by the original outcomes.

[2] Original R code available at https://github.com/vdorie/aciccomp/tree/master/2016 to generate 77 simulations settings.

[3] Using the scaling part of the data, from github.com/IBM-HRL-MLHLS/IBM-Causal-Inference-Benchmarking-Framework

[4] We obtained the dataset from https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/TWINS

Family of candidate estimators For these three datasets, the family of candidate estimators are gradient boosting trees for both the response surfaces and the treatment[5] with S-learner, learning rate in $\{0.01, 0.1, 1\}$, and maximum number of leaf nodes in $\{25, 27, 30, 32, 35, 40\}$ resulting in a family of size 18.

Nuisance estimators Drawing inspiration from the TMLE literature that uses combination of flexible machine learning methods (Schuler; Rose, 2017), we use as models for the nuisances $\check{e}$ (respectively $\check{m}$) a form of meta-learner: a stacked estimator of ridge and boosting classifiers (respectively regressions). We select hyper-parameters with randomized search on a validation set $\mathcal{V}$ and keep them fixed for model selection (E.4.2 lists hyperparameters). As extreme inverse propensity weights induce high variance, we use clipping (Swaminathan; Joachims, 2015; Ionides, 2008) to bound $min(\check{e}, 1-\check{e})$ away from 0 with a fixed $\eta = 10^{-10}$, ensuring strict overlap for numerical stability.

5.4.3 Measuring overlap between treated and non treated

Good overlap between treated and control population is crucial for causal inference as it is required by the positivity assumption 2. It is often assessed by comparing visually population distributions (as in Figure 5.2) or computing standardized difference on each feature (Austin, 2011; Austin; Stuart, 2015). While these methods are useful to decide if positivity holds, they do not yield a single measure. Rather, we compute the divergence between the population covariate distributions $\mathbb{P}(X|A=0)$ and $\mathbb{P}(X|A=1)$ (D'Amour et al., 2021; Johansson et al., 2022). We introduce the Normalized Total Variation (NTV), a divergence based on the sole propensity score (see E.3).

5.4.4 Results: factors driving good model selection

The R-risk is the best metric Each metric ranks differently the candidate models. Figure 5.5 shows the agreement between the ideal ranking of methods given the oracle τ-risk and the different feasible causal metrics. We measure this agreement with a relative[6] Kendall tau κ (eq. E.4) (Kendall, 1938). Given the importance of overlap in how well metrics approximate the oracle τ-risk (E.2.1), we separate strong and weak overlap.

Among all metrics, the classical mean squared error (ie. factual μ-risk) is worse and reweighting it with propensity score (μ-risk$_{IPW}$) does not bring much improvement. The R-risk, which includes a model of mean outcome and propensity scores, leads to the best performance. Interestingly, the U-risk, which uses the same nuisances, deteriorates in weak overlap, probably due to variance inflation when dividing by extreme propensity scores.

Beyond rankings, the differences in terms of absolute ability to select the best model are large: The R-risk selects a model with a τ-risk only 1% higher than the best possible candidate for strong overlap on Caussim, but selecting with the μ-risk or μ-risk$_{IPW}$ –as per machine-learning practice– leads to 10% excess risk and using τ-risk$_{IPW}$ –as in some causal-inference methods (Athey; Imbens, 2016; Gutierrez; Gerardy, 2016)–leads to 100% excess risk (Figure E.7). Across datasets, the R-risk consistently decreases the risk compared to the μ-risk: 0.1% compared to 1% on ACIC2016, 1% compared to 20% on ACIC2018, and 0.05% compared to 1% on Twins.

[5] Scikit-learn regressor, HistGradientBoostingRegressor, and classifier, HistGradientBoostingClassifier.

[6] To remove the variance across datasets (some datasets lead to easier model selection than others), we report values for one metric relative to the mean of all metrics for a given dataset instance: Relative $\kappa(\ell, \tau\text{−risk}) = \kappa(\ell, \tau\text{−risk}) - mean_{\ell}\big(\kappa(\ell, \tau\text{−risk})\big)$

Model selection is harder for low population overlap Model selection for causal inference becomes more and more difficult with increasingly different treated and control populations (Figure 5.6). The absolute Kendall's coefficient correlation with τ-risk drops from values around 0.9 (excellent agreement with oracle selection) to 0.6 on both Caussim and ACIC 2018 (E.4.3).

Nuisances can be estimated on the same data as outcome models Using the train set $\mathcal{T}$ both to fit the candidate estimator and the nuisance estimates is a form of double dipping which can lead to errors in nuisances correlated to that of outcome models (Nie; Wager, 2017). In theory, these correlations can bias model selection and, strictly speaking, push to split out a third separated data set –a "nuisance set"– to fit the nuisance models. The drawback is that it depletes the data available for model estimation and selection. However, Figure 5.7 shows no substantial difference between a procedure with a separated nuisance set and the simpler shared nuisance-candidate set procedure.

Stacked models are good overall estimators of nuisances For every risk, the oracle version recovers better the best estimator. However, stacked nuisances estimators (boosting and linear) lead to feasible metrics with close performance to the oracles ones: the corresponding estimators recover well-enough the true nuisances. One may wonder if simpler models for the nuisance could be useful, in particular in data-poor settings or when the true

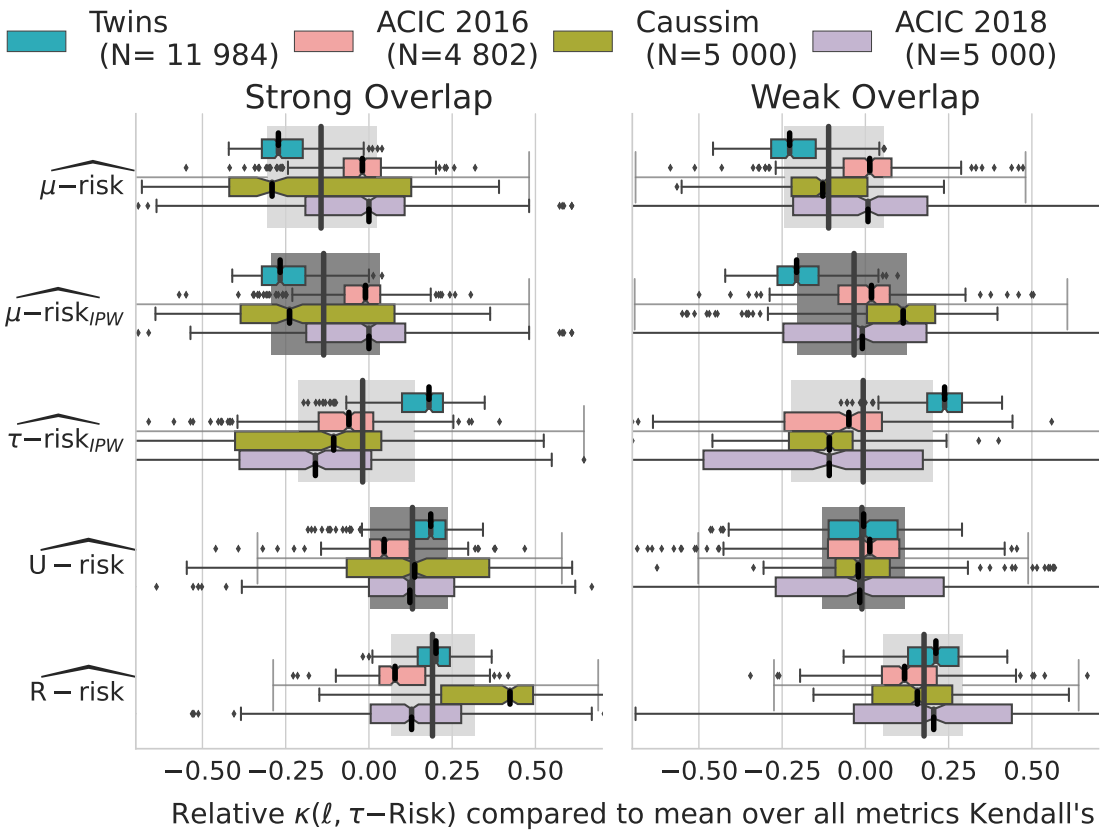


Fig. 5.5. The R-risk is the best metric: Relative Kendall's τ agreement with τ-risk. Strong and Weak overlap correspond to the first and last tertiles of the overlap distribution measured with Normalized Total Variation eq. E.2. E.4.3 presents the same results by adding semi-oracle risks in Figure E.5, measured with absolute Kendall's in Figure E.6 and with $\tau-$risk gains in Figure E.7. Table E.3 gives median and IQR of the relative Kendall.

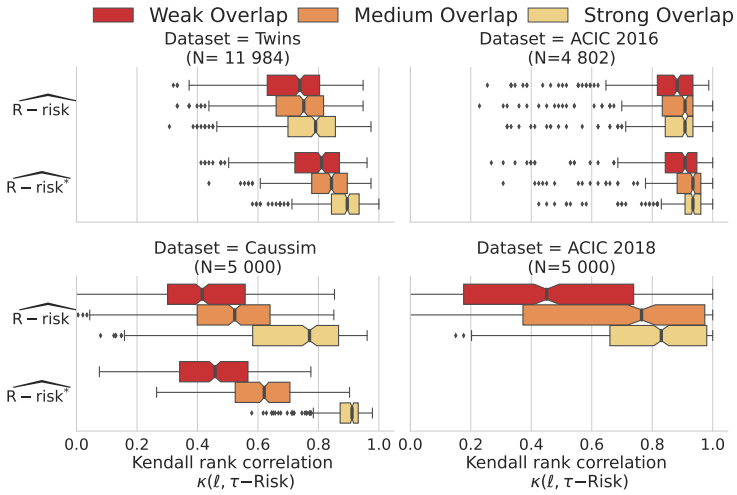


Fig. 5.6. Model selection is harder for low population overlap: Kendall's τ agreement with τ-risk. Strong, medium and Weak overlap are the tertiles of the overlap measured with NTV eq. E.2. E.4.3 presents results for all metrics in Figure E.9 in absolute Kendall's and continuous overlap values in Figure E.6.

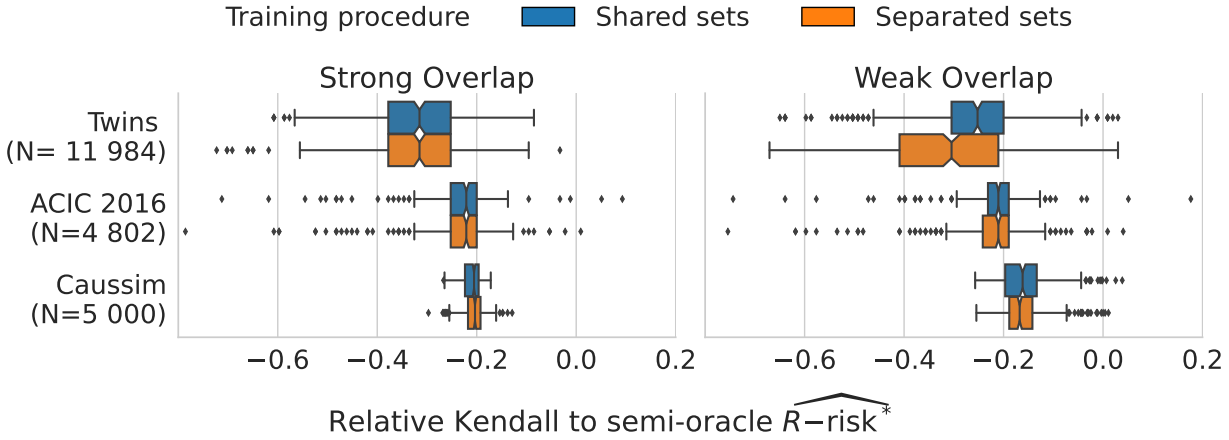


Fig. 5.7. Nuisances can be estimated on the same data as outcome models: Results for the R-risk are similar between the shared nuisances/candidate set and the separated nuisances set procedures. Figure E.8 details results for all metrics.

models are linear. Figure 5.8 compares causal model selection estimating nuisances with stacked estimators or linear model. It comprises the Twins data, where the true propensity model is linear, and a downsampled version of this data, to study a situation favorable to linear models. In these settings, stacked and linear estimations of the nuisances performs equivalently. Detailed analysis (Figure E.11) confirms that using adaptive models –as built by stacking linear models and gradient-boosted trees– suffices to estimate nuisance.

Use 90% of the data to estimate outcome models, 10% to select them The analyst faces a compromise: given a finite data sample, should she allocate more data to estimate the outcome model, thus improving the quality of the outcome model but leaving little data for model selection. Or, she could choose a bigger test set for model selection and effect estimation. For causal model selection, there is no established practice (as reviewed in E.5).

We investigate such tradeoff varying the ratio between train and test data size. For this, we first split out 30% of the data as a holdout set $\mathcal{V}$ on which we use the oracle response functions to derive silver-standard estimates of causal quantities. We then use the standard estimation procedure on the remaining 70% of the data, splitting it into train $\mathcal{T}$ and test $\mathcal{S}$ of varying sizes. We finally measure the error between this estimate and the silver standard.

We consider two different analytic goals: estimating a average treatment effect –a single number used for policy making– and a CATE –a full model of the treatment effect as a

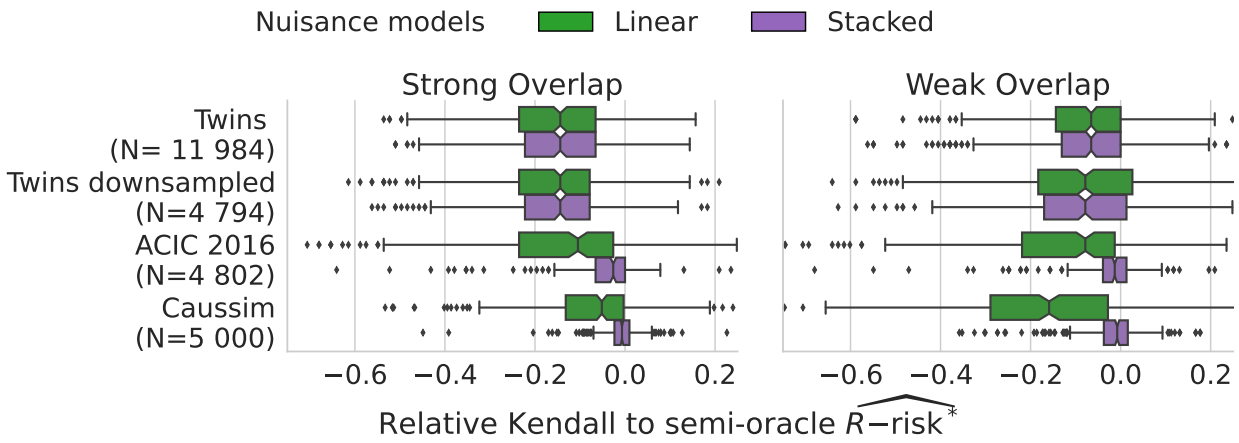


Fig. 5.8. Stacked models are good overall estimators of the nuisances: Results are shown only for the R-risk; Figure E.10 details every metrics. For Twins, where the true propensity model is linear, stacked and linear estimations of the nuisances performs equivalently, even for a downsampled version (N=4794).

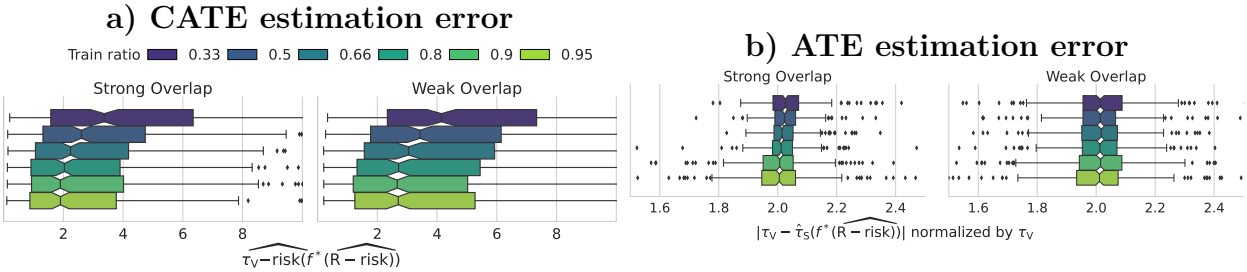


Fig. 5.9. a) For CATE, a train/test ratio of 0.9/0.1 appears a good trade-off. b) For ATE, there is a small signal pointing also to 0.9/0.1 (K=10). for ATE. Experiences on 10 replications of all 78 instances of the ACIC 2016 data.

function of covariates X. Given that the latter is a much more complex object than the former, the optimal train/test ratio might vary. To measure errors, we use for the ATE the relative absolute ATE bias between the ATE computed with the selected outcome model on the test set, and the true ATE as evaluated on the holdout set $\mathcal{V}$. For the CATE, we compare the τ-risk of the best selected model applied on the holdout set $\mathcal{V}$. We explore this trade-off for the ACIC 2016 dataset and the R-risk.

Figure 5.9 shows that a train/test ratio of 0.9/0.1 (K=10) or 0.8/0.2 (K=5) appears best to estimate CATE and ATE.

5.5 Discussion and conclusion

Predictive models are increasingly used to reason about causal effects, for instance in precision medicine to drive individualized decision. Our results highlight that they should be selected, validated, and tuned using different procedures and error measures than those classically used to assess prediction (estimating the so-called μ-risk). Rather, selecting the best outcome model according to the R-risk (eq. 6) leads to more valid causal estimates.

Nuisance models – More gain than pain Estimating the R-risk requires a more complex procedure than standard cross-validation used *e.g.,* in machine learning: it involves

fitting nuisance models necessary for model evaluation, though our results show that these can be learned on the same set of data as the outcome model evaluated.

The nuisance models must be well estimated (Figure 5.8). However these models are easier to select and control than a causally-valid outcome model, as they are associated to errors on observed distributions. Our results show that using for nuisance models a flexible stacking-based family of estimator suffices for good model selection. In fact, a feasible R-risk –where the nuisances are estimated– performs almost as well as an oracle R-risk –where the nuisances are known. This may be explained by results that suggest that estimation errors on both nuisances partly compensate out in the R-risk (Daniel, 2018; Kennedy, 2020; Nie; Wager, 2017; Chernozhukov et al., 2018a; Zivich; Breskin, 2021; Naimi et al., 2021).

Note that propensity score models must be selected to estimate the individual posterior probability. For this, we used the Brier score, which is minimized by the true individual probability. An easy mistake is to use calibration errors popular in machine learning (Platt; Platt, 1999; Zadrozny; Elkan, 2001; Niculescu-Mizil; Caruana, 2005; Minderer et al., 2021) as these select not for the individual posterior probability but for an aggregate error rate (Perez-Lebel et al., 2022).

Extension to binary outcomes While we focused on continuous outcomes, in medicine, the target outcome is often a categorical variable such as mortality status or diagnosis. In this case, it may be interesting to focus on other estimands than the Average Treatment Effect $\mathbb{E}[Y(1)]-\mathbb{E}[Y(0)]$, for instance the relative risk $\frac{\mathbb{P}(Y(1)=1)}{\mathbb{P}(Y(0)=1)}$ or the odd ratio, $\frac{\mathbb{P}(Y(1)=1)/[1-\mathbb{P}(Y(1)=1)]}{\mathbb{P}(Y(0)=1)/[1-\mathbb{P}(Y(0)=1]}$ are often used (Austin; Stuart, 2017). While the odds ratio is natural for case-control studies (Rothman et al., 2008), other measures can reduce heterogeneity (Colnet et al., 2023). In the log domain, the ratios are written as a difference, the framework studied here (section 5.2) can directly apply. In particular, the log odds ratio is estimated by the common cross-entropy loss (or log loss) as in logistic regression.

More R-risk to select models driving decisions Prediction models have flourished because their predictions can be easily demonstrated and validated on left-out data. But they require more careful validation for decision making, using a metric accounting for the putative intervention, the R-risk. Even when treated and untreated population differ little, as in RCTs, the R-risk brings a sizeable benefit. To facilitate better model selection, we provide Python code [7]. Using the R-risk does make evaluation more complicated not only because the procedure is more involved, but also because each intervention requires a dedicated evaluation. However, such off-policy evaluation remains much less costly than the recommended good practice of impact evaluation testing the ability of a prediction model to actually guide patient health (Hendriksen et al., 2013). Also, the model-selection procedure puts no constraints on the models used to build predictive models: it opens the door to evaluating a wide range of models, from gradient boosting to convolutional neutral, or language models.

[7] https://github.com/soda-inria/causal_model_selection

Chapter 6

Conclusion

6.1 Lessons learned

Considerable efforts are being made to prepare routine care data for reuse A vast amount of new data is being collected in healthcare, at the cost of new infrastructures: the Clinical Data Warehouses (CDWs). Navigating this mass of data is simplified, mainly by centralizing fragmented data from poor interoperability and the use of NLP techniques. The main aim of these CDWs is to improve clinical research. However, deriving new evidence from EHRs is currently hindered by a poorly standardized data access processes, heterogeneous local care practices, and the non-exhaustive collection of critical elements of care trajectories, even during hospital care.

These data repositories are not big enough for large neural networks trained on structured data Even the regional scale data center of the AP-HP with data on 10 millions individuals has only a few thousands of cases for common pathologies such as cardiovascular adverse events. On these medium sized datasets, flexible models on top of simple data representations –such as random forest with event counts– are performing better than transformer-based models while being simpler to deploy and more compute efficient. These results echo the recent benchmark from Grinsztajn et al., 2022 showing the superiority of tree-based methods over deep learning on tabular data.

Big data is no oracle, we need causal thinking Even with sufficient data, the observational nature of EHRs requires clear and robust workflows inspired by modern causal inference to derive unbiased intervention effects. Using flexible predictive models without a causal framework opens the door to collider biases, but using only linear models is prone to underfitting and biased estimates as well. Vibration analyses are key to better spot bias in observational studies.

Selecting outcome models for heterogeneous treatment effects benefits from flexible estimations of nuisances Causal model selection should not been performed with mean squared error as it is done in predictive modeling. Estimating nuisance parameters of the doubly robust R-risk thanks to flexible models is the most performant approach. In practice, a simple procedure where nuisances are estimated with the same train set as the candidate models introduces negligible bias.

6.2 Personal thoughts on perspectives

The unreasonable effectiveness of healthcare data is still out of reach Due to the impossibility to transfer models, administrative barriers to access data (linked with the multiplicity of the involved actors), and the difficulty to normalize healthcare data, predictive models are very seldom deployed (Kelly et al., 2019). This encourages us to rely

on sample efficient techniques that make the best of medium-sized data or rely on sharable sources of knowledge. This calls for more work adopting research network methodologies (Hripcsak et al., 2015a) and model sharing strategies such as federated learning (Rieke et al., 2020). Another interesting avenue is to leverage large language models pretrained on natural text to either extract information from clinical notes, or directly to build predictive models (Jiang et al., 2023). These last perspectives require having sufficient computing resources for finetuning the language models.

For treatment effect estimations, relying on text calls for better understanding of causal representation learning Using text for confounder adjustment in causal studies opens many research questions. Building a causal graph as in Figure 4.6 becomes tedious. Even if all information is present in the text –satisfying the ignorability assumption 1, should we extract every confounder or is it possible to build appropriate representations? Optimizing directly the R-loss by gradient descent could be an interesting avenue, which is closely related to the work of Johansson et al., 2022 and Chernozhukov et al., 2022 with non-text data. Vibration analysis is needed to confront these methods to applications and known effects such as the one presented in chapter 4.

Claims data are an interesting source for studying treatment effectiveness and public health policy Some difficulty to define precise populations make claims inappropriate for precise clinical questions, but they are a good source for many public health and medico-economic questions. They are well standardized, well documented, and contain most of the PICO elements needed to study chronic conditions (Caruana et al., 2023). Estimations methods borrowed to econometrics such as difference-in-difference (Athey; Imbens, 2006), regression discontinuity design (Bor et al., 2014) and instrumental variables (Greenland, 2000) might be also considered when appropriate for the question of interest.

Modern public health problems should account for the fragmented nature of healthcare Healthcare burdens and costs are driven by chronic diseases where death or rehospitalization are not the only outcomes of interest. Most of care is provided outside of the hospital (White et al., 1961), therefore requiring to link city care, and socio-economic features to hospital data. This would allow to study less critical outcomes, for which experiments are easier to conduct and where error is more acceptable. Adopting methods from policy learning (Athey; Wager, 2021) would allow better evaluation and continuous improvement for interventions, opening the door to potentially big prevention benefits. Such public health interventions should be conducted within clear ethical frameworks. They could draw inspiration from existing experimental processes such as the French Article 51 or the United States center for medicare and medicaid innovation, both focused on organisational innovations (Lenormand; Panteli, 2021). Ultimately, I am convinced that the resource constraints faced by modern healthcare systems link the issue care effectiveness to that of equity. Some patients will benefit more from a given intervention than others. Public health seeks to minimize these mistreated patients, hopefully without discrimination. More observational studies could help identify which subpopulations benefit the most from effective interventions.

Appendices

Appendix A

Chapter 1

A.1 Statistical learning theory

I recall here the formal definition of the regression problem in statistical learning and I refer to Hastie et al., 2009 for the classification problem.

Given n pairs of (features, outcome) noted $(x, y) \in \mathcal{X} \times \mathcal{Y}$ identically and independently distributed (i.i.d) , the goal is to find a function $f : \mathcal{X} \to \mathcal{Y}$ that approximates the true value of y ie. $f(x) \approx y$. One need to define a loss function ℓ that define proximity between the predicted value $\hat{y} = f(x)$ and the true value y. Usually, for continuous outcomes, the squared loss is used. Finally, when choosing among a family of functions $f \in \mathcal{F}$, the best possible function $f^\star$ minimizes the expected loss $\mathcal{E}(f^\star)$:

$$f^\star = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}\Big[(\hat{y} - y)^2\Big] \tag{A.1}$$

In finite sample regimes, the expectation is not accessible since we only have access to a finite number of data pairs $(x_i, y_i)_{i=1..n}$. So in practice, we aim to minimize the empirical loss $\mathcal{E}(f^\star)$::

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} \Big[(\hat{y}_i - y_i)^2\Big] \tag{A.2}$$

In most interesting problems, there is some randomness involved in the (x, y) association, either due to inherent randomness or to the lack of important information in x. This is modeled by assuming it independent of the features and with mean zero: $y = g(x) + e$, with $\mathbb{E}[e] = 0$. The best possible estimator is thus g, yielding the Bayes error $\mathcal{E}(g) = \mathbb{E}\Big[(g(x) + e - g(x))^2\Big] = \mathbb{E}[e^2]$. This error cannot be avoided.

Finally, the generalization error of the estimator $\hat{f}$ can be decomposed as:

$$\mathcal{E}(\hat{f}) = \mathcal{E}(g) + (\mathcal{E}(f^\star) - \mathcal{E}(g)) + (\mathcal{E}(\hat{f}) - \mathcal{E}(f^\star)) \tag{A.3}$$

The second term is the approximation error: the difference between the best estimator in the family of estimator considered and the Bayes estimator. It decreases for larger $\mathcal{F}$.

The third term is the estimation error related to the sampling noise of the data. It decreases with raising n –if we have a lot of data points. It increases for larger $\mathcal{F}$. This decomposition highlights the choice of a practitioner for applying regression: she must choose a function class $\mathcal{F}$ flexible enough to avoid underfitting the data but restrictive enough to avoid a high estimation error.

A.2 Statistical models

We recall briefly how trees, random forests and boosting work. For more details, see (Hastie et al., 2009).

A.2.1 Trees

Decision trees are a class of models that recursively split the feature space into a set of rectangles –called nodes, and assign a constant value to each rectangle. They can be use both for classification or regression. If used for regression, a new split from one region into two subregions is chosen as follows. Among all variable and split possibility, choose the split that minimizes the error –typically the squared error– between the average of the outcomes over the two newly regions and the individual outcomes. If used for classification, splits are chosen by minimizing some impurity measure of the class probabilities over the two new created nodes instead of the squared error. A typical impurity measure is the Gini index $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$ where $\hat{p}_{mk}$ is the empirical probability of class k in node m. The complexity of a tree is determined by its maximal depth –ie, the maximum number of splits before reaching a terminal node –called a leaf. Figure A.1 illustrates on a toy dataset, decision trees with depth 1, 2 and 5.

Trees have the advantages to have small biases and to be grown rapidly from both categorical or continuous variables. However, they suffer from instability and thus have high variance: Small changes in the data can yield to very different series of split.

A.2.2 Random Forests

Random forests (Breiman, 2001a) have been proposed to overcome the instability of trees. A random forest averages the results of B trees grown identically, thus not affecting the bias of the whole estimator. The variance reduction is performed by forcing the trees to be different from each other. This is achieved by introducing randomness in the tree growing procedure. At each split, only a random subset of features are selected. On average, the errors of each individual tree cancel each other, thus reducing the high variance of each individual tree without a high increase in bias. Figure A.2 shows on a toy dataset random forests with increasing number of trees of depth 4.

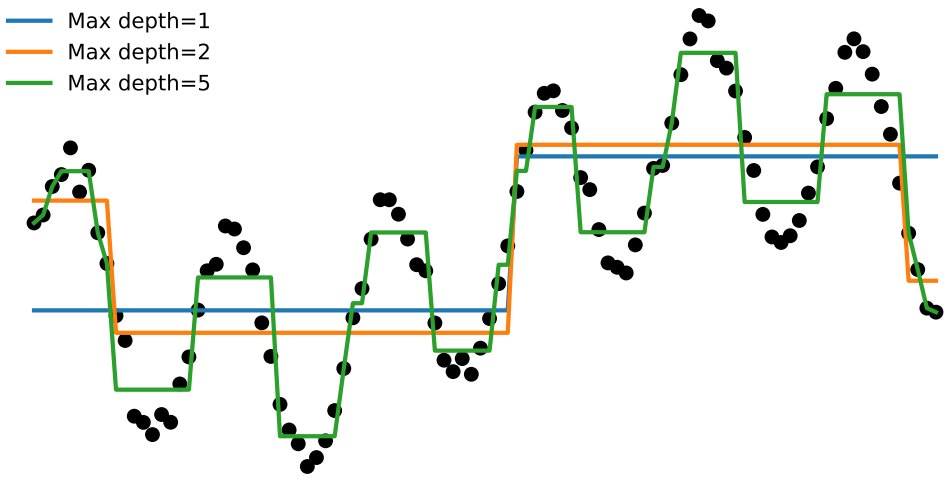


Fig. A.1. Illustration of regression trees with depth 1, 2 and 5.
Code adapted from `https://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_regression.html`.

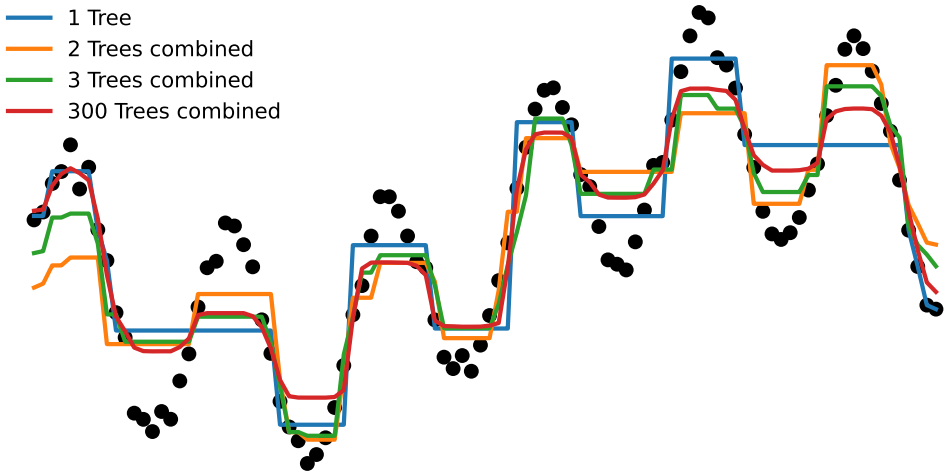


Fig. A.2. Illustration of random forest with increasing number of trees.
Code adapted from `https://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_regression.html`.

A.2.3 Gradient Boosting

Random forest is the aggregation of overfitting estimators. On the contrary, boosting (Freund; Schapire, 1995) is a method that iteratively sums underfitted learners –typically trees– to each other. Each new weak learner improves over the errors of the previous averaged estimators. Improvement on errors can be performed by solving analytically the error loss function or numerically by gradient descent (Friedman, 2001). The later has the advantage to avoid a new implementation for every new loss function. The final estimator has the same form than a random forest, but require substantially less trees to converge. Figure A.3 illustrates on a toy example of boosting trees with increasing number of trees of depth 4.

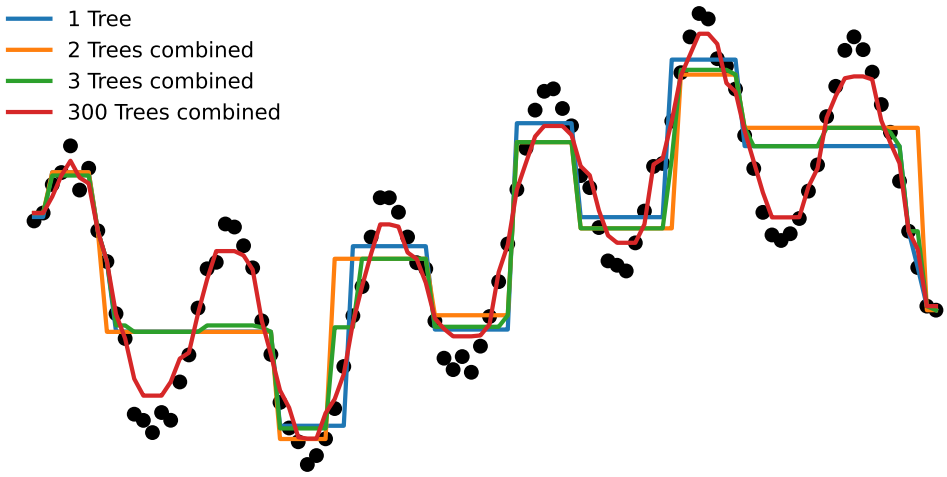


Fig. A.3. Illustration of boosting trees with increasing number of trees.
Code adapted from `https://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_regression.html`.

Appendix B

Chapter 2

B.1 List of interviewed stakeholders with their teams

Clinical Data Warehouse	Teams
CDW_AMIENS	IT : 1,MID : 1
CDW_ANGERS	Data Direction : 1
CDW_APHM	Clinician : 1,CDW team : 2
CDW_APHP	CDW team : 4,IT : 5
CDW_BORDEAUX	CDW team : 1,Inserm : 1,public health : 2
CDW_BREST	CDW team : 1,MID : 1
CDW_DIJON	CDW team : 1
CDW_EDSAN	CDW team : 2,MID : 1
CDW_HCL	Clinician : 1,Data Direction : 1,IT : 1,Inserm
CDW_INCLUDE_LILLE	Administration : 2,CDW team : 3,public health : 2
CDW_MARTINIQUE	CDW team : 1,public health : 1
CDW_MONTPELLIER	Data Direction : 2,MID : 1,public health : 1
CDW_NANCY	CDW team : 2,public health : 2
CDW_NANTES	CDW team : 2,public health : 1
CDW_POITIERS	IT : 2,CRD : 1
CDW_PREDIMED_CHUGA	CDW team : 3,public health : 2
CDW_REIMS	Clinician : 1,CDW team : 1
CDW_RENNES	CDW team : 2,public health : 2
CDW_STRASBOURG	CDW team : 2,public health : 2
CDW_TOULOUSE	CDW team : 1
CDW_TOURS	CDW team : 1

B.2 Interview form

Topics	Questions
Initiation and Construction of the Clinical Data Warehouse	How was the initiative born, when, which team(s) involved in the construction? A Data warehouse to meet what initial needs ?
	What was (is) the articulation between the medical informatics / engineer(s) / Clinical Research Department and user team(s), biostatistics ?
	Governance: How should the teams be organized for the creation and maintenance of the warehouse, data access, and project teams?
	What types of data are present in the warehouse from the following non-exhaustive list: Billing codes , other administrative data, other procedures, structured procedures and diagnoses, structured biology measures, structured drug treatments, emergencies, resuscitation, anesthesia, texts (letters, Clinician Reports), imaging, anatomopathology, sequencing.
	What are the medico-social/social data, especially from social and medico-social institutions ?
Current status - Ongoing and finished projects	Who are the main users? For what purposes (research, quality improvement, management, clinical usage)? Which therapeutic area(s)?
	What are the major types of projects from the following non-exhaustive list: Cohort development, descriptive epidemiology, analytical (comparative) epidemiology with/without randomization, monitoring and dashboards, indicators, inclusion in clinical trials.
	How many projects are completed / started / planned?
	What are the tools and methods used for these projects? Cohort building tool, standard data formats, NLPs, ...
	Is there a valorization strategy for the Clinical Data Warehouse?
	What connections with external sources such as the national health data platform, the outpatient data, the general practitioner data, the research cohorts ?
Opportunity and obstacles	What are the main difficulties encountered during data warehouse projects?
	Are there themes that deserve more encouragement from the HAS?
	What skills are needed? Are there any skills or technical resources missing?
Quality criteria for observational research	Coverage: How is it monitored? Geographically/by department? Time-wise? By what means?
	Cleaning: How are patient duplicates and source alignment managed?
	Database Network: Does the warehouse belong to a health database network?
	Data quality: Are there automatic reports on data quality? Frequency, design, code and documentation available? Presence of dedicated personnel or even a team to check the quality of the data continuously, and to carry out quality controls of the data on the central base, on the study bases?
	Data life cycle: Is there a reference document on the different stages of the data life cycle? How is this document kept up to date with the constant evolution of the warehouse? In what form? How is this documentation managed, accessed, updated and corrected? Precise description of the integrated fields?
	Harmonization procedure: What are the data structures/formats and coding systems used? (eHop, I2B2, OMOP, HL7 FHIR, other ?)
	Machine learning: If machine learning systems are used (e.g., for extracting and structuring information), is there specific documentation on their performance? For manual coding (e.g., labelling), is there a coding guide? Has a measurement of inter-coder consistency been conducted?
	De-identification: Elements on de-identification if applicable, performance metrics
	Constructed phenotypes: Are there operational definitions of target populations (study cohorts) and how are these compared to conceptual definitions, i.e., business and scientific definitions? Is there a study of FPR/TPR in relation to a reference standard? Are these definitions made public either with the study results or in the documentation of the warehouse?
	Transparency: Are the studies registered on a dedicated or pre-existing portal (epidemio-France, encepp (EU), clinicaltrials.gov (US))? Are the study codes made accessible as for opensafely? Are the publications accessible in open access, once the studies are completed?
	Multidisciplinarity: Are the project teams multidisciplinary? Specification of the participations for each part of the analysis from the data collection from the raw Information System.
Topics of interest to the HAS	Quality Department: quality indicators (french IQSS): coordination (patient assessment for discharge, patient contact at D+1), quality of the liaison letter), management (eligibility for the outpatient surgeries, pain management)
	Health Technology Assessment Department: hospital biology activity (description), adverse events associated with procedures, post-registration studies (procedures, early access oncology). Evaluation of procedures: e.g., biological and imaging procedures performed in hospitals, genetic tests in oncology and rare diseases.
open discussion	

B.3 Study data tables

The data tables used to produce the figures in the results section are available at the following url:
https://gitlab.has-sante.fr/has-sante/public/rapport_edsh/.

- The guests table concerns the individuals interviewed, the interview dates, the positions and the membership of a specific team: https://gitlab.has-sante.fr/has-sante/public/rapport_edsh/-/blob/master/data/cycle_eds/cycle_eds_intervenants.csv
- The warehouse table collects information about the CDW: https://gitlab.has-sante.fr/has-sante/public/rapport_edsh/-/blob/master/data/cycle_eds/cycle_eds_entrepots.csv
- The study table references the in-progress study titles and objectives from 10 public declarative portals in progress: https://gitlab.has-sante.fr/has-sante/public/rapport_edsh/-/blob/master/data/cycle_eds/cycle_eds_etudes.csv

Appendix C

Chapter 3

C.1 Code

The code for the experiments is available at https://github.com/soda-inria/predictive-ehr-benchm

A fork of the CEHR-BERT implementation is available at https://github.com/strayMat/cehr-bert. It was necessary to adapt the code to our data format and to the AP-HP computing environment.

All decayed counting and static embedding featurizers are available as a standalone package at https://gitlab.com/strayMat/event2vec.

C.2 Predictive models and tasks on EHRs

C.2.1 Why predictive models in healthcare ?

Risk stratification in the clinic Early prediction of a complication calls for early intervention. Focused on short term interventions in the clinic, these models seek to increase short or long terms outcomes of the patients thanks to early intervention before deterioration (Tang et al., 2007; Rothman et al., 2013; Wong et al., 2021). Those so-called alert systems (Yu et al., 2018) map complex inputs –e.g., a combination of nursing assessments, vital signs, laboratory results and cardiac rhythms– to simpler risk scores, allowing clinicians to rapidly judge the evolution of the patient. The same kind of risk stratification is also used for long term prevention under the term *screening*.

Predict to identify important risk factors Risk stratification is close to the research of risk factors. The Framingham risk score, one of the earliest predictive models in medicine was designed to predict Coronary heart disease risk by fitting a cox model using seven features on 5300 patients: age, cholesterol, systolic blood pressure, hematocrit, ECG status, smoking at intake, and relative body weight (Brand et al., 1976). The authors aim to identify important risk factors allowing the selection of individuals for intervention programs. Biostatistics also focuses on risk factors for subgroup identification, framing this task as therapeutics (Steyerberg, 2009) or heterogeneous treatment effect (Harrell et al., 2001).

Prognosis for automated decision-making Multiple applications of Artificial Intelligence in Medicine were already discussed in the early 80s (Szolovits, 1982, Chapter 1): diagnostic and therapeutic program for glaucoma (CASNET), diagnostic and therapy for infectious diseases (MYCIN) (SHORTLIFFE, 1976), therapic advice for patients with heart disease, diagnosis in general internal medicine (INTERNIST-I). In this line of work, a good predictive model often serves as a module in a larger decision-making system aiming at personalized medicine (Topol, 2019).

Predict to better plan and pilot In complex healthcare organizations, accurate individual predictions help to use efficiently constrained medical resources (Topol, 2019). One of such operational tasks is Length Of Stay (LOS) prediction, allowing to plan the number of beds and members of staff required, identify individual outliers (Verburg et al., 2017). Identified as a quality of care indicator, unplanned readmission at 30 days is used to benchmark and finance hospitals in several countries (CMS for Medicare, 2019; Kristensen et al., 2015). In addition to the risk reduction for patients, it incentivized the hospital to develop better prediction models for unplanned readmission.

Interestingly, recent developments of predictive models on EHRs (described in next section) focused mainly on administrative tasks and less on the risk factors or the intervention aspect.

C.2.2 Predictive models on EHRs: from simple to complex

Predictions on EHRs originally used linear models on few carefully selected statics variables Early work for predictive models on EHR used parsimonious logistic regression (selecting 10 variables on average) to predict Heart Failure (12% case prevalence) within 6 months reaching 0.77 AUC (Wu et al., 2010).

Goldstein et al., 2017 identified key characteristics of 107 studies on predictive models on EHRs: a very large study size (median=26,100), few predictors are included (median = 27 variables), few longitudinal data (37 studies), half multi-center studies. The tasks and performance were mortality with 0.84 AUC, clinical prediction (various clinical endpoints) with 0.83 AUC, hospitalization with 0.71 AUC, service utilization with 0.71 AUC. Generalized linear models such as logistic regression or Cox regression (87 studies) were the most common, followed by Bayes methods (11 studies), random forests (10 studies) and regularized regressions (7 studies).

Including high-cardinality and time-varying features thanks to neural networks Acknowledging the high cardinality of medical vocabularies used in EHRs, medical informatics leveraged representation learning (mostly algorithms used in Natural Language Processing) to embed them into low dimensional feature spaces (Shickel et al., 2017). The goal was to reduce the costly feature engineering work used by traditional predictive models.

This line of work focused first on concept representations: restricted bolzmann machines for suicide predictions (13.1% prevalence) (Tran et al., 2015).

Then it included time: word2vec with temporal contexts (Beam et al., 2019), convolutional neural networks for unplanned readmission at 6 months (balanced case/control) with 0.82 AUC (Nguyen et al., 2016), LSTM for diagnosis from physiological signals with 0.86 weighted ROC AUC and 0.12 precision at 10 (Lipton et al., 2016), recurrent neural networks for heart failure (selected 9 to 10 control to cases) with 0.88 AUC (Choi et al., 2017), recurrent neural networks for Length Of Stay over 7 days (20.68% prevalence), mortality (1.74% prevalence) and 30-day readmission (12.93% prevalence) with respectively 0.79, 0.87 and 0.70 AUC all using 24 first hours of observation data (Beaulieu-Jones et al., 2021).

Rare benchmarks of these different methods include :

- Harutyunyan et al., 2019 benchmarked four clinical tasks on MIMIC-III database, a EHR rich in signals since it covers Intense Care Units. The tasks were in-hospitality prediction based on the first 48 hours of data, decompensation prediction at each hour, Length-Of-Stay prediction at each remaining hour of stay, phenotyping of 25 acute care conditions. This benchmark uses 17 time-varying clinical measurements but with a measure each hours, focusing on the temporal part of EHRs. They found that LSTM processing separately each signal were the mist performant for all tasks.

- Solares et al., 2021 benchmarked different concept embedding methods to predict the presence of three ICD10 codes (level 2) at six months in General practitioner data. They use either Auto Encoders to reconstruct the patient histories, Neural collaborative filtering with positive/negative sampling, Continuous Bag of Words (context window is not precised), CBOW with time aware attention for the context window (Cai et al., 2018), and BEHRT (Li et al., 2020b). This benchmark covers the high-cardinality part of EHRs. There is less emphasis on temporality since patients only have up to 10 different visits.

The age of foundation models for EHRs? Foundation models (FMs) are machine learning models capable of performing many different tasks after being trained on large, typically unlabeled datasets (Wornow et al., 2023). Leveraging the transformer architecture Vaswani et al., 2017 that proved efficient for Natural Language Processing, large EHR modeling models are pre-trained on large volume of data, then evaluated on downstream tasks.

BEHRT focus on 301 diseases predictions in next general practitioner visits within 6 months with 0.958 AUROC and 0.525 APS. It was pretrained on 1.6 million patients from UK general practitioners encounters diagnoses. Med-BERT focus on heart failure for diabetes patients (DHF) (6.2% prevalence) and pancreatic cancer prediction (30% prevalence) with respectively 84 and 74 AUROC (Rasmy et al., 2021). It was pretrained on 28.5 million patients from IBM MarketScan billing codes. Cerh-Bert predicts 30-days all-cause readmission in heart failure (24.116% prevalence), mortality within 1 year since discharge to home (4.85% prevalence), heart failure for DT2 patients (9.38% prevalence) and 2 year risk of hospitalization starting from the 3rd year since the initial entry into the EHR (10.9% prevalence) with respective AUROC/APS 66/38.6, 94.6/52.7, 80.7/32.3, 75.9/31.1 (Pang et al., 2021). It was pretrained on 2.4 million patients from the Columbia University Irving Medical Center-New York Presbyterian Hospital.

Wornow et al., 2023 review other published FMs for EHRs highlighting the small scale of patients used for pretraining, and the lack of model weights accessibility. To improve upon the existing they advise: 1) for better predictive performance (measured both on AUOC, AUPRC and ranking metrics) as well as calibration performance; 2) to detail the performance with the number of labelled cases; 3) simplified model deployment; 4) Emergent clinical applications; 5) Multimodality.

C.3 Number of cases used in foundation models downstream tasks

A strong argument in favor of foundation models trained on EHRs is their ability to transfer well to downstream once pretrained. However, in most of the papers presenting these models, the number of cases used for the downstream tasks is out of reach for numerous medical applications. Table C.1 shows the number of cases for downstream tasks of Med-BERT and CEHR-BERT. All tasks exceed 10,000 cases, a number of cases out of reach for the vast majority of healthcare centers. Figure C.1 outlines the performance of BEHRT on every diagnosis. There is no clear trend that outlines better performance for higher number of cases. However, the number of cases is still greater than 5000 for almost all targeted diseases.

Coupled to the current impossibility to share (and transfer) the weights of these models, this need for large downstream datasets is a strong limitation to their use in practice.

Model	Task	Cohort size	Prevalence	Number of Cases	ROC AUC	AUPRC
CEHR-BERT (Pang et al., 2021)	Heart Failure readmission	97,758	24.16%	23,618	66.3 (0.2)	38.6 (0.1)
CEHR-BERT (Pang et al., 2021)	Discharge home Death	207,919	4.85%	10,084	94.6 (0.1)	52.7 (0.4)
CEHR-BERT (Pang et al., 2021)	T2DM Heart Failure	114,564	9.38%	10,746	80.7 (0.6)	32.3 (1.0)
CEHR-BERT (Pang et al., 2021)	Hospitalization	590,578	10.90%	64,373	75.9 (0.1)	31.1 (0.4)
Med-BERT (Rasmy et al., 2021)	Pancreatic cancer Cerner	29,405	39%	11,486	82.23 (0.29)	75.08 (0.36)
Med-BERT (Rasmy et al., 2021)	Pancreatic cancer Truven	42,721	40%	17,088	80.57 (0.21)	71.54 (0.45)
Med-BERT (Rasmy et al., 2021)	T2DM Heart Failure	672,647	6.2%	39,727	85.39 (0.05)	83.8 (0.05)

Table C.1. For downstream tasks of both CEHR-BERT and Med-BERT, the number of cases (ie. number of patients with the positive class) is out of reach for numerous medical applications were the number of positive classes is closer to the thousand –in the best cases.

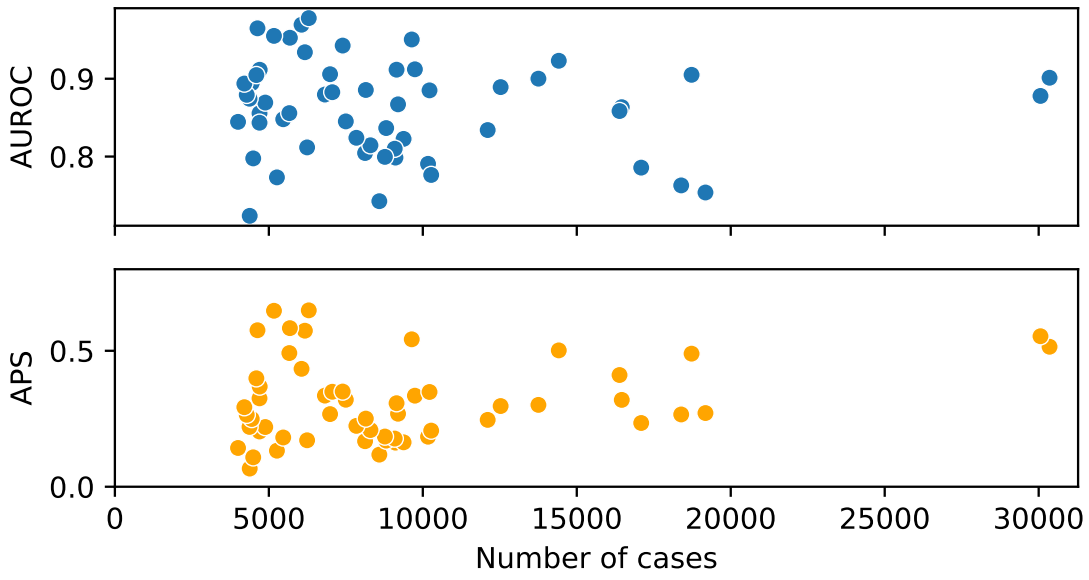


Fig. C.1. BEHRT (Li et al., 2020b) performance for every diagnosis target. Number of positive cases is above 5000 for almost every disease.

Data Type	GPU number	GPU model	Training time (hours)	Pretraining	Sample size (millions)	Server details	Reference
Claims	4	Nvidia V100	<24	No	43	72 CPU cores	Beaulieu-Jones et al., 2021
EHR	NA	Nvidia Titan Xp	NA	Yes	1.6	NA	Li et al., 2020b
EHR	2	Nvidia RTX 20280 Ti	45	Yes	2.4	768 GB memory	Pang et al., 2021
EHR	1	Nvidia V100 32 GB	168	Yes	28	NA	Rasmy et al., 2021
Text	24	Nvidia A100 40GB	504	Yes	0.39	Pre-training on NYU Langone High-Performance Computing; Deployment on 128 GB RAM servers with 2 RTX 3090 GPUs	Jiang et al., 2023
Text	992	Nvidia A100	144	Yes	2	HiperGator-AI cluster	Yang et al., 2022

Table C.2. Computing resources for modern large scale predictive models.

C.4 Review of computing resources for modern predictive models in healthcare

The increasing architecture sizes of predictive models require appropriate computing infrastructures to pre-train and deploy those models. Table C.2 reviews the computing requirements for some of these modern predictive models. Computing resources for Large Language Models trained with clinical notes are greater than for EHR or claims models. Numbers are rarely provided at deployment , since they require smaller computing resources.

C.5 Detailed pipelines

For all methods except demographics, the event features are the following structured events: billing codes (ICD-10), procedure codes (CCAM nomenclature), drugs administration (ATC7 nomenclature). Despite their high predictive potential, we did not consider biology since the number of events was too big for our memory capacity.

C.5.1 Demographics

All static features correspond to the index visit of the task (T1=target stay, T2=first included stay, T3=???): age, gender, admission reason, discharge destination, type and value.

The feature time of day was built as followed: morning between 7am and 12pm, afternoon between 12pm and 20pm and night between 20pm and 7am. This feature was mostly set at night, because almost all times are set to 22:00pm in our data extraction. More fine grained details should be available in the information system but were not communicated to us.

C.5.2 Decayed counting

During cross-validation, the explored decay parameters are: $\Big[[0], [0,1], [0,7], [0,30], [0,90]\Big]$.

C.5.3 Static Embeddings of event features

SVD-PPMI We recall the SVD-PPMI algorithm developed by Beam et al., 2019 and used for transferring phenotyping in Hong et al., 2021.

The algorithm takes a sequence of coded events as input and outputs vector representations.

As shown in Figure C.2, it builds a context window around every event $e = (i, t, c)$, then updates the cooccurrence matrix for the corresponding medical code $P(c, c_j) \, \forall \, c_j \in Vocabulary$.

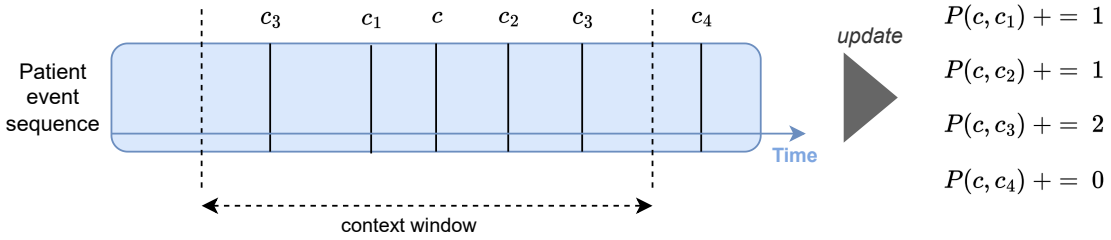


Fig. C.2. The cooccurrence matrix is updated when two events occur in a specified time window.

The PPMI matrix is then computed as the logged-shifted version of the cooccurrence matrix.

$$PMI(c_i, c_j) = log \frac{P(c_i, c_j)N}{P(c_i)P(c_j)} \tag{C.1}$$

where $P(c_i)$ is the total count of the event c_i and N is the total count of events.

$$PPMI = max(0, PMI - log(k)) \tag{C.2}$$

Finally, concept embeddings are recovered by SVD factorization and taking the mean of the context and target word matrix.

$$PPMI = U \cdot \Sigma \cdot V \tag{C.3}$$

$$embeddings = U_d \cdot \sqrt{\Sigma_d} + V_d \cdot \sqrt{\Sigma_d} \tag{C.4}$$

Where the subscript d indicates the restriction to the first d components of the matrices.

Aggregation of the static embbedings Figure C.3 details how the aggregation on a patient sequence is performed before feeding the vector to a scikit-learn estimator.

C.5.4 CEHR-BERT

CEHR-BERT is based on the original BERT architecture (Devlin et al., 2018). Pang et al., 2021 modified previous transformers applied on EHR in two ways. They embed absolute and relative time with age of the patients, using a fourier transform of time numerical values. Secondly, they add the Visit Type Prediction objective. In addition to the usual sequence reconstruction task (masked language model) for concept embeddings, the model also tries to reconstruct the type of the visit associated with the masked concepts (inpatient, outpatient,emergency).

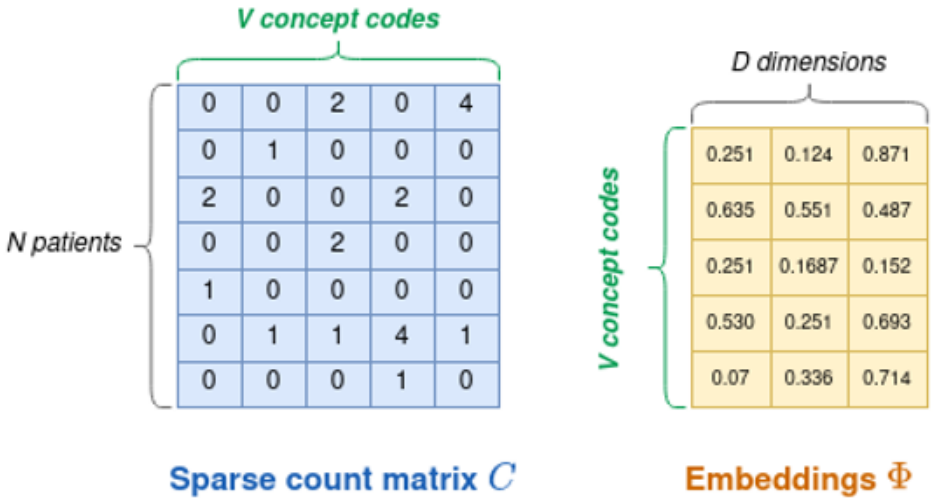


Fig. C.3. Aggregation of the embeddings to yield one vector representing each patient sequence. The left blue matrix is the same as the one computed in the decayed method. It is multiplied by the embedding matrix Φ either learned on the training set or transferred from the SNDS claims database.

C.6 Experimental study

C.6.1 Database description: two extractions from the Paris hospitals data warehouse

We use two data extractions from the clinical data warehouse of the Greater Hospital of Paris (AP-HP) hosting routine care data from 38 hospitals formatted in the OHDSI OMOP format (Hripcsak et al., 2015a). The first extraction containing 200,000 randomly sampled patients, is used for the LOS (C.6.2) and Prognosis (C.6.2) tasks. For the MACE task (C.6.2), the number of case events –prevalence– was lower than for the two other tasks, so we had to work on a bigger extraction containing 2.1 million patients. In both cases, raw data contained diagnoses and procedures billing codes, prescriptions and administrations of drugs, administrative information, laboratory results for inpatient only and clinical notes.

We sessionize the visits to merge indices that are closer than one day into one unique stay. This avoids to consider transfers as two different hospitalizations.

C.6.2 Tasks descriptions

Selection flowcharts The Figure C.4 shows the selection flowcharts for the three tasks.

LOS interpolation

Plan – Long Length Of Stay interpolation (LOS) During complete hospitalization, stays with extreme LOS values are responsible for a large share of hospital costs and resource uses. The ability to predict extreme LOS is useful for resource managements (Omachonu et al., 2004; Caetano et al., 2014; Jiang et al., 2023). We define LOS as a binary task categorizing each inpatient stay as long if LOS greater than 7 days or short if LOS shorter than 7 days. The *study period* was january 2017 to june 2022. The *population at risk* are the patients aged above 18, with at least one hospitalization lasting at least 24 hours. Stays with in-hospital mortality are discarded. The *index visit* is the first included visit for each

Acute Myocardial Infarction	I210, I211, I219, I220, I221, I229
Unstable Angina	I200, I208, I209
Acute Heart Failure	I501, I5020, I5021, I5022, I5023, I5030, I5031, I5032, I5033
Acute Cerebrovascular Events (Stroke)	I60, I61, I62, I630, I631, I632, I633, I634, I64
Other codes	I24, I23

Table C.3. ICD10 codes used for MACE definition.

individual. LOS is defined as a binary classification task for each index visit. The label is 0 if the stay lasts less than 7 days and 1 if the stay lasts more than 7 days. The *horizon* is the end of the index visit and the *observation period* is the full index visit. This task is an interpolation since we classify the length of the stay with data from the whole stay.

Stratify – Prognosis Prognosis is important for prevention as it can influence shared decision-making between the clinician and the patient. We define a proxy for prognosis as a multi-label binary classification task where we predict the next stay ICD10 codes starting from a random non final stay for each patient. We reduce the granularity of the codes to only 21 chapters at the highest level of the hierarchy. For this task, we realized that providing the index stay ICD10 chapters was a strong baseline. We therefore concatenated these codes as separated demographic features for all pipelines after aggregation.

The study period was january 2017 to june 2022. The *cohort* are the patients aged above 18, with at least two hospitalization (inpatient or outpatient). Stays with in-hospital mortality are discarded. The *index stay* is defined as a random stay for each individual before its last stay. The *horizon* is the beginning of the next stay and the *observation period* is the full patient trajectory up to the end of the index visit. Prognosis is defined as a multi-label classification task. In the following stay after the index visit, for each of the 20 ICD10 chapters with a prevalence greater than 1%, the label is 1 if the stay contains a diagnosis in this chapter, otherwise it is 0.

Prevent – Predict Major Adverse Cardiovascular Events Prediction of all-chapters billing codes does not focus on a clinically well-defined population. On the contrary, the composite outcome of Major Adverse Cardiovascular Events (MACE) is often used in clinical trials targets the cardio-vascular risk. We define MACE prediction as the prognosis of incident MACE at one year for a randomly chosen stay for each patient. The study period was january 2018 to december 2020. The *population at risk* are the patients aged above 18, with at least two hospitalization (inpatient or outpatient). *The index visits* are for each patients a randomly selected stay without in-hospital mortality, with at least an *horizon* of 12 months between the end of the stay and the end of the study period. The *observation period* is the whole patient trajectory up to the end of the index visit. MACE is defined as a binary classification task. The label is 1 if a MACE billing code is observed (see Table C.3) and 0 otherwise. We used billing codes defined by Bosco et al., 2021 and complement them with other codes specific to the French healthcare system Caisse Nationale d'Assurance Maladie, n.d.

For well-defined clinical tasks such as MACE, the study population is only a small part of the initial general population. Even when studying regional data warehouses, the relevant population for the algorithm is an order of magnitude smaller than the full population, requiring better sample efficiency than general tasks such as LOS or prognosis.

C.6.3 Training procedure

The temporal split was designed for the train/test ratio to be 0.8/0.2. This resulted in different split dates for each dataset. For LOS, training period covers 2017-01-01 to 2021-04-06 and test period covers 2021-04-07 to 2022-05-01. For Prognosis, train period covers 2017-01-01 to 2021-04-21 and test period covers 2021-04-22 to 2022-05-20. For MACE, train period covers 2018-01-01 to 2019-08-11 and test period covers 2019-08-11 to 2020-01-05.

Note that CEHR-BERT is also pre-trained on growing parts of the train set but only uses a validation set for stopping pretraining. Due to the high computing cost of transformers, we could not cross-validate its internal parameters.

C.7 Supplementary results for temporal split

C.7.1 LOS interpolation

C.7.2 Prognosis

Performance is reported with Area Under the Receiver Operating Characteristic Curve (ROC AUC) and Area Under the Precision Recall Curve (AUPRC) using `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html` and `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html#sklearn.metrics.average_precision_score` implementations to compute the scores.

ROC AUC Figure C.6 displays the AUPRC averaged with equal weight for each chapter (Macro) and with the chapter prevalence as the weight (weighted).

AUPRC scores for every ICD10 chapter Figure C.7 displays the AUPRC averaged with equal weight for each chapter (Macro).

AUPRC scores for every ICD10 chapter Figure C.21 to C.23 display AUPRC scores for every ICD10 chapters with respect to number of patients in the train set. Recall that contrary to ROC AUC where the random baseline is 0.5, the random baseline for AUPRC is the prevalence of the target class.

Prevalence results Figure C.29 shows the prevalence results for the AUPRC curve and the linear estimator.

Figure C.30 shows the prevalence results for the ROC AUC curve and the linear estimator.

Figure C.31 shows the prevalence results for the ROC AUC curve and the random forest estimator.

C.7.3 MACE

C.8 Results for the geographic split

C.8.1 Dataset split by hospital

To evaluate the geographic validity of our results, we proceed to a geographic split for train and test, according to hospitals. In the test set, the patients visited one of the following

hospitals: Avicenne, Jean Verdier, Cochin, Ambroise Pare and Raymond Poincare-Berck. In the train set, patients visited another of the AP-HP 38 hospitals. Patients having a visit in both hospital groups were removed from both the train and test data. Figure C.32 shows the hospital cooccurrence matrix for patient populations.

C.8.2 Hospital split results

LOS interpolation This task is saturated by the best machine learning models so we did not evaluate the transformer on it. Instead, we benchmarked the transfer from cui2vec embeddings, trained on american claims (Beam et al., 2019). To be fair to these embeddings, we restricted the vocabulary of medical codes to the common intersection of 2100 codes occurring both in cui2vec and in our dataset. The decay was set to 7 days (not cross-validated as in the main experiments). Figure C.33 shows all embeddings methods outperforming the count models.

Prognosis Figure C.34 shows the results on the prognosis task, with train/test split by hospital. We average the results for the 21 target diagnosis chapters: either with the same weight for each chapter (macro) or weighted by each chapter prevalence (weighted). The Random forest estimators with SNDS embeddings seem to be the best performing method. The transformer method begins to be competitive only with 8000 thousands patients in the train set. An interesting result is the performance of the naive method that predicts a diagnosis if it is present in the index visit (last stay before prediction target). It outperforms all other methods by a large margin, suggesting than ICD10 predictions might not be a useful predictive task. This result made us add the codes from the previous visit as new static feature to the prediction matrix in the main analysis.

C.9 Vibration study on the effects of the decay

Using the LOS task, we investigated if the decay hyperparameter played an important role in the performance of the different pipelines. We evaluated this vibration analysis on a randomly chose train set (no temporal or hospital data shift). Figure C.35 shows the big impact of different decay hyperparameters on ROC AUC. This led us to cross-validate the decays in the main analyses. Interestingly, concatenating multiple decays only had a significant impact on the count pipelines.

C.10 Medical concept embeddings

C.10.1 Previous work and motivation

Building on a formulation of word2vec (Mikolov et al., 2013) as the factorization of the Positive Pointwise Information Matrix –PPMI eq. C.2–, Beam et al. (2019) built medical concept embeddings from both text and structured data. These embeddings are non-contextual, meaning that they do not take into account the full sequence of event at the time of inference. These make them less powerful than contextual embeddings such as the ones from the transformer-based models (Li et al., 2020b; Rasmy et al., 2021; Pang et al., 2021).

However, the information at the basis of the non-contextual embeddings is a sharable aggregated data: the cooccurrence matrix. The easy creation and exchange of medical concept embeddings opens a wide range of applications that efficiently pulls data from

multiple sites, overcoming both the heterogeneity of the data, and the administrative barriers to access individual patient data. Transferring these dense representations of healthcare events to specific low-sampled cohorts might provide *efficient* pretrained data representations even in setups where very few patients are available.

C.10.2 Background on medical concept embeddings for prediction

Only few works have studied the transfer capabilities of clinical concept embeddings for relevant prognosis tasks. None of them studied how such representations can be used in a combination with traditional machine learning techniques such as random forests or logistic regressions.

Huang et al., 2018 used the same embeddings method on the two EHR in MIMIC, fusing them with Procruste. Utility is evaluated on diagnoses prediction tasks, using Patient Diagnosis Projection Similarity, top-K cosine distance between a temporally weighted average of the embedding for a given patient and the diagnoses embeddings. They show that Procruste fused embeddings are almost as performant (measured with ROC-AUC) as embeddings derived from the two databases.

Hong et al., 2021 used SVD-PPMI factorization on two databases. They evaluated the utility of embeddings on a phenotyping downstream task for eight diseases: coronary artery disease (CAD), type I diabetes mellitus (T1DM), type II diabetes mellitus (T2DM), depression, rheumatoid arthritis (RA), multiple sclerosis (MS), Crohn's disease (CD) and ulcerative colitis (UC). They did not evaluate them for predictive tasks, only for differential diagnosis.

Xiang et al., 2019 fine-tuned a LSTM on top of different medical concept representations (FastText (Bojanowski et al., 2017), SVD-PPMI or Skip-Gram with a time context window (Beam et al., 2019) and), for heart failure (11.7% prevalence) reaching respectively for each representation 84.9, 82.4 and 85.4 ROC AUC.

C.10.3 Embeddings implementation

We published event2vec [1], a python package to quickly build medical concept embeddings using the SVD-PPMI algorithm Beam et al., 2019; Levy; Goldberg, 2014. Our package proposes two backends for the computation of this cooccurrence matrix: pandas for small datasets (ten thousand of patients), spark for big datasets (Salloum et al., 2016).

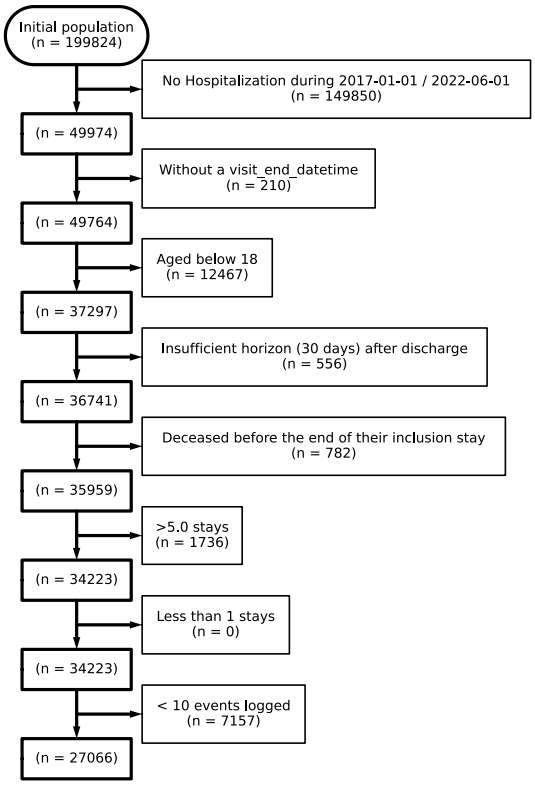
C.10.4 Qualitative assessment of the embeddings

Figure C.36 displays 2D Tsne projections Van der Maaten; Hinton, 2008 of the embeddings. Clear groups of pathologies has been recovered by the algorithm. Table C.4 shows the ten closest neighbors of the *Diabete type 1 with acidosis* concept for each dataset (by cosine distance), reflecting discrepancies between both dataset for this billing code. Interactive Tsne plots and full embeddings (SNDS only) are available on our gitlab repository [2].

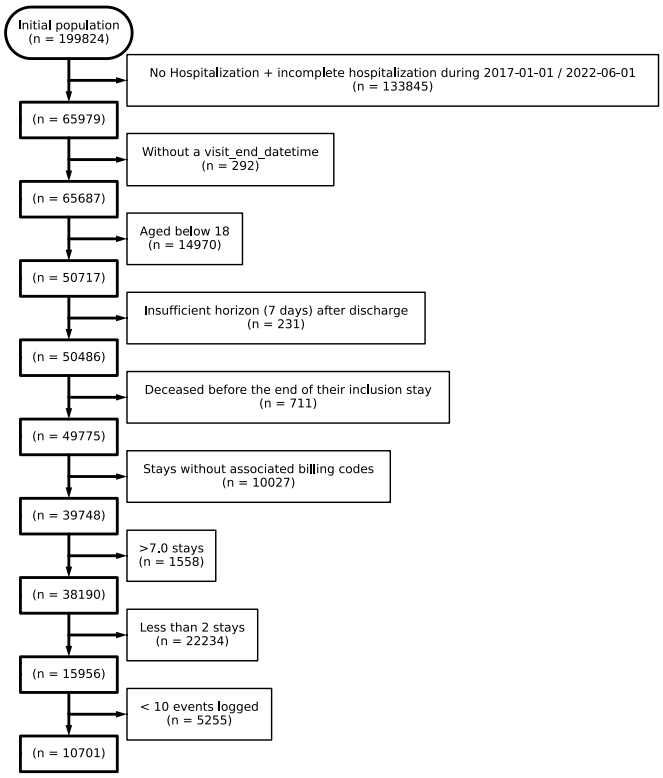
[1] https://gitlab.com/strayMat/event2vec/
[2] https://gitlab.com/strayMat/event2vec/-/tree/main/data/results/

Concept code	Concept label	medical vocabulary	similarity
I210	Acute myocardial infarction	ICD10:diagnosis	1.000000
I25	Chronic ischemic heart disease	ICD10:diagnosis	0.491964
DDAF006	Intraluminal dilatation of a coronary vessel with stenting	CCAM:procedures	0.459349
DDQJ001	Intra-arterial coronary ultrasound and/or Doppler ultrasound	CCAM:procedures	0.455709
B01AC22	Prasugrel	ATC7:drugs	0.426467
C03DA04	Eplerenone	ATC7:drugs	0.385482
TNS	Nicotine replacement therapy	NGAP:GP	0.399372
SRA	Resuscitation supplement	NGAP:GP	0.392237
1526	Creatine phosphokinase	NABM:biology	0.385801
521	Lactate deshydrogenase (LDH)	NABM:biology	0.315254

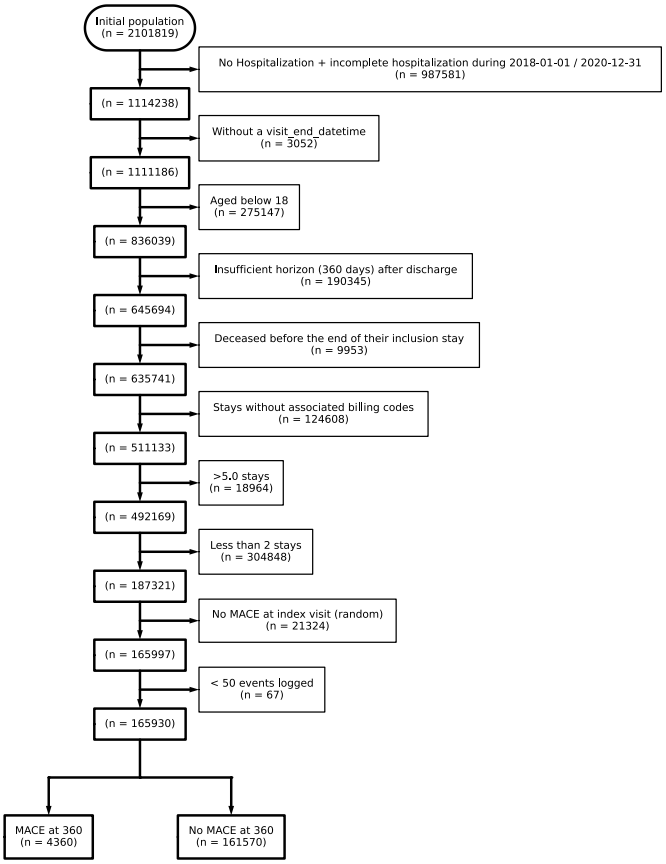
Table C.4. Two closest concepts for each vocabulary for the medical concept I210 of Transmural infarction with the claims SNDS embeddings. Prasugrel is a medication used to prevent formation of blood clots. Eplerenone is an aldosterone antagonist type of potassium-sparing diuretic that is used to treat chronic heart failure and high blood pressure. Creatine phosphokinase used to be determined specifically in patients with chest pain. Because Lactate deshydrogenase is released during tissue damage, it is a marker of common injuries and disease such as heart failure



(a) Selection flowchart for Length Of Stay interpolation task.



(b) Selection flowchart for ICD-10 chapter prediction task.



(c) Selection flowchart for MACE prognosis.

Fig. C.4. Selection flowcharts for the three tasks.

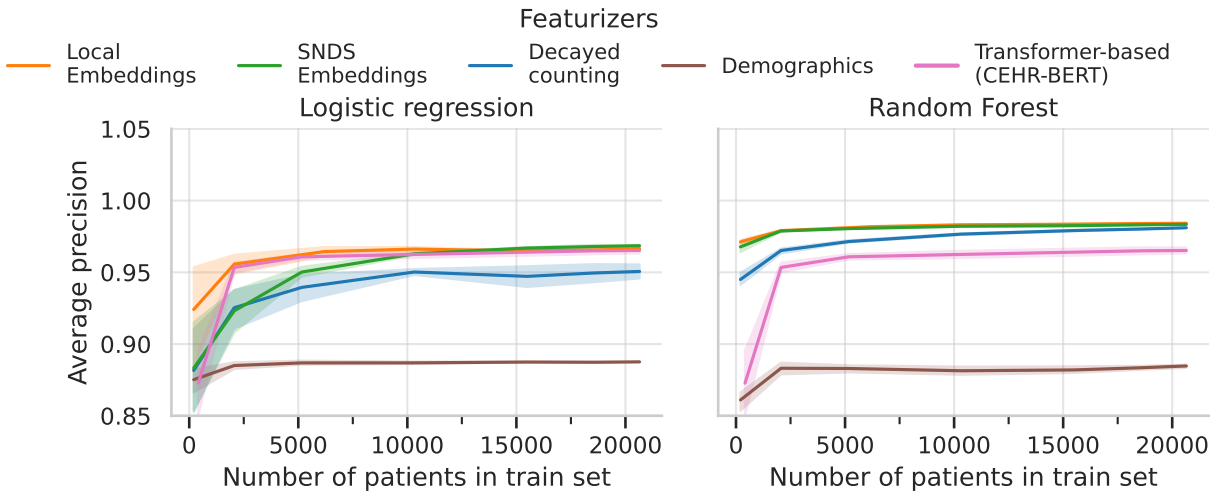


Fig. C.5. LOS AUPRC for different featurizers and estimators. The performance is averaged over 5 folds. The shaded area represents the standard deviation. The task performance seems to saturate at 98% average precision for random forest and all featurizers but the demographics, suggesting that the Bayes error rate is reached. However, for lower sample regimes –below 12,500 patients, we see an advantage of static embeddings over event counts (both for logistic regression and random forest).

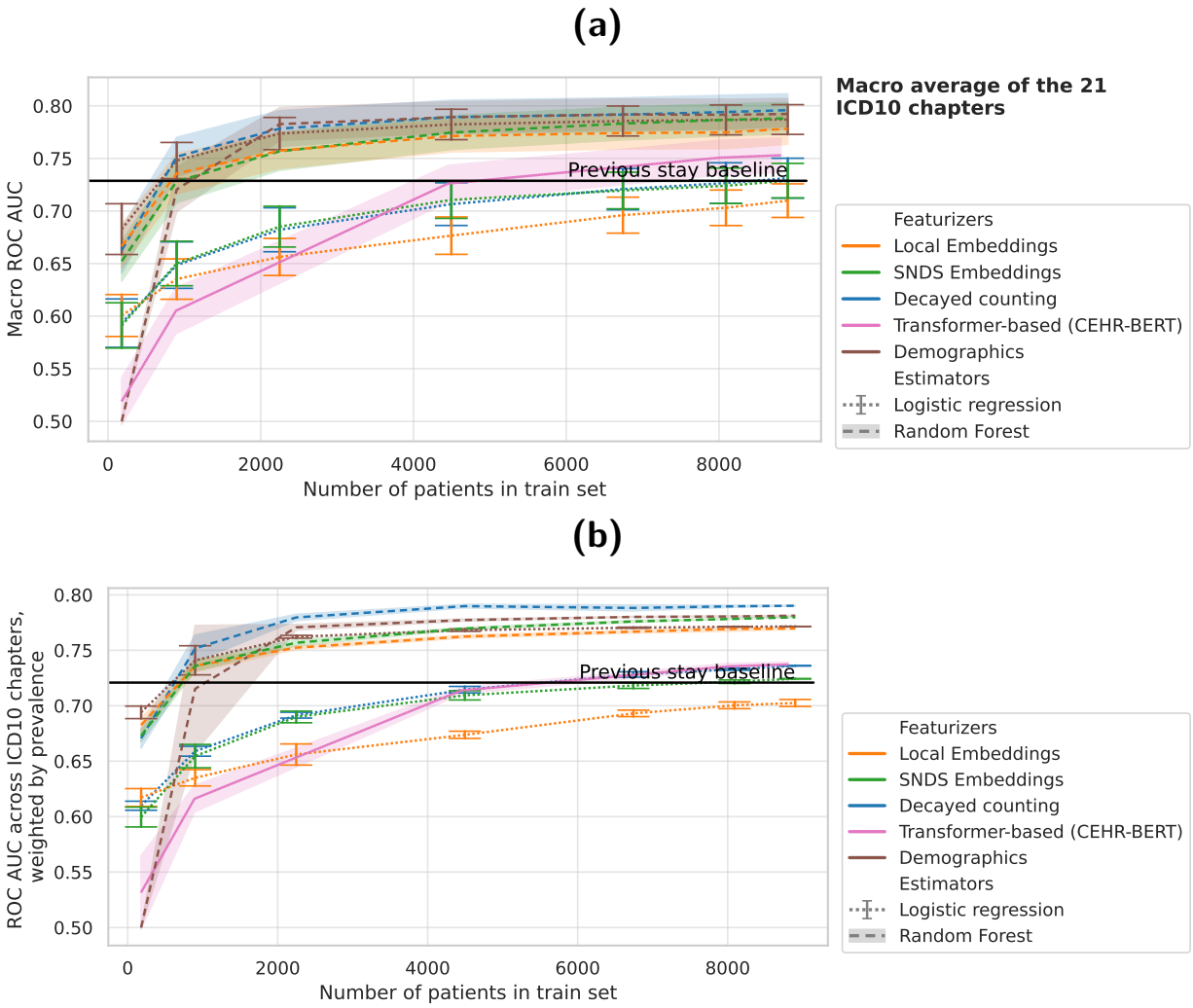


Fig. C.6. Prognosis ROC AUC averaged over chapters for different featurizers and estimators. The performance is averaged over 5 folds. The shaded area represents the standard deviation. The horizontal black lines display the naive baseline that predicts the previous stay codes for the target stay. Figure C.6a average all chapter with equal weight whereas C.6b averages chapters by prevalences. Random forest have better performance. Count encoder outperform other featurizers, suggesting the importance of low count events that are smoothed out in embedding methods.

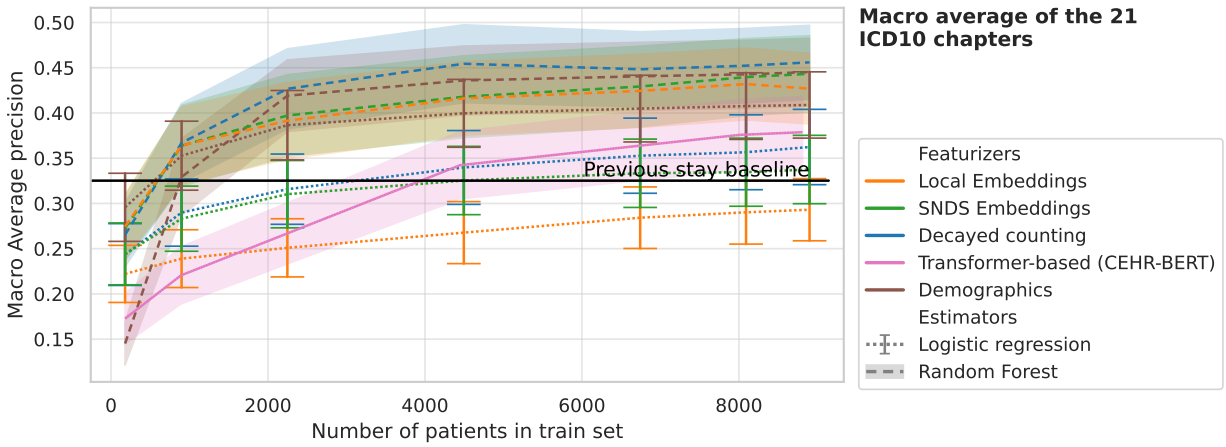


Fig. C.7. Prognosis AUPRC averaged by prevalence over chapters for different featurizers and estimators. The performance is averaged over 5 folds. The shaded area represents the standard deviation. The horizontal black line displays the naive baseline that predicts the previous stay codes for the target stay.

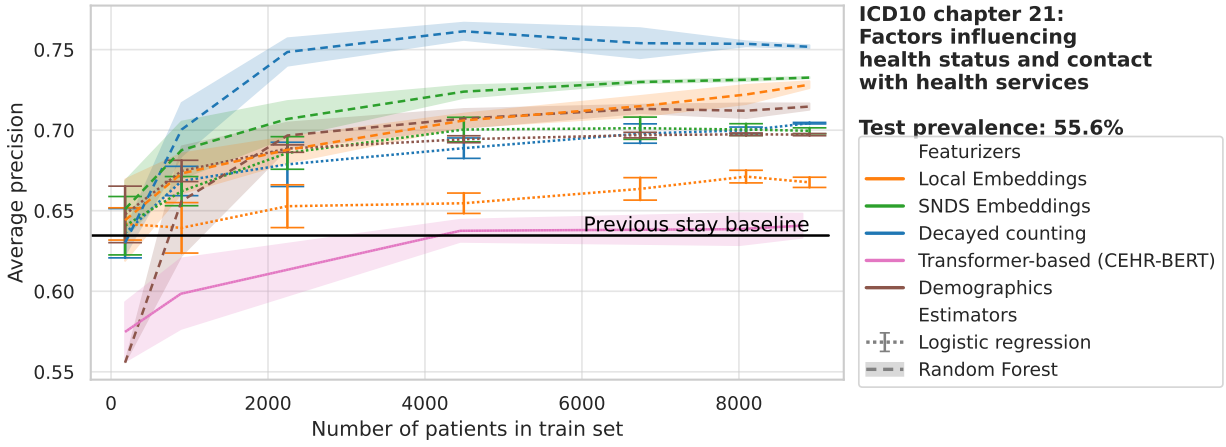


Fig. C.8. ICD10 chapter 21

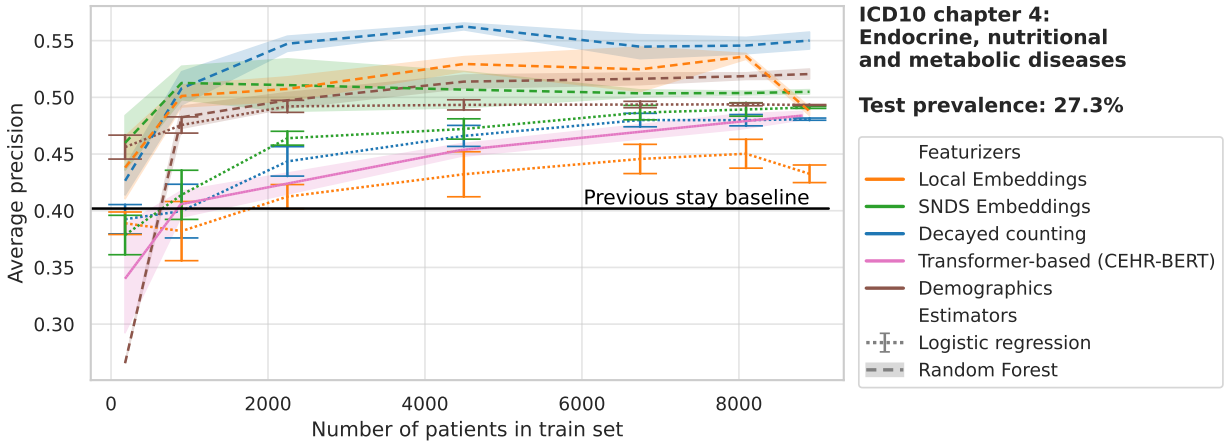


Fig. C.9. ICD10 chapter 4

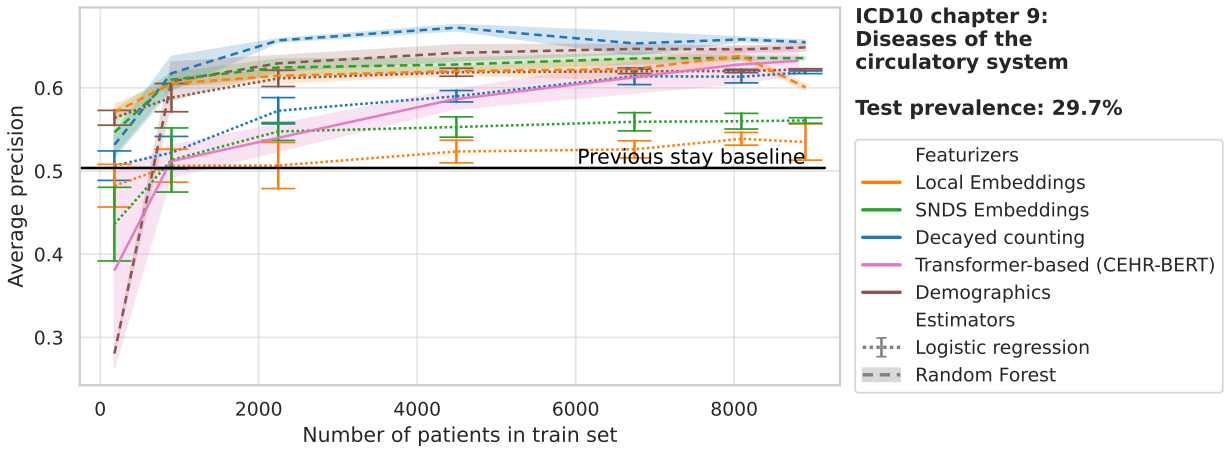


Fig. C.10. ICD10 chapter 9

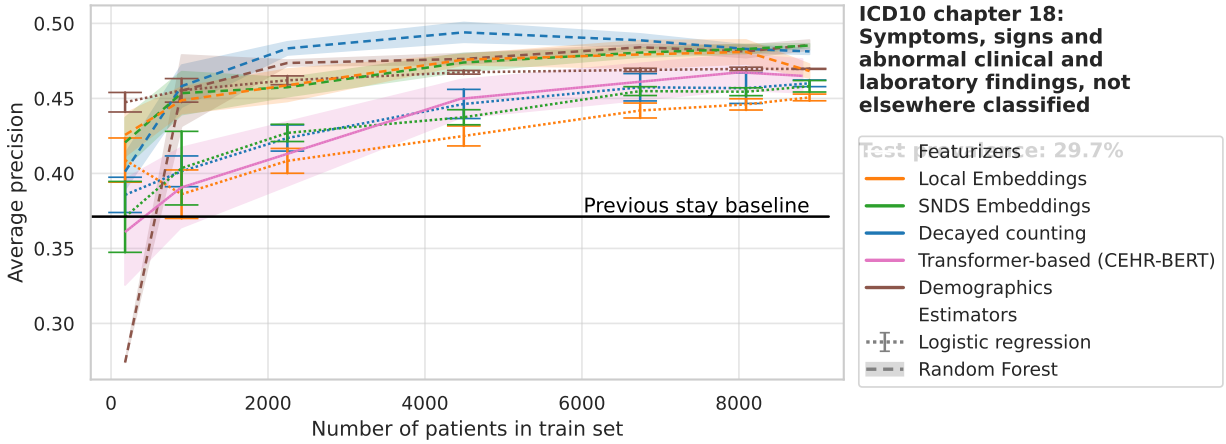


Fig. C.11. ICD10 chapter 18

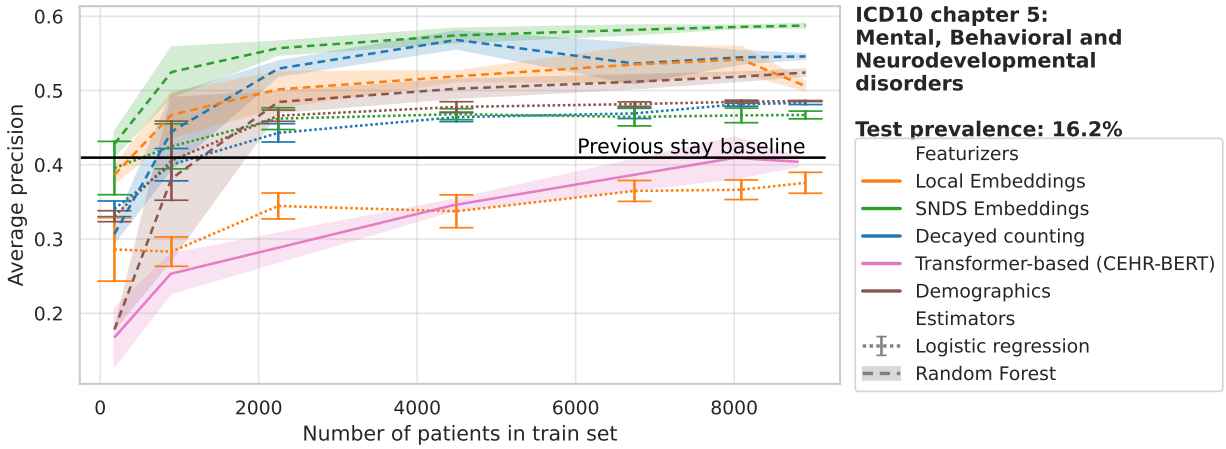


Fig. C.12. ICD10 chapter 5

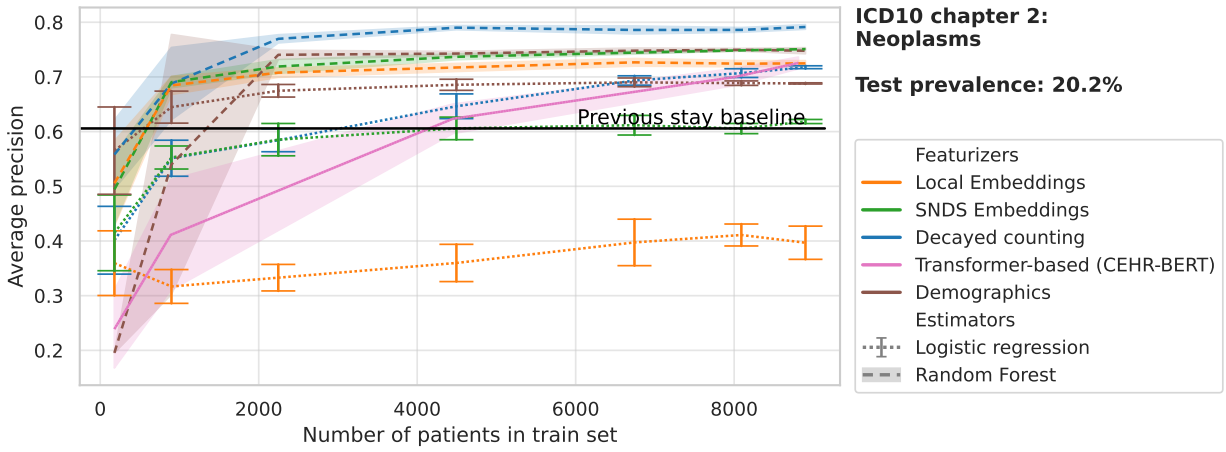


Fig. C.13. ICD10 chapter 2

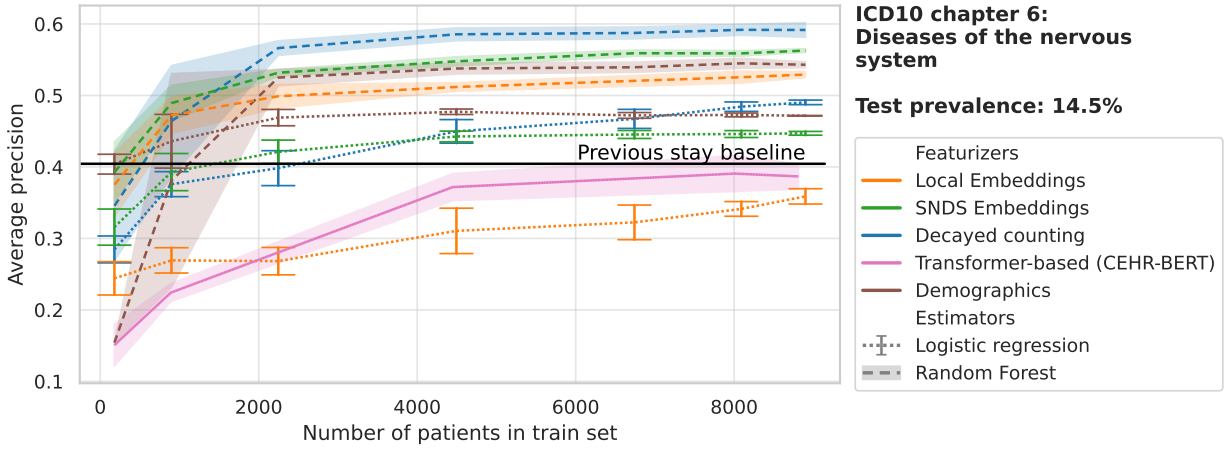


Fig. C.14. ICD10 chapter 6

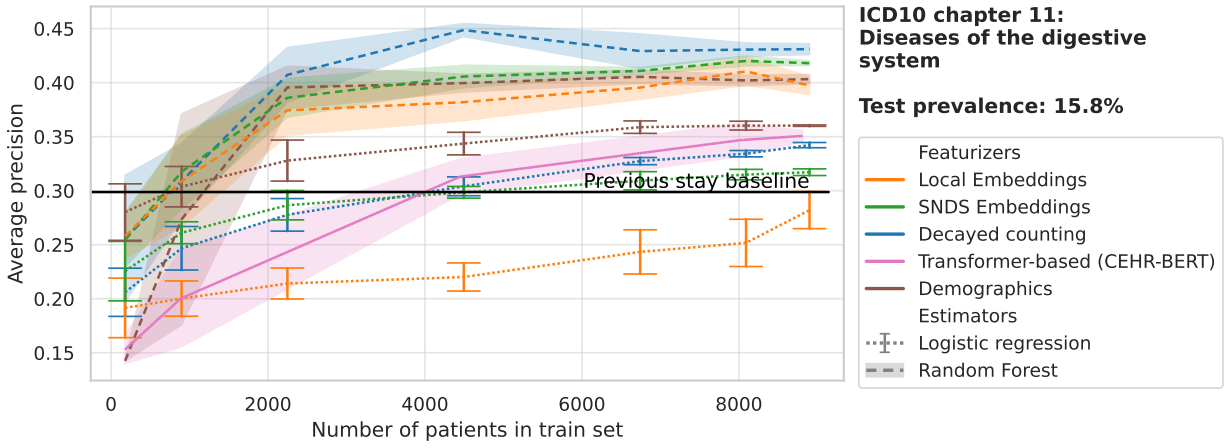


Fig. C.15. ICD10 chapter 11

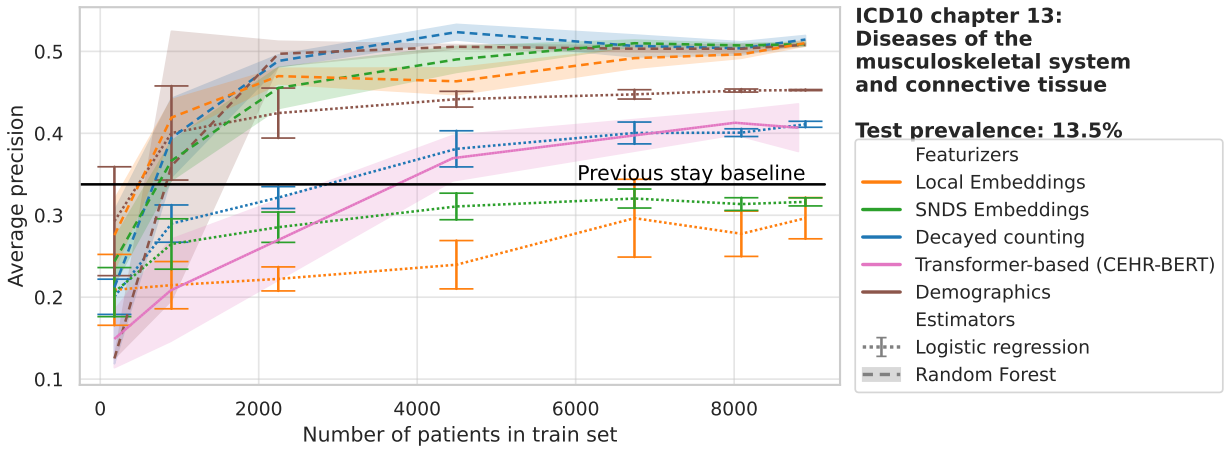


Fig. C.16. ICD10 chapter 13

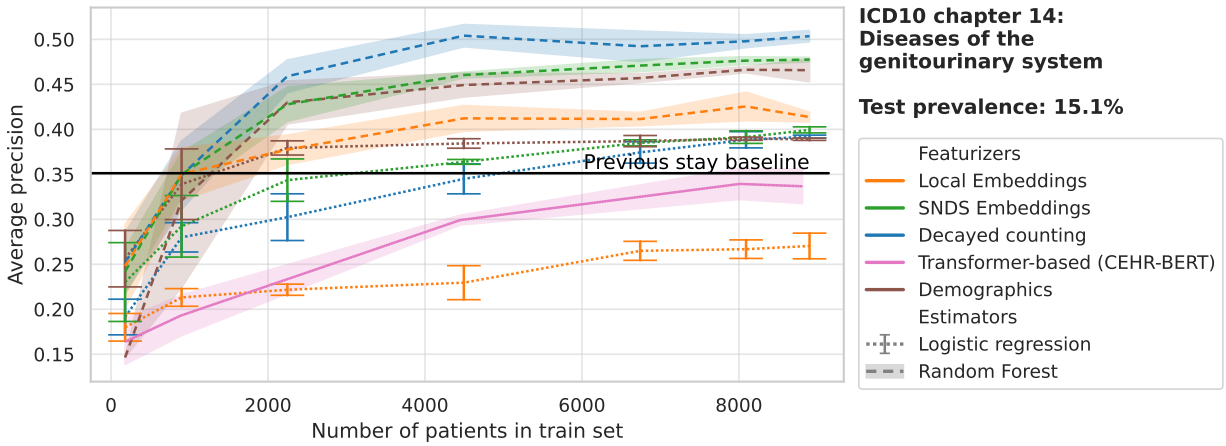


Fig. C.17. ICD10 chapter 14

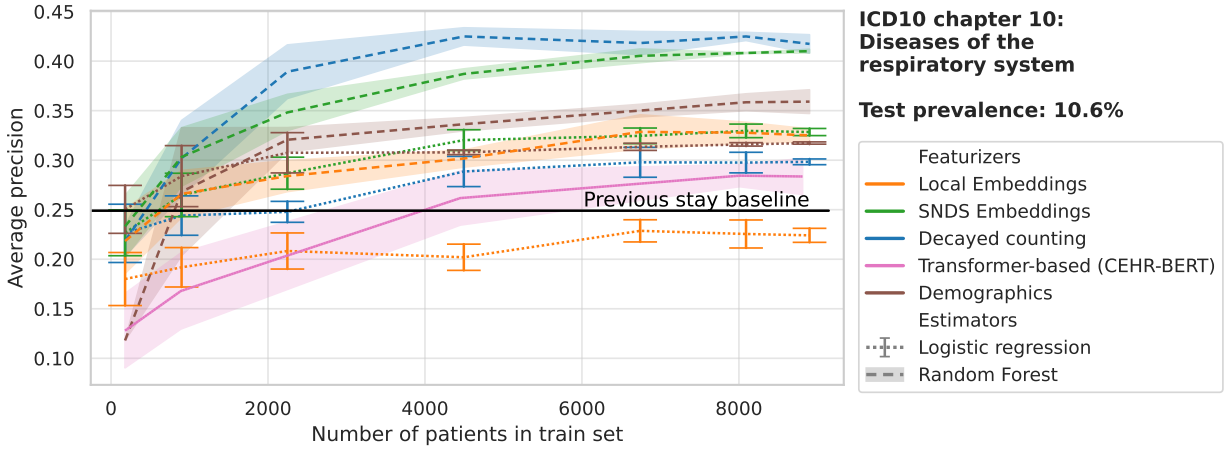


Fig. C.18. ICD10 chapter 10

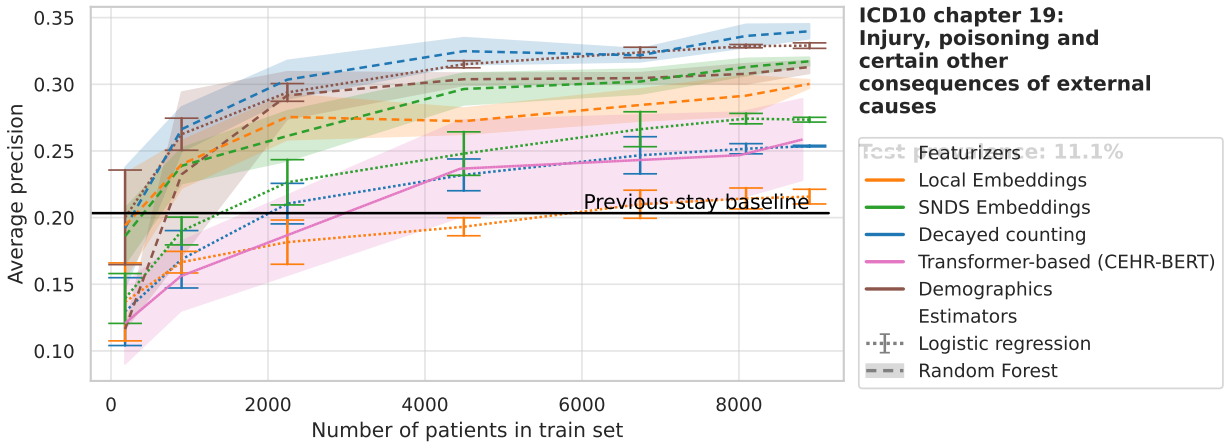


Fig. C.19. ICD10 chapter 19

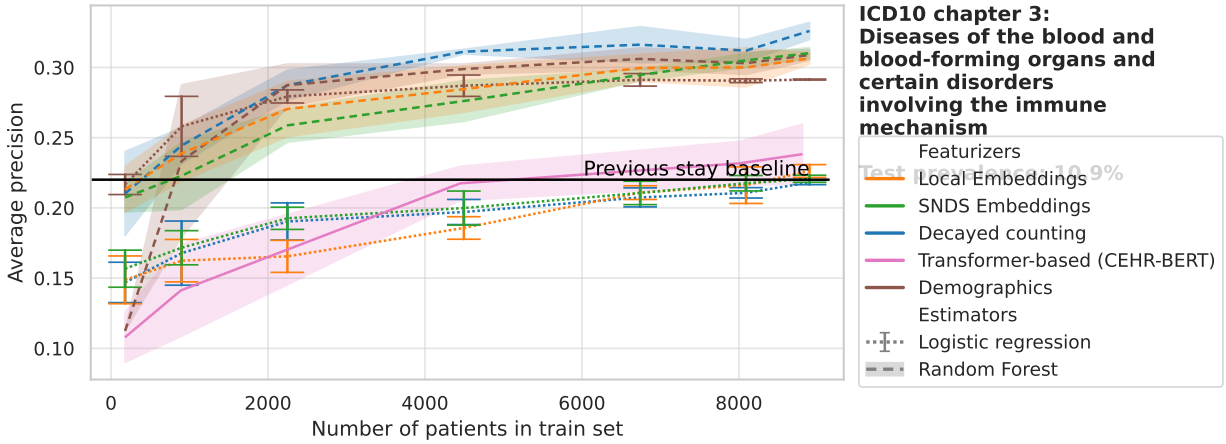


Fig. C.20. ICD10 chapter 3

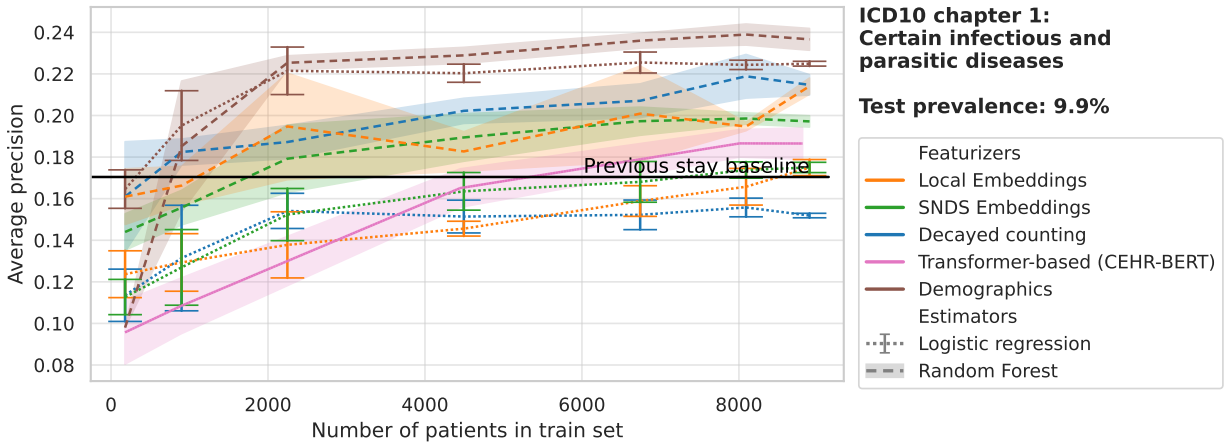


Fig. C.21. ICD10 chapter 1

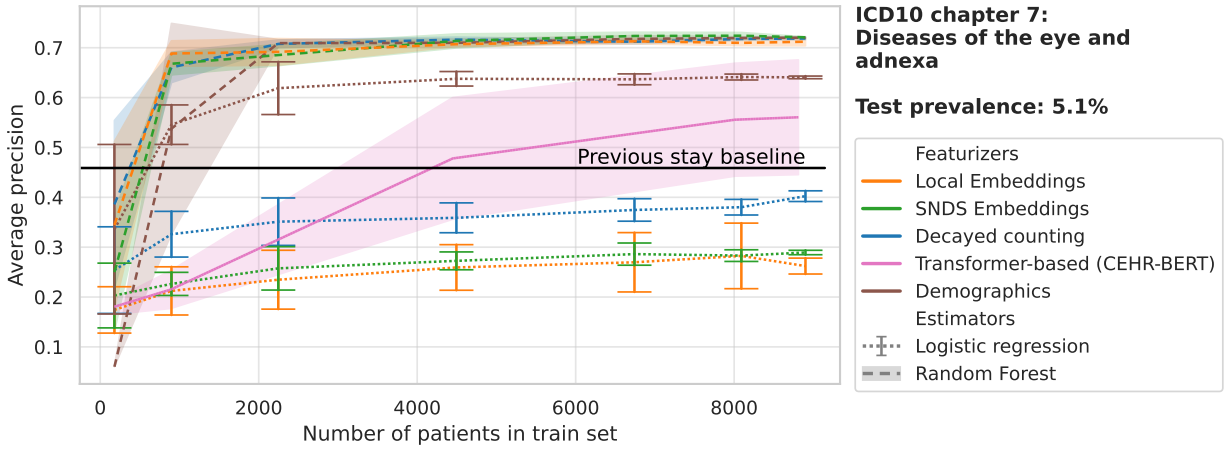


Fig. C.22. ICD10 chapter 7

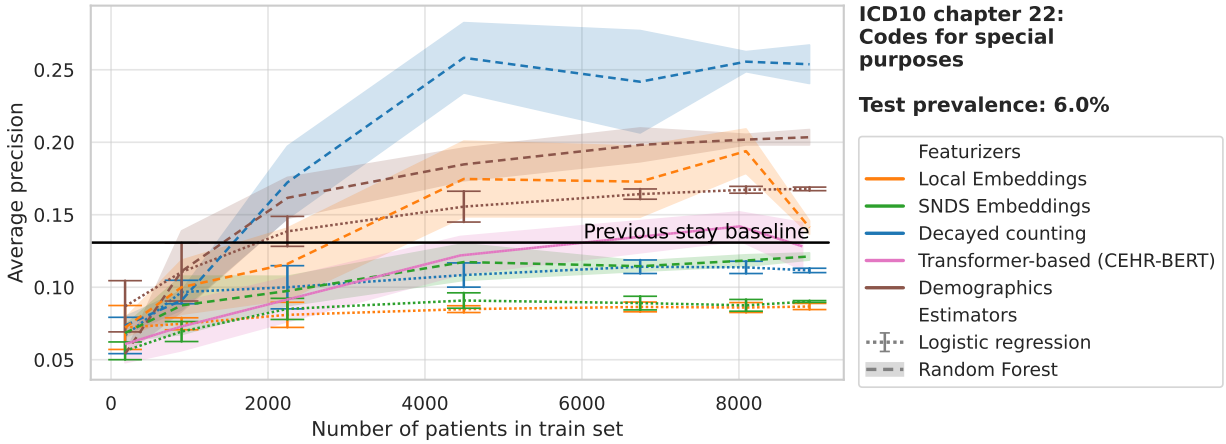


Fig. C.23. ICD10 chapter 22

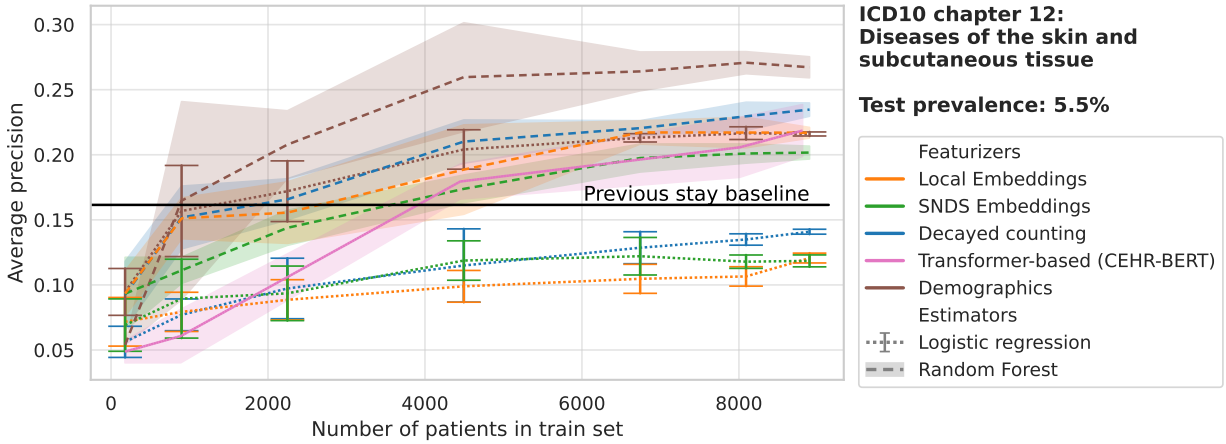


Fig. C.24. ICD10 chapter 12

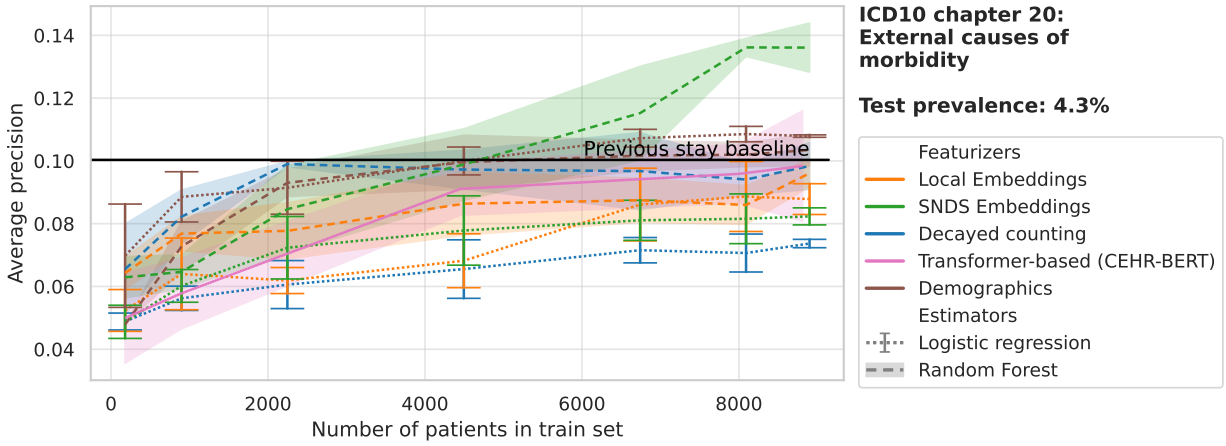


Fig. C.25. ICD10 chapter 20

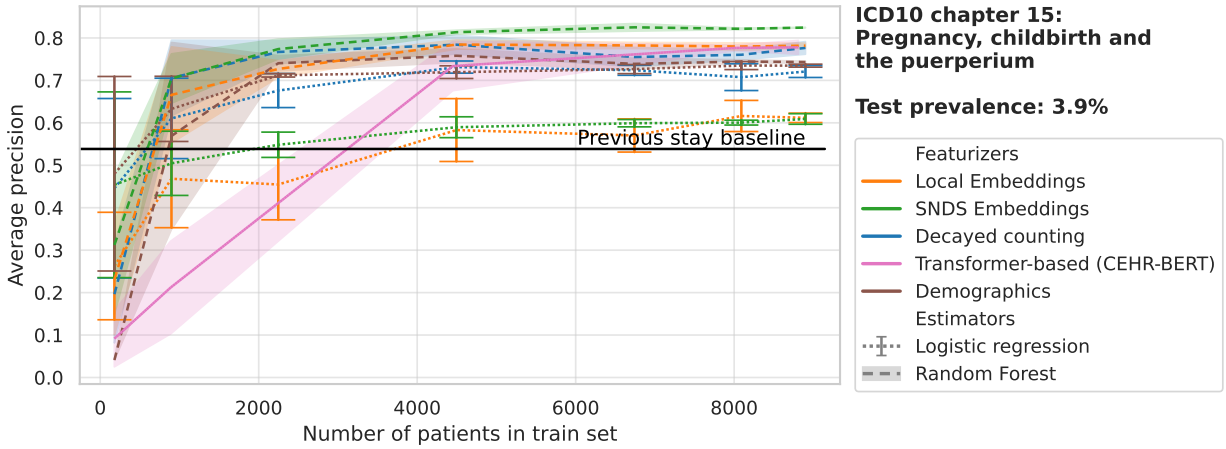


Fig. C.26. ICD10 chapter 15

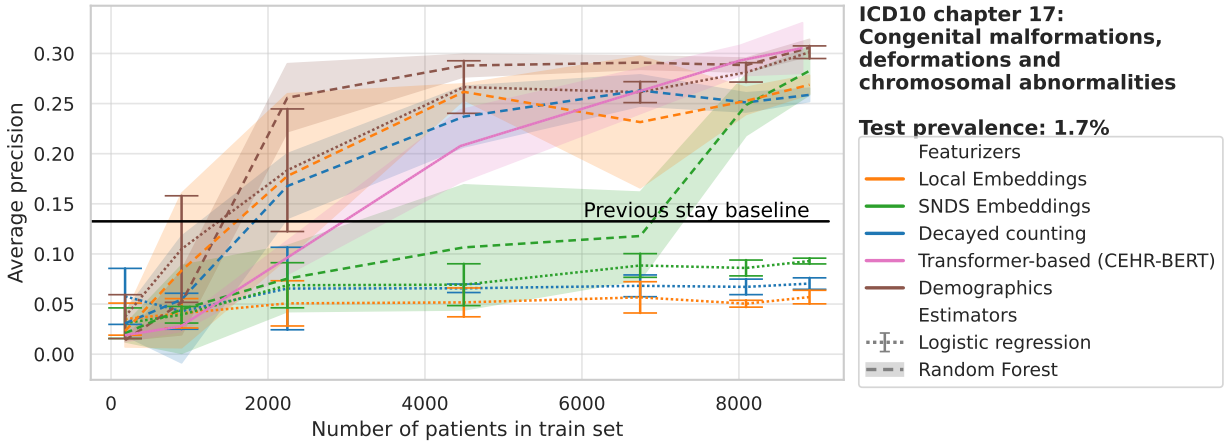


Fig. C.27. ICD10 chapter 17

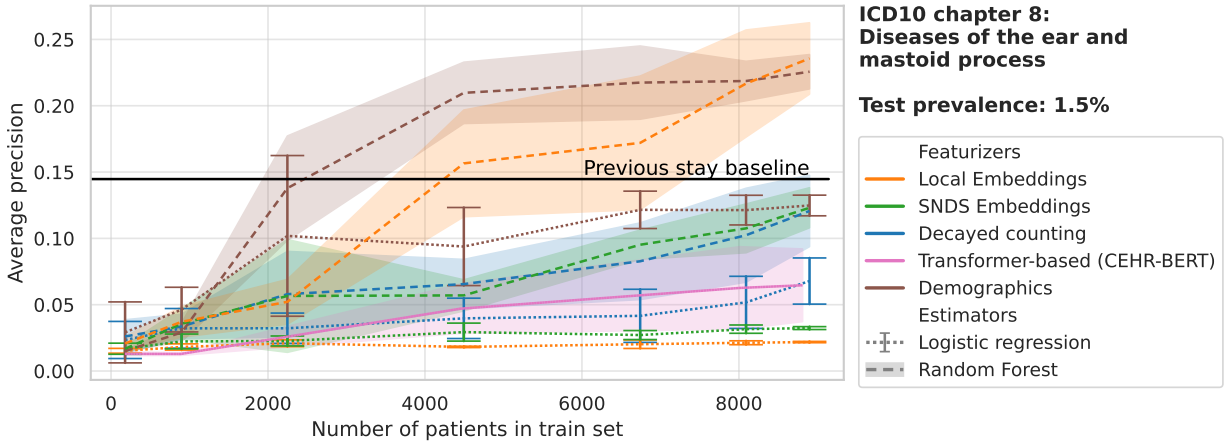


Fig. C.28. ICD10 chapter 8

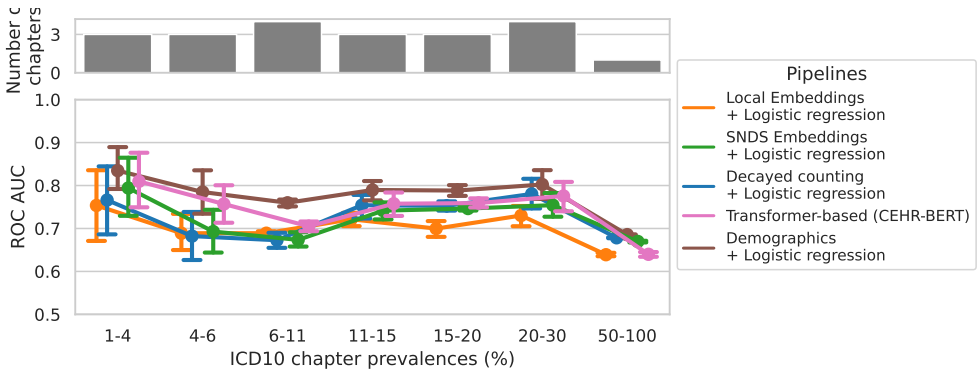


Fig. C.29. Bigger target prevalences yield better ROC AUC. The different chapters are binned in prevalence bins. The estimator used for this plot is a penalized linear model trained on the full effective train set. Each box contour represents Q1-Q3 inter-quartile range.

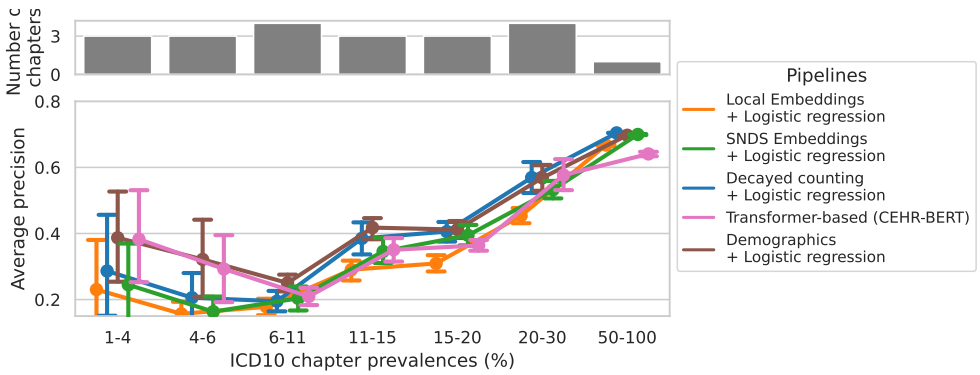


Fig. C.30. Bigger target prevalences yield better AUPRC. The different chapters are binned in prevalence bins. The estimator used for this plot is a penalized linear model trained on the full effective train set.

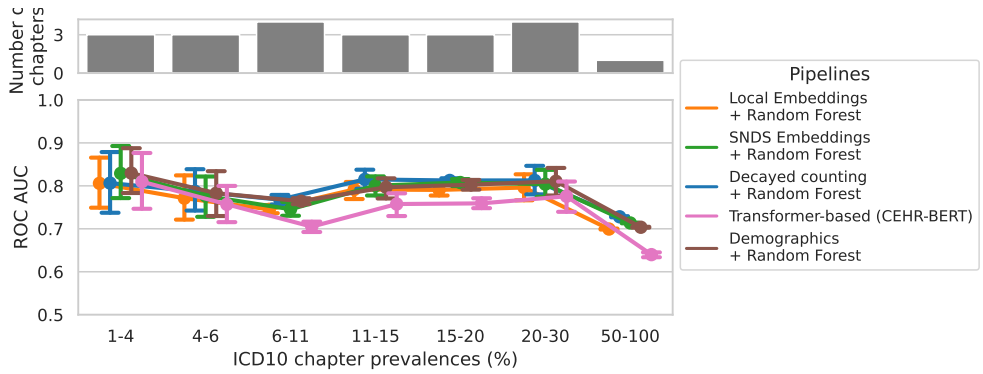


Fig. C.31. Bigger target prevalences yield better ROC AUC. The different chapters are binned in prevalence bins. The estimator used for this plot is random forest trained on the full effective train set.

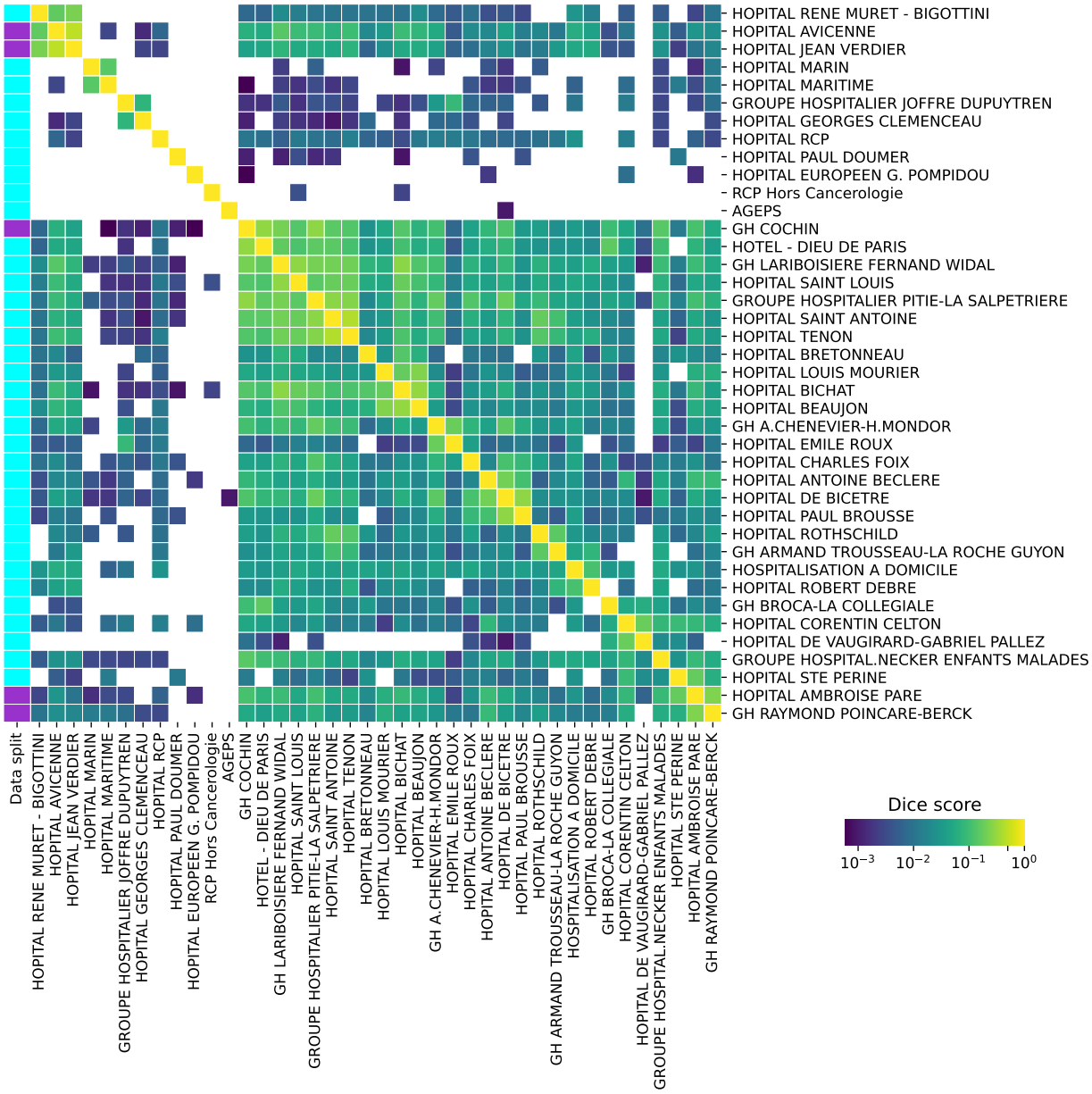


Fig. C.32. Hospitals clustered by average euclidean distance in common patients from the LOS cohort. Scale in dice score: $2 \cdot |A \cap B|/(|A| + |B|)$. Test set hospitals appear in purple on the left.

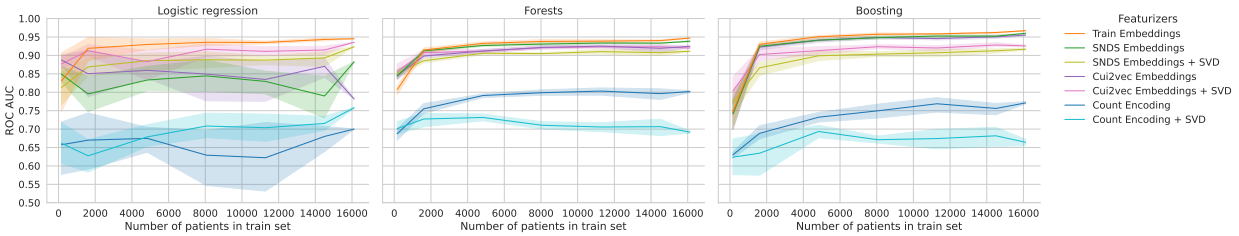


Fig. C.33. LOS task, transfer between hospitals with 2100 codes only, ROC AUC: The local, SNDS or cui2vec embeddings with forest and boosting are all equivalent. For logistic regression, the local embeddings remain the best performing method. We also compare the effect of reducing the dimension by applying a Singular Value Decomposition with 30 components kept. This increases a little bit the performance for logistic regression but decreases the performance of the other estimators.

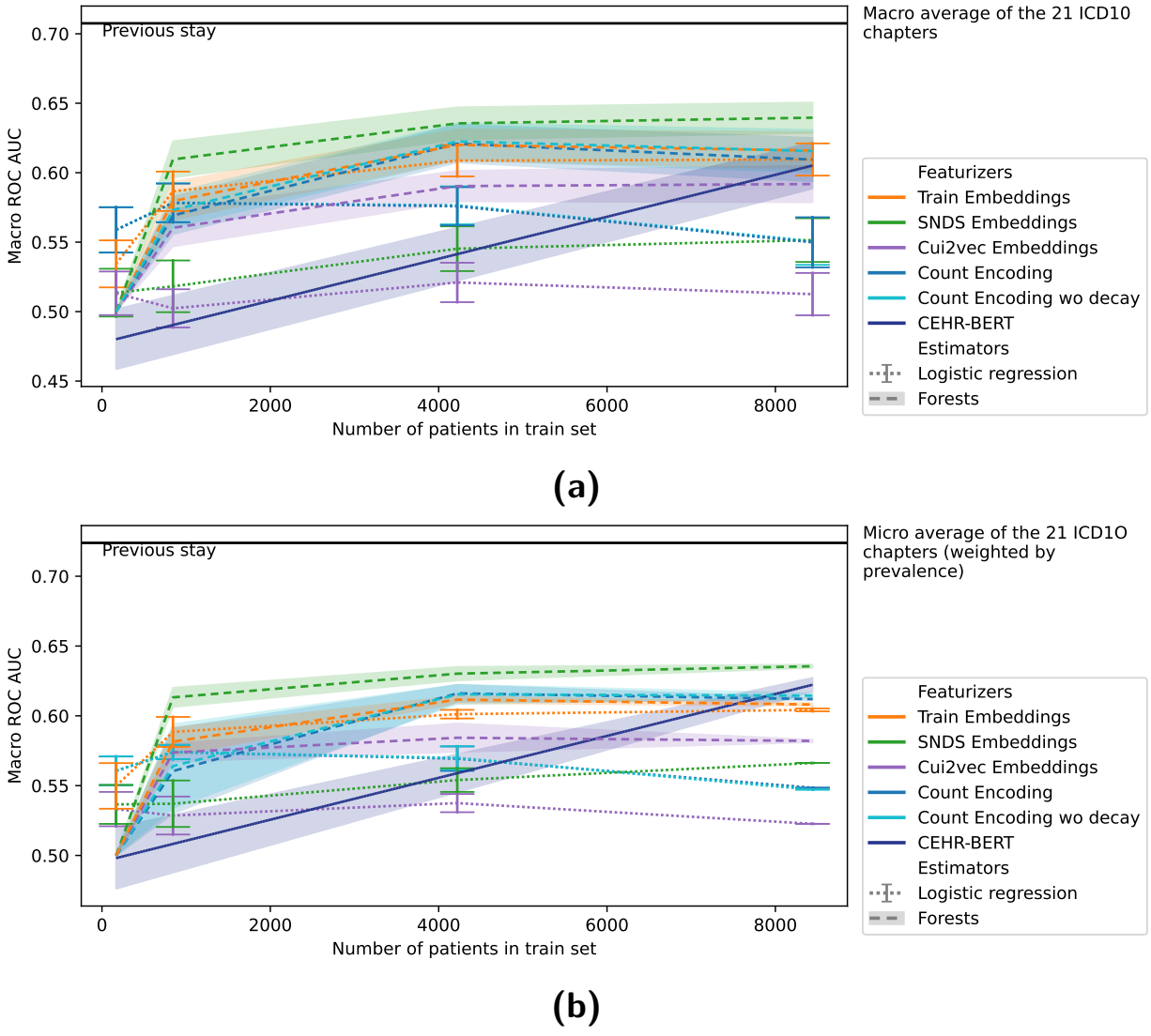


Fig. C.34. Prognosis task, transfer between hospitals, weighted average (by chapter prevalence) and macro average ROC AUC: The black vertical line shows the results from predicting the next diagnoses if they appear in the last visits in the observation period.

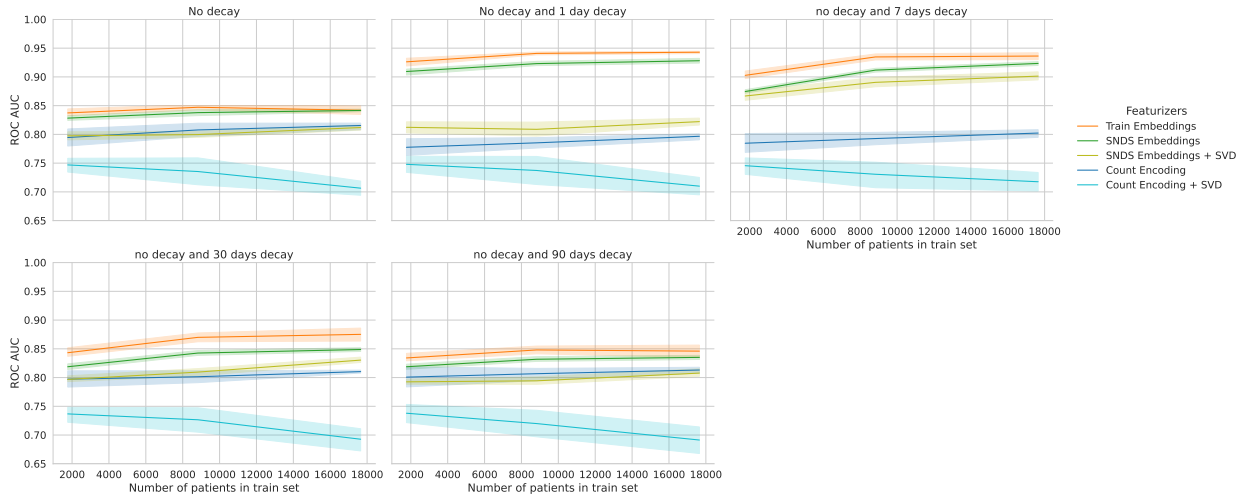


Fig. C.35. LOS task, random test set, ROC AUC, estimator is random forest: Concatenating a one day decay add 10 points of ROC AUC to the embedding pipeline.

(a) AP-HP **(b)** SNDS

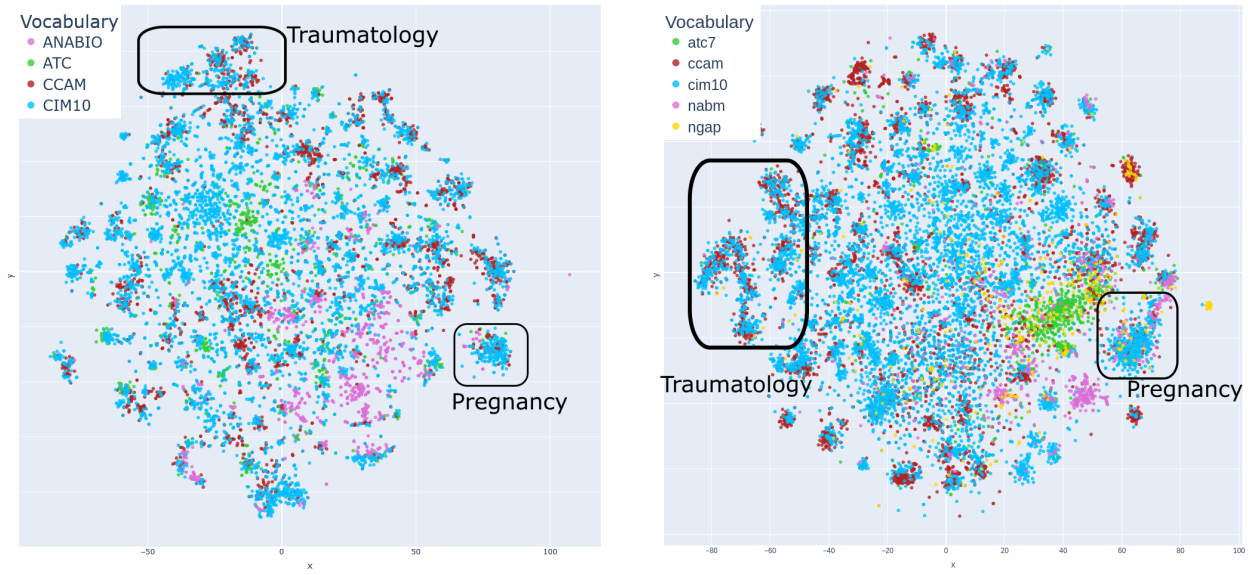


Fig. C.36. TSNE projection of medical event embeddings. Each point is a projection in 2D of the embedded vector for a given medical concept: a) in the AP-HP Clinical Data Warehouse (200, 000 random patients extractions), b) in the French Medical Claims (SNDS). Colors correspond to different medical vocabularies: drugs in green, billing diagnoses in blue, billing procedures in red, biology in pink (different vocabulary for AP-HP and SNDS), general practitioner (GP) activity in yellow. Interactive versions of these plots are available at: `https://straymat.gitlab.io/event2vec/visualizations.html`

Appendix D

Chapter 4

D.1 Motivating example: Failure of predictive models to predict mortality from pretreatment variables

To illustrate how machine learning frameworks can fail to inform decision making, we present a motivating example from MIMIC-IV. Using the same population and covariates as in the main analysis (described in Table D.5), we train a predictive model for 28-day mortality. We split the data into a training set (80%) and a test set (20%). The training set uses the last measurements from the first 24 hours, whereas the validation set only uses the last measurements before the administration of crystalloids. We split the train set into a train and a validation set. We fit a HistGradientBoosting classifier [1] on the train set and evaluate the performance on the validation set and on the test set. We see good area under the Precision-recall curve (PR AUC) on the validation set, but a deterioration of 10 points on the test set (Figure D.1a). The same is seen in Figure D.1b when measuring performance with Area Under the Curve of the Receiving Operator Characteristic (ROC AUC). In the contrary, a model trained on pre-treatment features yield competitive performance. This failure illustrates well the shortcuts on which predictive models could rely to make predictions. A clinically useful predictive model should support decision making –in this case, addition of albumin to crystalloids– rather than maximizing predictive performance. In this example, causal thinking would have helped to identify the bias introduced by post-treatment features. In fact, these features should not be included in a causal analysis since they are post-treatment colliders.

D.2 Estimation of Treatment effect with MIMIC data

We searched for causal inference studies in MIMIC using PubMed and Google scholar with the following search terms ((MIMIC-III OR MIMIC-IV) AND (causal inference OR treatment effect)). We retained eleven treatment effect studies clearly following the PICO framework:

- Liu et al., 2021 studied the effect of High-flow nasal cannula oxygen (HFNC) against noninvasive mechanical ventilation on 801 patients with hypoxemia during ventilator weaning on 28-day mortality. They used propensity score matching, and found non-negative effects as previous RCTs reported – though those were focused on reintubation as the main outcome (Stéphan et al., 2015; Hernandez et al., 2016).
- Yarnell et al., 2023 studied the effect of lower hypoxemia vs higher hypoxemia thresholds for the initiation of invasive ventilation (defined with saturation-to-inspired oxygen ratio (SF)) for 3,357 patients from MIMIC receiving inspired oxygen fraction >= 0.4 on 28-day moratlity. Using bayesian G-computation (time-varying treatment model with gaussian process and outcome-model with BART, taking the treatment model

[1] https://scikit-learn.org/stable/modules/ensemble.html#histogram-based-gradient-boosting

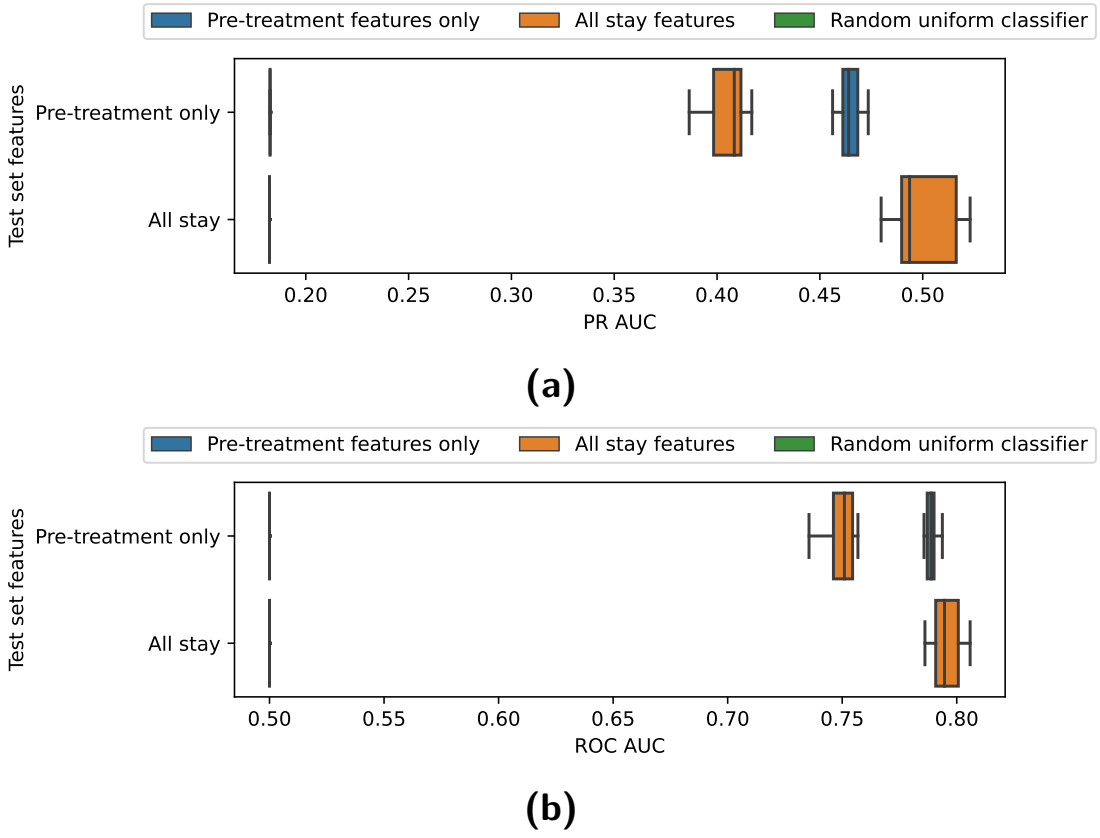


Fig. D.1. Failure to predict 28-day mortality from a model fitted on pre-treatment variables. The model is trained on the last features from the whole stay and tested on two validation sets: one with all stay features and one with last features before crystalloids administration (Pre-treatment only). The all-stay model performance markedly decreases in the pre-treatment only dataset.

as entry), they found protective effects for initialization at low hypoxemia. However, when externally validation their findings in the AmsterdamUMCdb dataset, they found the highest mortality probability for patients with low hypoxemia. Authors concluded that their model was heavily dependent on clinical context and baseline caracteristics. There might be some starting-time bias in this study since it is really close

- Hsu et al., 2015 studied the effect of indwelling arterial catheters (IACs) vs non-IAC for 1,776 patients who are mechanically ventilated and did not require vasopressor support on 28-day mortality. They used propensity score matching and found no effect. A notebook based on google cloud access to MIMIC-IV replicating the study is available here.
- Feng et al., 2018 studied the effect of transthoracic echocardiography vs no intervention for 6,361 patients with sepsis on 28-day mortality. They used IPW, PSM, g-formula and a doubly robust estimation. The propensity score was modeled with boosting and the outcome model with a logistic regression. They found a significant positive reduction of mortality (odd ratio 0.78, 95% CI 0.68-0.90). Study code is open source.
- Gani et al., 2023 studied the effect of liberal –target SpO2 greater than 96%– vs conservative oxygenation –target SpO2 between 88-95%– in 4,062 mechanically ventilated patients on 90-day mortality. They found an advantage of the liberal strategy over liberal (ATE=0.13) by adjusting on age and apsii. This is not consistent with previous RCTs where no effects have been reported (Panwar et al., 2016; Mackle et al., 2019).

- Shahn et al., 2020 studied the effect of fluid-limiting treatment –caped between 6 and 10 L– vs no cap on fluid administration strategies for 1,639 sepsis patients on 30 day-mortality. Using a dynamic Marginal Structural Model with IPW, they found a protective effect of fluid-limitation on ATE -0.01 (95%CI -0.016, -0.03). This is somehow concordant with the RIFTS RCT that found no effect of fluid limitation (Corl et al., 2019) and two previous meta-analyses (Malbrain et al., 2014; Meyhoff et al., 2020).

- Chinaeke et al., 2021 studied the effect of statin use prior to ICU admission vs absence of pre-ICU prescription for 8,200 patients with sepsis on 30-day mortality. Using AIPW (no estimator reported) and PSM (logistic regression), they found a decrease on mortality (ATE -0.039, 95%CI -0.084, -0.026). This partly supports previous findings in Propensity Matching bases observational studies (Lee et al., 2017; Kyu Oh et al., 2019). But all RCTs (National Heart; Network, 2014; Singh et al., 2017) found no improvement for sepsis (not pre-admission administration though). The Wan et al., 2014 meta-analysis concludes that there is lack of evidence for the use of statins in sepsis with inconsistent results between RCTs (no effect) and observational studies (protective effect).

- Adibuzzaman et al., 2019 studied the effect of higher vs lower positive end-expiratory pressures (PEEP) in 1,411 patients with Acute Respiratory Distress Syndrome (ARDS) syndrome on 30 day mortality. Very few details on the methods were reported, but they found a protective effect for higher PEEP consistent results from a target trial (National Heart; Network, 2004).

- Adibuzzaman et al., 2019 also studied the effect of early use of a neuromuscular blocking agent vs placebo in 752 patients moderate-severe ARDS on 30 day mortality. Very few details on the methods were reported, but they found a protective effect for the use of a neuromuscular blocking agent, consistent with the results from a target trial (Papazian et al., 2010).

- Zhou et al., 2021b studied the administration of a combination of albumin within the first 24-h after crystalloids vs crystalloids alone for 6,641 patients with sepsis on 28-day mortality. Using PSM, they found protective effect of combination on mortality, but insist on the importance of initialization timing. This is consistent with Xu et al., 2014, who found a non-significant trend in favor of albumin used for severe sepsis patients and a significant reduction for septic shock patients, both on 90-day mortality. These results are aligned with Caironi et al., 2014 that found no effect for severe sepsis patient but positive effect for septic shock patients.

- Wang et al., 2023a studied early enteral nutrition (EN) –<=53 ICU admission hours– vs delayed EN for 2,364 patients with sepsis and EN on acute kidney injury. With PSM, IPW and g-formula (logistic estimator each time), they found a protective effect (OR 0.319, 95%CI 0.245, 0.413) of EEN.

These eleven studies mainly used propensity score matching (6) and IPW (4), two of them used Double robust methods, and only one included a non-linear estimator in either the outcome or the treatment model. None of them performed a vibration analysis on the confounders selection or the feature transformations. They have a strong focus on sepsis patients. Only four of them found concordant results with previous RCTs (Liu et al., 2021; Shahn et al., 2020; Adibuzzaman et al., 2019).

D.3 Target trials proposal suitable to be replicated in MIMIC

Celi et al., 2016 suggested the creation of a causal inference database based on MIMIC with a list of replicable RCTs, which has not been accomplished yet. We reviewed the following RCTs, which could be replicated within the MIMIC-IV database. Table D.1 details the sample sizes of the eligible, control and treated populations for the identified RCTs.

Trial name	Criteria description	Number of patients	Criteria status	Implemented	Meta-analysis or target RCT 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)
	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), (Li et al., 2020a), (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	✓	

Table D.1. Eligibility criteria and resulting populations for potential target trials in MIMIC-IV.

D.4 Major causal-inference methods

D.4.1 Causal estimators: When to use which method ?

Difference in Mean, (Splawa-Neyman et al., 1990) This is the most intuitive method to estimate the ATE. It processes by comparing the mean of the outcome between both populations:

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{A_i=1} Y_i(1) - \frac{1}{n_0} \sum_{A_i=0} Y_i(0) \tag{D.1}$$

In the case of randomization, we have an independence between the treatment and the potential outcomes: $\left\{Y_i^{(0)}, Y_i^{(1)}\right\} \perp\!\!\!\perp A_i$. We can thus show that this estimator is unbiased, ie. $\mathbb{E}[\hat{\tau}_{DM}] = \tau$:

$$
\begin{aligned}
\mathbb{E}\left[Y \mid A=1\right] &= \mathbb{E}\left[YA \mid A=1\right] && \text{Binary nature of } A \\
&= \mathbb{E}\left[Y^{(1)}A^2 \mid A=1\right] && \text{Consistency } Y_i = A_i\, Y_i^{(1)} + (1-A_i)\, Y_i^{(0)} \\
&= \mathbb{E}\left[Y^{(1)} \mid A=1\right] && \text{Binary nature of } A \\
&= \mathbb{E}\left[Y^{(1)}\right] && \text{Randomization,}
\end{aligned}
$$

The same calculation with $A = 0$ give the result. Note that this result does not apply to observational data, where we only have conditional randomization: $\left\{Y_i^{(0)}, Y_i^{(1)}\right\} \perp\!\!\!\perp A_i | X_i$. Thus, we should rely on more elaborated estimation strategies if the data is observational.

G-formula This estimator is also called conditional mean regression (Wendling et al., 2018b), g-computation (Robins; Greenland, 1986), or Q-model (Snowden et al., 2011). It is directly modeling the outcome, also referred to as the response surface: $\mu_{(a)}(x) = \mathbb{E}\left(Y \mid A=a, \mathbf{X}=x\right)$

Using an outcome estimator to learn a model for the response surface $\hat{\mu}$ (e.g., a linear model), the ATE estimator is an average over the n samples:

$$
\hat{\tau}_G(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(x_i, 1) - \hat{\mu}(x_i, 0) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_{(1)}(x_i) - \hat{\mu}_{(0)}(x_i) \tag{D.2}
$$

This estimator is unbiased if the model of the conditional response surface $\hat{\mu}_{(a)}$ is well-specified. This approach assumes than $Y(a) = \mu_a(X) + \epsilon_a$ with $\mathbb{E}[\epsilon|X] = 0$. The main drawback is the extrapolation of the learned outcome estimator from samples with similar covariates X but different intervention A. Figure D.2 shows the intuition of g-formula on a example with a single confounder.

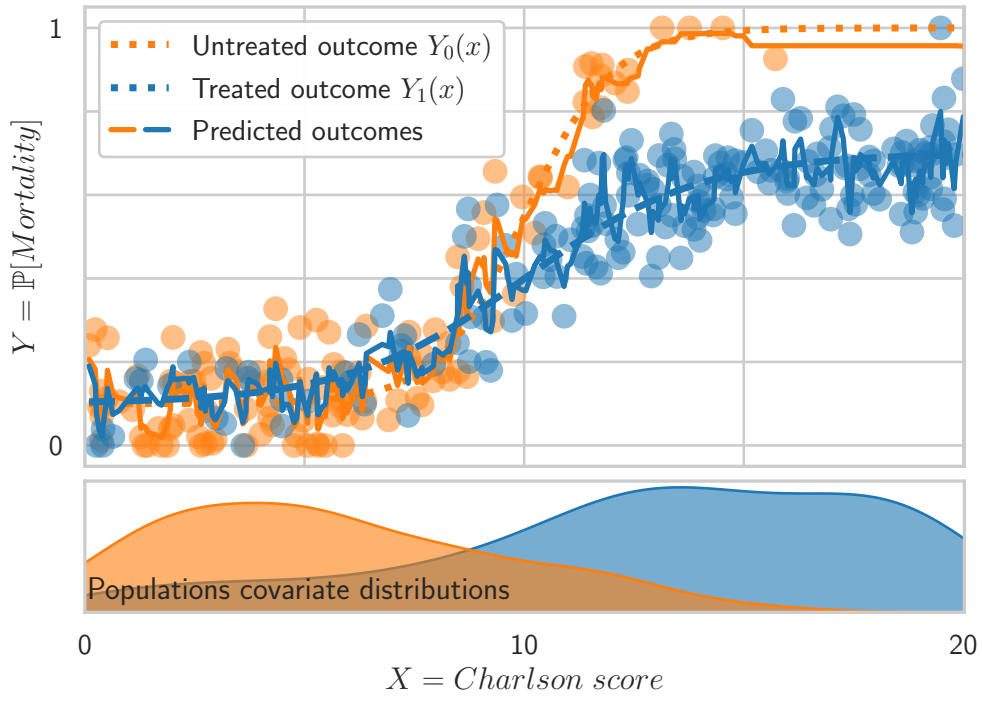


Fig. D.2. G-formula fit a model on the outcome against the confounders and the treatment.

Proof 1 (Unbiasdness of the G-formula, (Robins; Greenland, 1986)) *Suppose, we have access to the oracle mean response surfaces $\mu_{(a)}$, then the finite sample estimator is $\hat{\tau}_G(\mu) = \frac{1}{n}\sum_{i=1}^{n} \mu_{(1)}(x_i) - \mu_{(0)}(x_i)$*

$$
\begin{aligned}
\tau(x) &= \mathbb{E}\left[Y(1) - Y(0) \mid X = x\right] && \textit{Definition of CATE} \\
&= \mathbb{E}\left[Y(1) \mid X = x, A = 1\right] - \mathbb{E}\left[Y(0) \mid X = x, A = 0\right] && \textit{Ignorability, eq. 1} \\
&= \mathbb{E}\left[AY(1) \mid X = x, A = 1\right] - \mathbb{E}\left[(1-A)Y(0) \mid X = x, A = 0\right] && \textit{Binary nature of } A \\
&= \mathbb{E}[Y \mid X = x, A = 1] - \mathbb{E}[Y \mid X = x, A = 0] && \textit{Consistency, eq. 3} \\
&= \mathbb{E}[\hat{\tau}_G(\mu)]
\end{aligned}
$$

Note that without further parametric assumption on the conditional response surface $\mu_{(a)}$, the asymptotic properties of this estimator are unknown.

Propensity Score Matching (PSM) To avoid confounding bias, the ignorability assumption 1) requires to contrast treated and control outcomes only between comparable patients with respect to treatment allocation probabilities. A simple way to do this is to group patients into bins, or subgroups, of similar confounders and contrast the two population outcomes by matching patients inside of these bins (Stuart, 2010). However, the number of confounder bins grows exponentially with the number of variables. Rosenbaum; Rubin, 1983 proved that matching patients on the individual probabilities to receive treatment –propensity scores– is sufficient to verify ignorability. PSM is a conceptually simple method, but has delicate parameters to tune such as choosing a model for the propensity score, deciding what is the maximum distance between two potential matches (the caliper width), the number of matches by sample, and matching with or without replacement. It also prunes data not meeting the caliper width criteria, and suffers form high estimation variance in highly-dimensional data where extreme propensity weights are common. Finally, the simple bootstrap confidence intervals are not theoretically grounded (Abadie; Imbens, 2008), making PSM more difficult to use for applied practitioners.

Inverse Propensity Weighting (IPW) A simple alternative to propensity score matching is to weight the outcome by the inverse of the propensity score: Inverse Propensity Weighting (Austin; Stuart, 2015). It relies on the same idea than matching but builds automatically balanced population by reweighting the outcomes with the propensity score model $\hat{e}$ to estimate the ATE:

$$
\hat{\tau}_{IPW}(\hat{e}) = \frac{1}{n}\sum_{i=1}^{N} \frac{A_i Y_i}{\hat{e}(X_i)} - \frac{(1-A_i)Y_i}{(1-\hat{e}(X_i))} \tag{D.3}
$$

This estimate is unbiased if $\hat{e}$ is well-specified. IPW suffers from high variance if some weights are too close to 0 or 1. In high-dimensional cases where poor overlap between treated and control is common, one can clip extreme weights to limit estimation instability. Figure D.3 illustrates the intuition of IPW on an example with a single confounder.

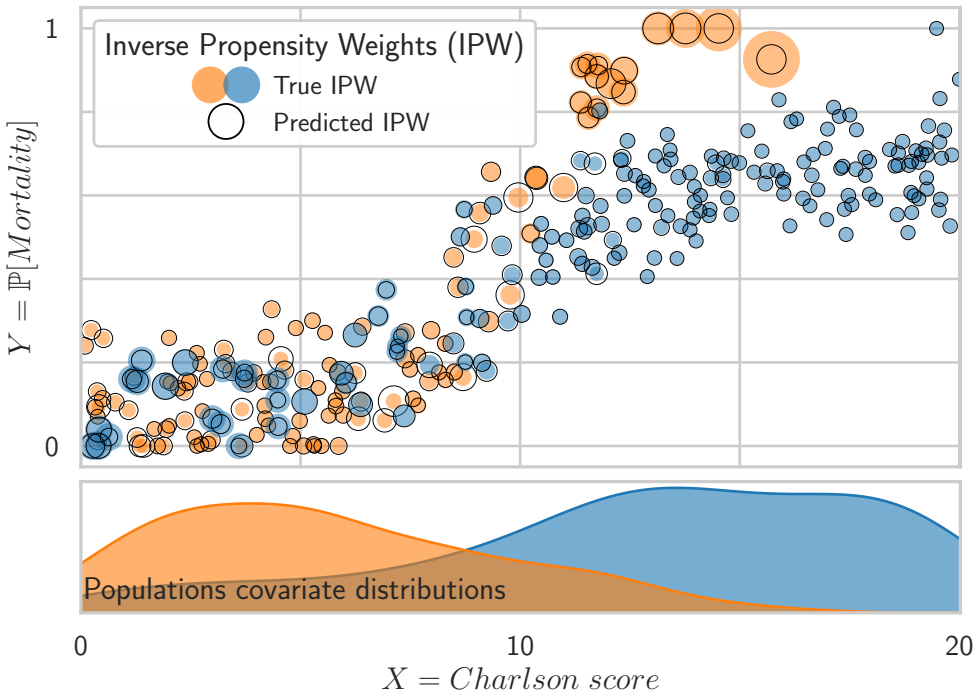


Fig. D.3. IPW reweights individual to build a balanced population with respect to the treatment. It gives more weight to treated samples having a small chance to receive the treatment (bigger blue points on the left), and more weight to controls having a small chance to not receive the treatment (bigger orange points on the right).

Proof 2 (Unbiasdness of IPW, (Rosenbaum; Rubin, 1983)) *As an intermediary result, we need to show that the propensity score e is a balancing score, ie. if ignorability (1) holds then $\{Y(1), Y(0)\} \perp\!\!\!\perp A | e(X)$. Thus, need to prove that $\mathbb{P}[A = 1 \mid Y(1), Y(0), e(X)] = \mathbb{P}[A = 1 \mid e(X)]$ First remark that $\mathbb{P}[A = 1 \mid e(X)] = e(X)$:*

$$
\begin{aligned}
\mathbb{P}[A = 1 \mid e(X)] &= \mathbb{E}\Big[\mathbb{P}[A = 1 \mid X, e(X)] \mid e(X)\Big] && \textit{Law of total expectation} \\
&= \mathbb{E}\Big[\mathbb{P}[A = 1 \mid X] \mid e(X)\Big] && \textit{X contains all information of e(X)} \\
&= \mathbb{E}\Big[e(X) \mid e(X)\Big] && \textit{by definition of the propensity score} \\
&= e(X) && \textit{e(X) contains all information of e(X)}
\end{aligned}
$$

Thus, we only have to prove that $\mathbb{P}[A = 1 \mid Y(1), Y(0), e(X)] = e(X)$:

$$
\begin{aligned}
&\mathbb{P}[A = 1 \mid Y(1), Y(0), e(X)] \\
&= \mathbb{E}\Big[\mathbb{P}[A = 1 \mid Y(1), Y(0), X, e(X)] \mid Y(1), Y(0), e(X)\Big] && \textit{Law of total expectation} \\
&= \mathbb{E}\Big[\mathbb{P}[A = 1 \mid Y(1), Y(0), X] \mid Y(1), Y(0), e(X)\Big] && \textit{X contains all information of e(X)} \\
&= \mathbb{E}\Big[\mathbb{P}[A = 1 \mid X] \mid Y(1), Y(0), e(X)\Big] && \textit{ignorability} \\
&= \mathbb{E}\Big[e(X) \mid Y(1), Y(0), e(X)\Big] && \textit{by definition of the propensity score} \\
&= e(X) && \textit{e(X) contains all information of e(X)}
\end{aligned}
$$

Suppose that we have access to the oracle propensity score e, then the finite sample estimator is $\hat{\tau}_{IPW}(e) = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{e(X_i)} - \frac{(1-A_i)Y_i}{1-e(X_i)}$.

$$
\begin{aligned}
\mathbb{E}\Big[\frac{AY}{e(X)}\Big] &= \mathbb{E}\Big[\frac{A^2Y(1)}{e(X)}\Big] && \textit{Consistency, eq. 3}\\
&= \mathbb{E}\Big[\frac{A^2Y(1)}{e(X)}\Big] && \textit{Binary nature of } A\\
&= \mathbb{E}\Big[\mathbb{E}\Big[\frac{AY(1)}{e(X)}|e(X)\Big]\Big] && \textit{Law of total expectation}\\
&= \mathbb{E}\Big[\mathbb{E}\Big[\frac{A}{e(X)}|e(X)\Big]\mathbb{E}\Big[Y(1)|e(X)\Big]\Big] && \textit{Ignorability and } e \textit{ is a balancing score}\\
&= \mathbb{E}[Y(1)] && \textit{since} e(X) = \mathbb{P}[A=1|X]
\end{aligned}
$$

Doing the same calculation for $A = 0$ gives the result.

Doubly Robust Learning, DRL It is also called Augmented Inverse Probability Weighting (AIPW) (Robins et al., 1994).

The underlying idea of DRL is to combine the G-formula and IPW estimators to protect against a mis-specification of one of them. It first requires to estimate the two nuisance parameters: a model for the intervention $\hat{e}$ and a model for the outcome f.

$$
\begin{aligned}
\widehat{\tau}_{AIPW} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}\left(x_i\right) - \hat{\mu}_{(0)}\left(x_i\right) + \frac{a_i}{\hat{e}\left(x_i\right)}\left(y_i - \hat{\mu}_{(1)}\left(x_i\right)\right) - \frac{1-a_i}{1-\hat{e}\left(x_i\right)}\left(y_i - \hat{\mu}_{(0)}\left(x_i\right)\right)\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_i y_i}{\hat{e}(x_i)} - \frac{(1-a_i)y_i}{1-\hat{e}(x_i)} + \hat{\mu}_{(1)}(x_i)\Big(1 - \frac{a_i}{\hat{e}(x_i)}\Big) - \hat{\mu}_{(0)}(x_i)\Big(1 - \frac{1-a_i}{1-\hat{e}(x_i)}\Big)\right)
\end{aligned}
$$

If one of the two nuisance is unbiased, the following ATE estimator is as well. We see it easily in the previous equation. Suppose that the outcome model is unbiased. Then, looking at the first line, we see that the first term is the G-formula estimator so it is unbiased. And the second term has mean zero since $\mathbb{E}[Y - \mu_{(A)} \mid X, A = 1] = 0$ and $\mathbb{E}[Y - \mu_{(0)} \mid X, A = 0] = 0$. Suppose now that the propensity score model is unbiased. Then, looking at the second line, we see that the first term is the IPW estimator so it is unbiased. And the second term has mean zero since $\mathbb{E}[1 - \frac{A}{e(X)}|X] = 0$ and $\mathbb{E}[1 - \frac{1-A}{1-e(X)}|X] = 0$.

More interestingly, despite the need to estimate two models, this estimator is more efficient in the sense that it converges quicker than single model estimators (Wager, 2020b). For this propriety to hold, one need to fit and apply the two nuisance models in a cross-fitting manner. This means that we split the data into K folds. Then for each fold, we fit the nuisance models on the K-1 complementary folds, and predict on the remaining fold.

To recover Conditional Treatment Effects from the AIPW estimator, Foster; Syrgkanis, 2019 suggested to regress the Individual Treatment Effect estimates from AIPW on potential sources of heterogeneity X^{cate}: $\hat{\tau} = \text{argmin}_{\tau\in\Theta}(\hat{\tau}_{AIPW}(X) - \tau(X^{cate}))$ for Θ some class of model (e.g., linear model).

Double Machine Learning (Chernozhukov et al., 2018b) It is also known as the R-learner (Nie; Wager, 2021). It is based on the R-decomposition, (Robinson, 1988), and the modeling of the conditional mean outcome, $m(x) = \mathbb{E}[Y|X = x]$ and the propensity score, $e(x) = \mathbb{E}[A = 1|X = x]$:

$$
y_i - m\left(x_i\right) = \left(a_i - e\left(x_i\right)\right)\tau\left(x_i\right) + \varepsilon_i \quad \text{with } \varepsilon_i = y_i - \varepsilon\left[4_i \mid x_i, a_i\right] \tag{D.4}
$$

Note that we can impose that the conditional treatment effect $\tau(x)$ only relies on a subset of the features, x^{cate} on which we want to study treatment heterogeneity.

From this decomposition, we can derive an estimation of the ATE τ, where the right hand-side term is the empirical R-Loss:

$$\hat{\tau}(\cdot) = \text{argmin}_{\tau} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left((y_i - m(x_i)) - (a_i - e(x_i)) \tau\left(x_i^{cate}\right) \right)^2 \right\} \tag{D.5}$$

The full procedure for R-learning is:

- Fit nuisances: $\hat{m}$ and $\hat{e}$
- Minimize the estimated R-loss eq.D.5, where the oracle nuisances (e, m) have been replaced by their estimated counterparts $(\hat{e}, \hat{m})$. Minimization can be done by regressing the outcome residuals weighted by the treatment residuals
- Get the ATE by averaging conditional treatment effect $\tau(x^{cate})$ over the population

This estimator has also the doubly robust proprieties described for AIPW. it should have less variance than AIPW since it does not use the propensity score in the denominator.

D.4.2 Statistical considerations when implementing estimation

Counterfactual prediction lacks off-the-shelf cross-fitting estimators Doubly robust methods use cross-fit estimation of the nuisance parameters, which is not available off-the-shelf for IPW and T-Learner estimators. For reproducibility purposes, we did not reimplement internal cross-fitting for treatment or outcome estimators. However, when flexible models such as random forests are used, a fairer comparison between single and double robust methods should use cross-fitting for both. This lack in the scikit-learn API reflects different needs between purely predictive machine learning focused on generalization performance and counterfactual prediction aiming at unbiased inference on the input data.

Good practices for imputation not implemented in EconML Good practices in machine learning recommend to input distinctly each fold when performing cross-fitting [2]. However, EconML estimators test for missing data at instantiation preventing the use of scikit-learn imputation pipelines. We thus have been forced to transform the full dataset before feeding it to causal estimators. An issue mentioning the problem has been filed, so we can hope that future versions of the package will comply with best practices. [3]

Bootstrap may not yields the more efficient confidence intervals To ensure a fair comparison between causal estimators, we always used bootstrap estimates for the confidence intervals. However, closed form confidence intervals are available for some estimators – see Wager, 2020b for IPW and AIPW (DRLeaner) variance estimations. These formulas exploit the estimator properties, thus tend to have smaller confidence intervals. On the other hand, they usually do not include the variance of the outcome and treatment estimators, which is naturally dealt with in bootstrap confidence intervals. Closed form confidence intervals are rarely implemented in the packages. Dowhy did not implement the well-known confidence interval method for the IPW estimator, nor did EconML for the AIPW confidence intervals.

Bootstrap was particularly costly to run for the EconML doubly robust estimators (AIPW and Double ML), especially when combined with random forest nuisance estimators (from 10 to 47 min depending on the aggregation choice and the estimator). See Table D.2 for details.

[2] https://scikit-learn.org/stable/modules/compose.html#combining-estimators
[3] https://github.com/py-why/EconML/issues/664

	estimation_method	compute_time (sec)	outcome_model	event_aggregations
2	LinearDML	1128	Forests	['first', 'last']
3	backdoor.propensity_score_matching	200	Forests	['first', 'last']
4	backdoor.propensity_score_weighting	86	Forests	['first', 'last']
5	TLearner	284	Forests	['first', 'last']
6	LinearDRLearner	2855	Forests	['first', 'last']
7	LinearDML	50	Regularized LR	['first', 'last']
8	backdoor.propensity_score_matching	128	Regularized LR	['first', 'last']
9	backdoor.propensity_score_weighting	6	Regularized LR	['first', 'last']
10	TLearner	7	Regularized LR	['first', 'last']
11	LinearDRLearner	81	Regularized LR	['first', 'last']

Table D.2. Compute times for the different estimation methods with 50 bootstrap replicates.

D.4.3 Packages for causal estimation in the python ecosystem

We searched for causal inference packages in the python ecosystem. The focus was on the identification methods. Important features were ease of installation, sklearn estimator support, sklearn pipeline support, doubly robust estimators, confidence interval computation, honest splitting (cross-validation), Targeted Maximum Likelihood Estimation. These criteria are summarized in D.3. We finally chose EconML despite lacking `sklearn._BaseImputer` support through the `sklearn.Pipeline` object as well as a TMLE implementation.

The zEpid package is primarily intended for epidemiologists. It is well documented and provides pedagogical tutorials. It does not support sklearn estimators, pipelines and honest splitting.

EconML implements almost all estimators except propensity score methods. Despite focusing on Conditional Average Treatment Effect, it provides all. One downside is the lack of support for scikit-learn pipelines with missing value imputers. This opens the door to information leakage when imputing data before splitting into train/test folds.

Dowhy focuses on graphical models and relies on EconML for most of the causal inference methods (identifications) and estimators. Despite, being interesting for complex inference –such as mediation analysis or instrumental variables–, we considered that it added an unnecessary layer of complexity for our use case where a backdoor criterion is the most standard adjustment methodology.

Causalml implements all methods, but has a lot of package dependencies which makes it hard to install.

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 D.3. Selection criteria for causal python packages

D.4.4 Hyper-parameter search for the nuisance models

We followed a two-step procedure to train the nuisance models (e.g., $(\hat{e}, \hat{\mu})$ for the AIPW causal estimator), taking inspiration from the computationally cheap procedure from Bouthillier

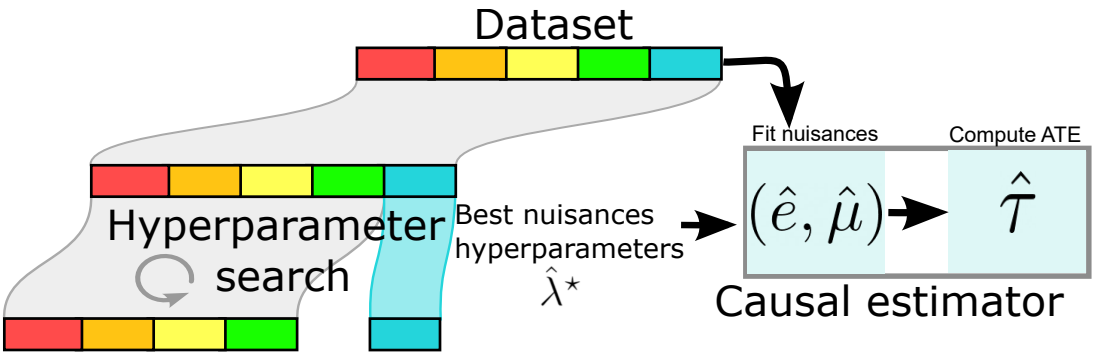


Fig. D.4. Hyper-parameter search procedure.

Estimator type	estimator	nuisance	Grid
Linear	LogisticRegression	treatment	{'C': logspace(-3, 2, 10)}
Linear	Ridge	outcome	{'alpha': logspace(-3, 2, 10)}
Forest	RandomForestClassifier	treatment	{'n_estimators': ['10', '100', '200'], 'max_depth': ['3', '10', '50']}
Forest	RandomForestRegressor	outcome	{'n_estimators': ['10', '100', '200'], 'max_depth': ['3', '10', '50']}

Table D.4. Hyper-parameter grid used during random search optimization.

et al., 2021a, section 3.3. First, for each nuisance model, we fit a random parameter search with 5-fold cross validation and 10 iterations on the full dataset. Each iteration fit a model with a random combination of parameters in a predefined grid, then evaluate the performance by cross-validation. The best hyper-parameters $\hat{\lambda}^{\star}$ are selected as the ones reaching the minimal score across all iterations. Then, we feed this parameters to the causal estimator. The single robust estimators (matching, IPW and TLearner) refit the corresponding estimator only once on the full dataset, then estimate the ATE. The doubly-robust estimators use a cross-fitting procedure (K=5) to fit the nuisances then estimate the ATE. Figure D.4 illustrates the procedure and Table D.4 details the hyper-parameters grid for the random search.

D.5 Computing resources

The whole project was run on a laptop running Ubuntu 22.04.2 LTS with the following hardware: CPU 12th Gen Intel(R) Core(TM) i7-1270P with 16 threads and 15 GB of RAM.

D.6 Selection flowchart

Figure D.5 details the selection flowchart for the emulated trial.

D.7 Complete description of the confounders for the main analysis

Figure D.5 detail the characteristics of the emulated trial population with all confounders used in our study.

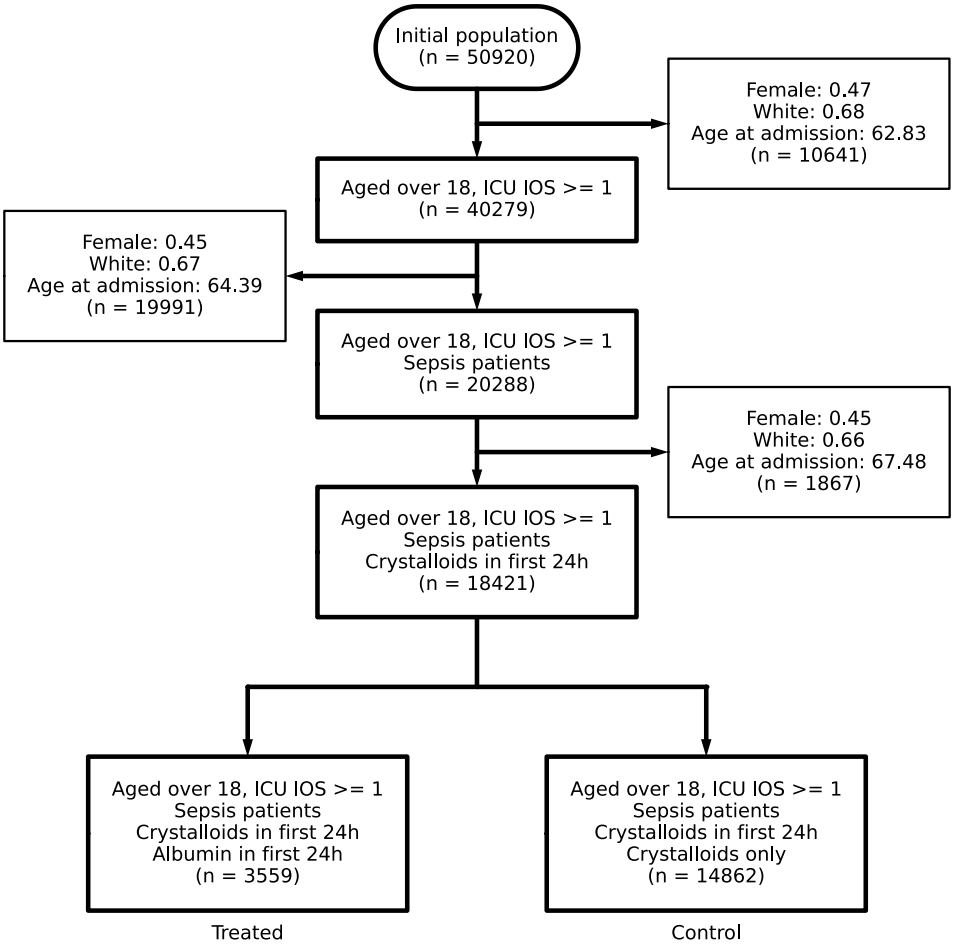


Fig. D.5. Selection flowchart on MIMIC-IV for the emulated trial.

D.8 Complete results for the main analysis

Compared to figure 4.7, we also report in figure D.6 the estimates for Causal forest estimators and other choices of feature aggregation (first and last).

D.9 Complete results for the Immortal time bias

Compared to figure 4.3, we also report in figure D.7 the estimates for Double Machine Learning, Inverse Propensity Weighting for both Random Forest and Ridge Regression. Feature aggregation was concatenation of first and last for all estimates.

D.10 Vibration analysis for aggregation

We conducted a dedicated vibration analysis on the different choices of features aggregation, studying the impact on the estimated ATE. We also studied if some choices of aggregation led to substantially poorer overlap.

We assessed overlap with two different methods. As recommended by (Austin; Stuart, 2015), we did a graphical assessment by plotting the distribution of the estimated. The treatment model hyper-parameters were chosen by random search, then predicted propensity

scores were obtained by refitting this estimator with cross-fitting on the full dataset.

As shown in Figure D.8, we did not find substantial differences between methods when plotting graphically the distribution of the estimated propensity score.

We also used normalized total variation (NTV) as a summary statistic of the estimated propensity score to measure the distance between treated and control population (Doutreligne; Varoquaux, 2023). This statistic varies between 0 – perfect overlap – and 1 – no overlap at all. Fig D.9 shows no marked differences in overlap as measured by NTV between aggregation choices, comforting us in our expert-driven choice of the aggregation: a concatenation of first and last feature observed before inclusion time.

	Missing	Overall	Cristalloids only	Cristalloids + Albumin	P-Value
n		18421	14862	3559	
Glycopeptide, n (%)		9492 (51.5)	7650 (51.5)	1842 (51.8)	
Beta-lactams, n (%)		5761 (31.3)	5271 (35.5)	490 (13.8)	
Carbapenems, n (%)		727 (3.9)	636 (4.3)	91 (2.6)	
Aminoglycosides, n (%)		314 (1.7)	290 (2.0)	24 (0.7)	
suspected_infection_blood, n (%)		170 (0.9)	149 (1.0)	21 (0.6)	
RRT, n (%)		229 (1.2)	205 (1.4)	24 (0.7)	
ventilation, n (%)		16376 (88.9)	12931 (87.0)	3445 (96.8)	
vasopressors, n (%)		9058 (49.2)	6204 (41.7)	2854 (80.2)	
Female, n (%)		7653 (41.5)	6322 (42.5)	1331 (37.4)	
White, n (%)		12366 (67.1)	9808 (66.0)	2558 (71.9)	
Emergency admission, n (%)		9605 (52.1)	8512 (57.3)	1093 (30.7)	
Insurance, Medicare, n (%)		9727 (52.8)	7958 (53.5)	1769 (49.7)	
myocardial_infarct, n (%)		3135 (17.0)	2492 (16.8)	643 (18.1)	
malignant_cancer, n (%)		2465 (13.4)	2128 (14.3)	337 (9.5)	
diabetes_with_cc, n (%)		1633 (8.9)	1362 (9.2)	271 (7.6)	
diabetes_without_cc, n (%)		4369 (23.7)	3532 (23.8)	837 (23.5)	
metastatic_solid_tumor, n (%)		1127 (6.1)	1016 (6.8)	111 (3.1)	
severe_liver_disease, n (%)		1289 (7.0)	880 (5.9)	409 (11.5)	
renal_disease, n (%)		3765 (20.4)	3159 (21.3)	606 (17.0)	
aki_stage_0.0, n (%)		7368 (40.0)	6284 (42.3)	1084 (30.5)	
aki_stage_1.0, n (%)		4019 (21.8)	3222 (21.7)	797 (22.4)	
aki_stage_2.0, n (%)		6087 (33.0)	4605 (31.0)	1482 (41.6)	
aki_stage_3.0, n (%)		947 (5.1)	751 (5.1)	196 (5.5)	
SOFA, mean (SD)	0	6.0 (3.5)	5.7 (3.4)	6.9 (3.6)	<0.001
SAPSII, mean (SD)	0	40.3 (14.1)	39.8 (14.1)	42.8 (13.6)	<0.001
Weight, mean (SD)	97	83.3 (23.7)	82.5 (24.2)	86.4 (21.2)	<0.001
temperature, mean (SD)	966	36.9 (0.6)	36.9 (0.6)	36.8 (0.6)	<0.001
mbp, mean (SD)	0	75.6 (10.2)	76.3 (10.7)	72.4 (7.2)	<0.001
resp_rate, mean (SD)	9	19.3 (4.3)	19.6 (4.4)	18.0 (3.8)	<0.001
heart_rate, mean (SD)	0	86.2 (16.3)	86.2 (16.8)	86.5 (14.3)	0.197
spo2, mean (SD)	4	97.4 (2.2)	97.3 (2.3)	98.0 (2.1)	<0.001
lactate, mean (SD)	4616	3.0 (2.5)	2.8 (2.4)	3.7 (2.6)	<0.001
urineoutput, mean (SD)	301	24.0 (52.7)	24.7 (58.2)	21.1 (16.6)	<0.001
admission_age, mean (SD)	0	66.3 (16.2)	66.1 (16.8)	67.3 (13.1)	<0.001
delta mortality to inclusion, mean (SD)	11121	316.9 (640.2)	309.6 (628.8)	365.0 (708.9)	0.022
delta intervention to inclusion, mean (SD)	14862	0.3 (0.2)	nan (nan)	0.3 (0.2)	nan
delta inclusion to intime, mean (SD)	0	0.1 (0.2)	0.1 (0.2)	0.1 (0.1)	0.041
delta ICU intime to hospital admission, mean (SD)	0	1.1 (3.7)	1.0 (3.7)	1.6 (3.4)	<0.001
los_hospital, mean (SD)	0	12.6 (12.5)	12.6 (12.5)	12.9 (12.4)	0.189
los_icu, mean (SD)	0	5.5 (6.7)	5.5 (6.5)	5.5 (7.2)	0.605

Table D.5. Characteristics of the trial population measured on the first 24 hours of ICU stay. Risk scores (AKI, SOFA, SAPSII) and lactates have been summarized as the maximum value during the 24 hour period for each stay. Total cumulative urine output has been computed. Other variables have been aggregated by taking mean during the 24 hour period.

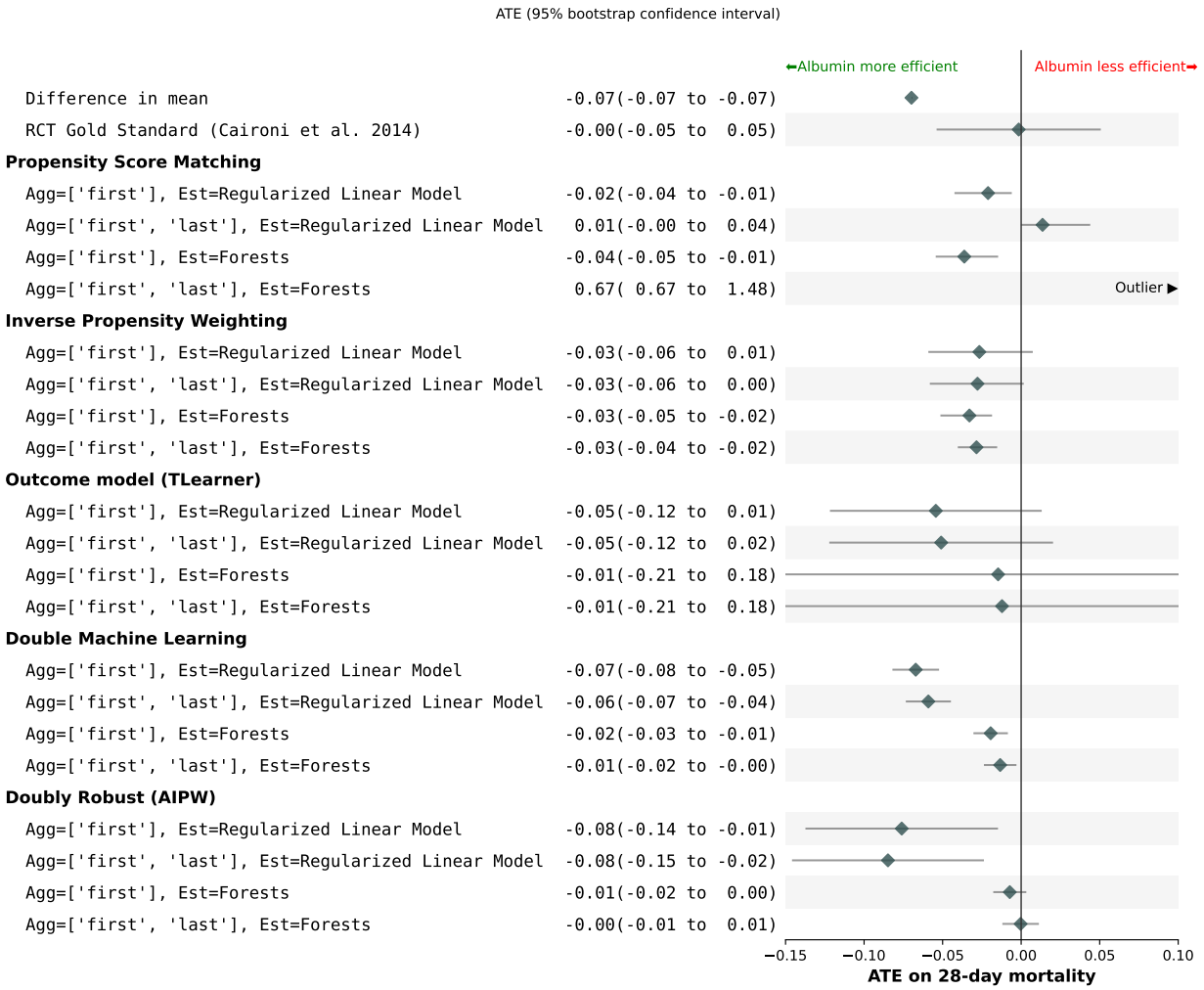


Fig. D.6. Full sensitivity analysis: The estimators with forest nuisances point to no effect for almost every causal estimator consistently with the RCT gold standard. Only matching with forest yields an unconvincingly high estimate. Linear nuisance used with doubly robust methods suggest a reduced mortality risk for albumin. The choices of aggregation only marginally modify the results expect for propensity score matching. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 50 bootstrap repetitions.

D.11 Details on treatment heterogeneity analysis

D.11.1 Detailed estimation procedure

The estimation of heterogeneous effect based on Double Machine Learning adds another step after the computation, regressing the residuals of the outcome nuisance $\tilde{Y} - \mu(X)$ against the residuals of the treatment nuisance $\tilde{A} = A - e(X)$ with the heterogeneity features X_{CATE}. Noting the final CATE model θ, Double ML solves:

$$\underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_n\left[(\tilde{Y} - \theta(X_{CATE}) \cdot \tilde{A})^2\right]$$

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

To avoid the over-fitting of this last regression model, we split the dataset of the main analysis into a train set (size=0.8) where the causal estimator and the final model are learned, and a test set (size=0.2) on which we report the predicted Conditional Average Treatment

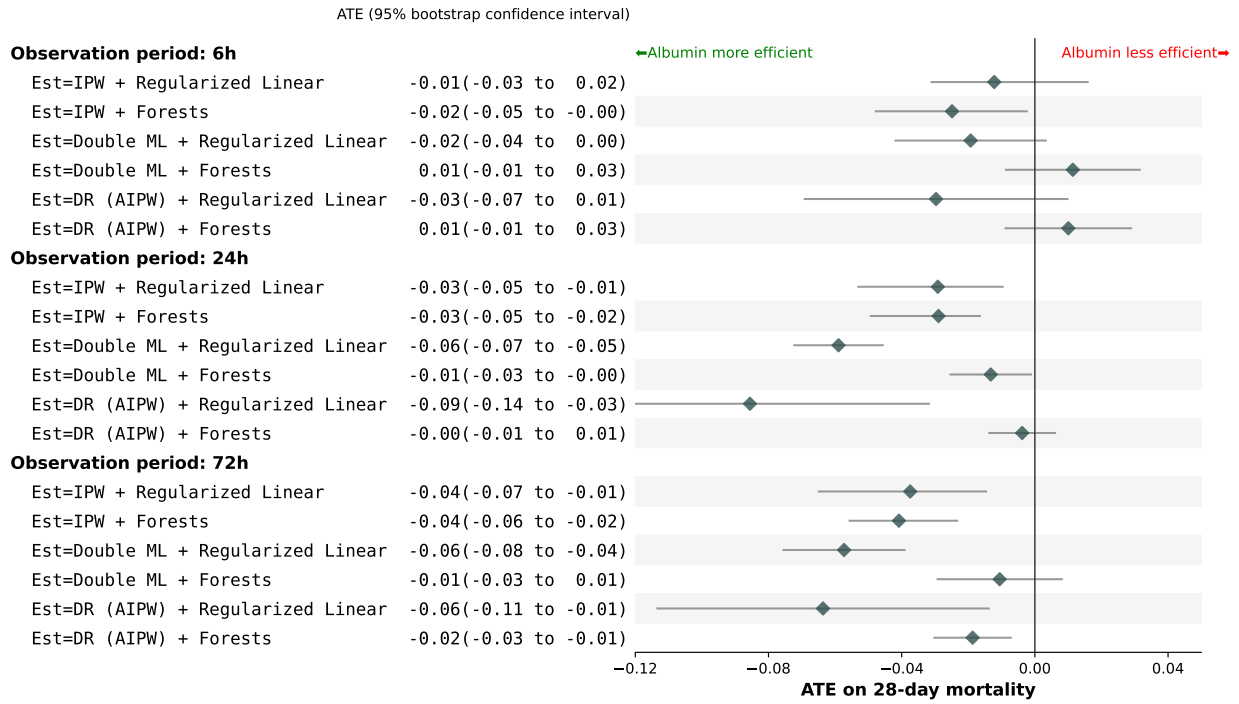


Fig. D.7. Sensitivity analysis for immortal time bias: Every choice of estimates show an improvement of the albumin treatment when increasing the observation period, thus increasing the blank period between inclusion and administration of albumin. Aggregation was concatenation of first and last features. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 50 bootstrap repetitions.

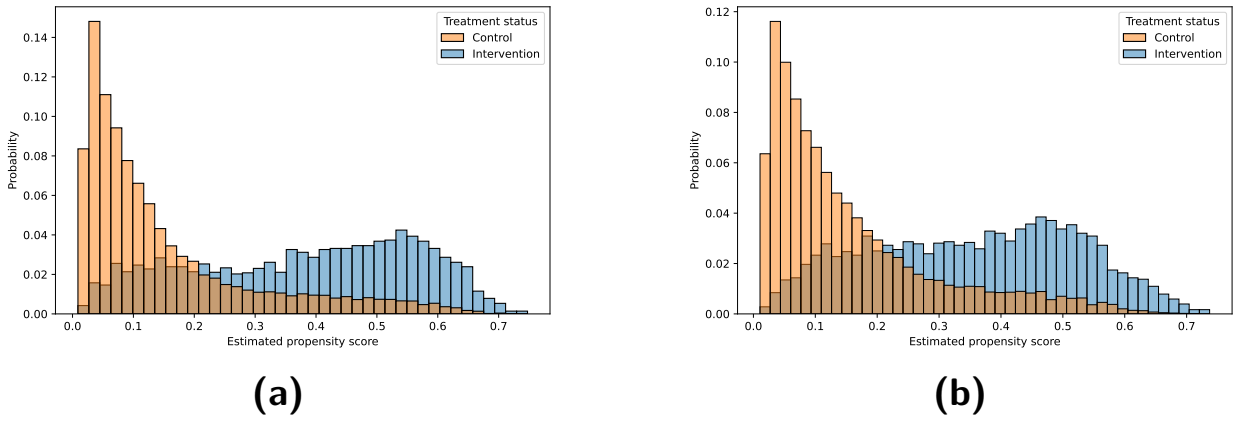


Fig. D.8. Different choices of aggregation yield qualitatively close distributions of the propensity score: Figure D.8a)a) shows a concatenation of first, last and median measures whereas Figure D.8b)b) shows an aggregation by taking the first measure only. The underlying treatment effect estimator is a random forest.

Effects.

D.11.2 Known heterogeneity of treatment for the emulated trial

Caironi et al., 2014 observed statistical differences in the post-hoc subgroup analysis between patient with and without septic shock at inclusion. They found increasing treatment effect measured as 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).

Investigators, 2007 conducted a post-hoc subgroup analysis of patients with or without brain injury –defined as Glasgow Coma Scale between 3 to 8–. The initial population was patients with traumatic brain injury (defined as history or evidence on A CT scan of head trauma, and a GCS score <= 13). They found higher mortality rate at 24 months in the albumin group for patients with severe head injuries.

Zhou et al., 2021b conducted a subgroup analysis on age (<60 vs >60), septic shock and sex. They conclude for increasing treatment effect measured as Restricted Mean Survival Time for Sepsis vs septic shock (3.47 vs. 2.58), for age >=60 (3.75 vs 2.44), for Male (3.4 vs 2.69). None of these differences were statistically significant.

D.11.3 Vibration analysis

The choice of the final model for the CATE estimation should also be informed by statistical and clinical rationals. Figure D.10 shows the distribution of the individual effects of a final random forest estimator, yielding CATE estimates that are not consistent with the main ATE analysis. Figure D.11 shows that the choice of this final model imposes a inductive bias on the form of the heterogeneity and different sources of noise depending of the nature of the model. A random forest is more noisy than a linear model. Figure D.11 shows the difference of modelization on the subpopulation of non white male patients without septic shock. One see that the downside linear trend is reflected by the forest only for patients aged between 55 and 80.

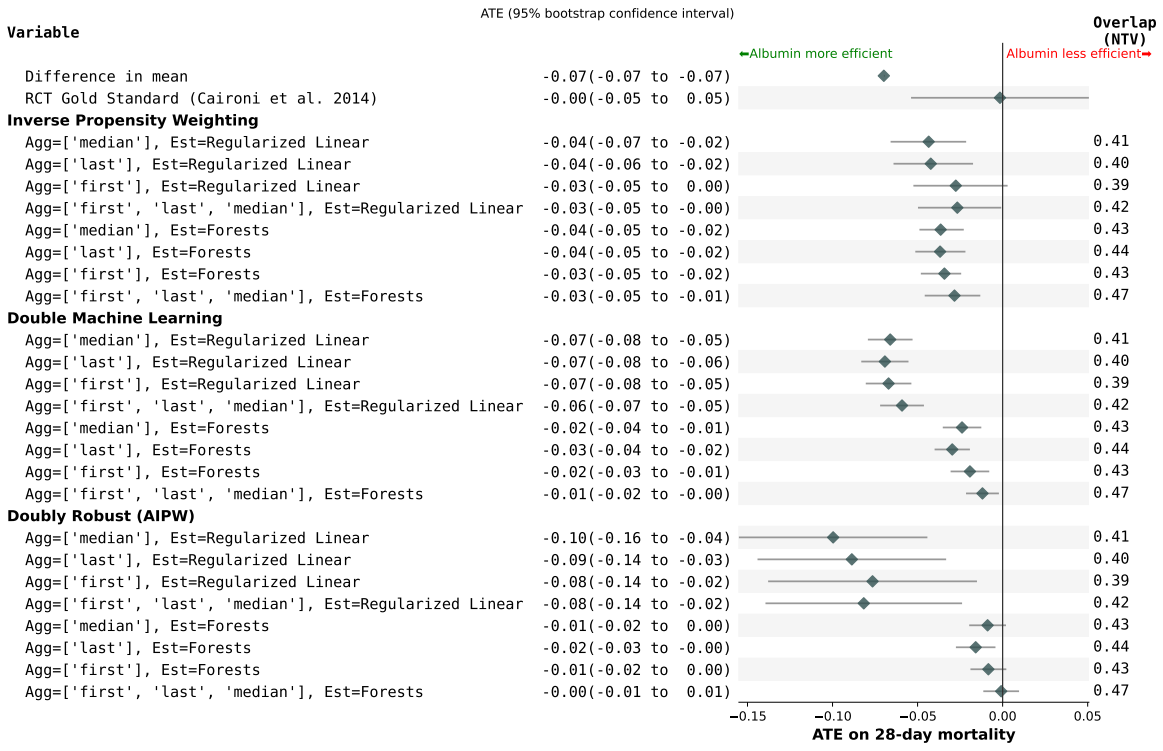


Fig. D.9. Vibration analysis dedicated to the aggregation choices. The choices of aggregation only marginally modify the results. When assessed with Normalized Total Variation, the overlap assumption is respected for all our choices of aggregation. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 50 bootstrap repetitions.

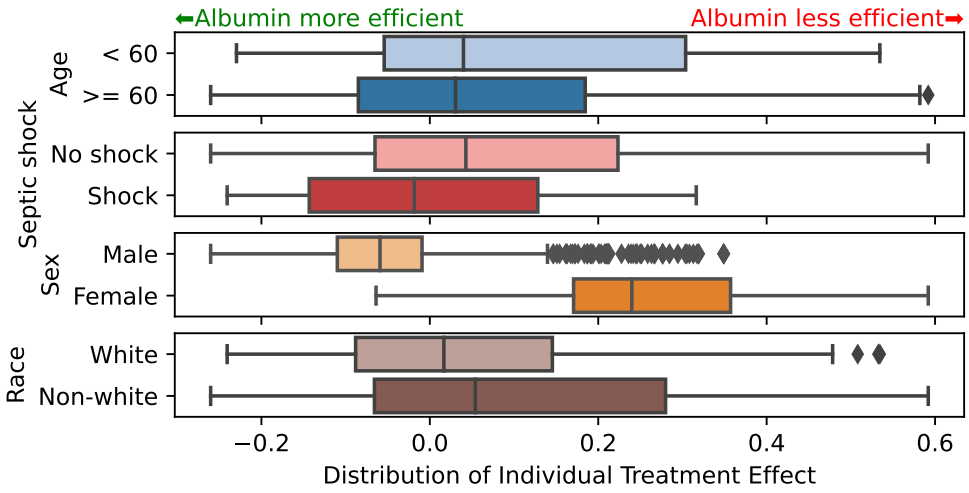


Fig. D.10. Distribution of Conditional Average Treatment effects on sex, age, race and pre-treatment septic shock estimated with a final forest estimator. The CATE are positive for each subgroups, which is not consistent with the null treatment effect obtained in the main analysis. The boxes contain between the 25th and 75th percentiles of the CATE distributions with the median indicated by a vertical line. The whiskers extends to 1.5 the inter-quartile range of the distribution.

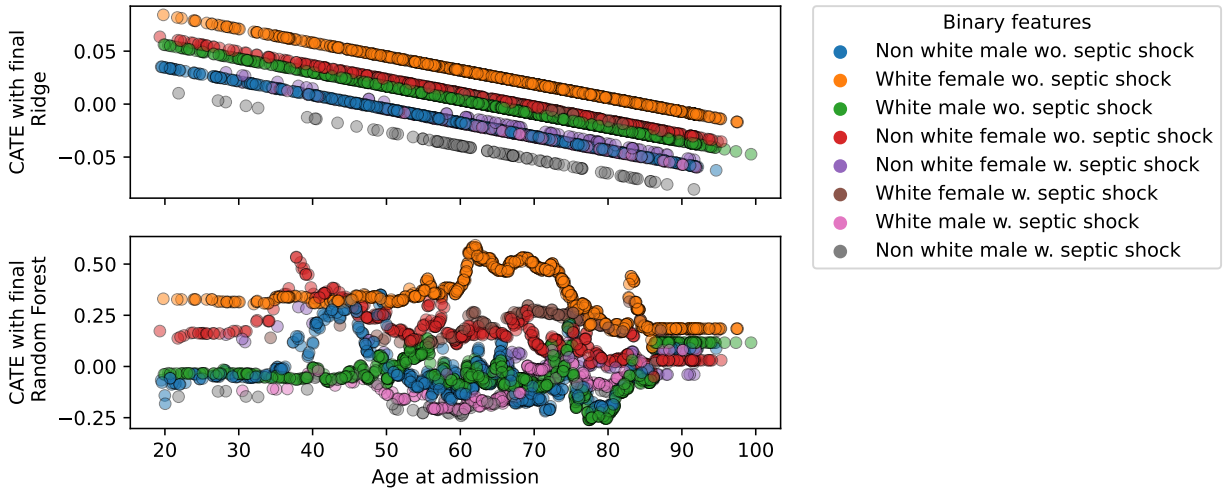


Fig. D.11. Distribution of Conditional Average Treatment effects on sex, age, race and pre-treatment septic shock plotted for different ages. On the top the final estimator is a linear model; on the bottom, it is a random forest. The forest-based CATE displays more noisy trends than the linear-based CATE. This suggest that the flexibility of the random forest might be underfitting the data.

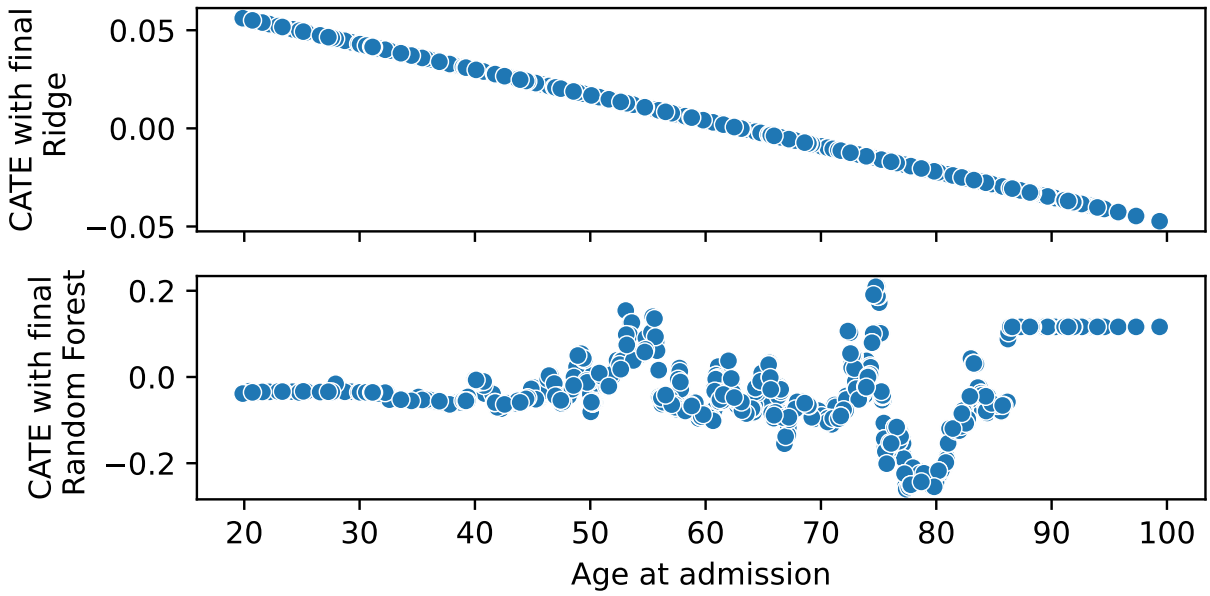


Fig. D.12. Figure D.11 on the subpopulation of white male patients without septic shock. Contrary to the ridge regression (on top) inducing a nicely interpretable trend, using random forests as the final estimator failed to recover CATE on ages: the predicted estimates do not exhibit any trend and display inconsistently large effect sizes, suggesting data underfitting.

Appendix E

Chapter 5

E.1 Variability of ATE estimation on ACIC 2016

Figure 5.1 shows ATE estimations for six different models used in g-computation estimators on the 76 configurations of the ACIC 2016 dataset. Outcome models are fitted on half of the data and inference is done on the other half –ie. train/test with a split ratio of 0.5. For each configuration, and each model, this train test split was repeated ten times, yielding non parametric variance estimates (Bouthillier et al., 2021b).

Outcome models are implemented with scikit-learn (Pedregosa et al., 2011) and the following hyper-parameters:

Outcome Model	Hyper-parameters grid
Random Forests	Max depth: [2, 10]
Ridge regression without treatment interaction	Ridge regularization: [0.1]
Ridge regression with treatment interaction	Ridge regularization: [0.1]

Table E.1. Hyper-parameters grid used for ACIC 2016 ATE variability

E.2 Proofs: Links between feasible and oracle risks

E.2.1 Upper bound of τ-risk with μ-risk$_{IPW}$

For the bound with the μ-risk$_{IPW}$, we will decompose the CATE risk on each factual population risks:

Definition 7 (Population Factual μ-risk) *(Shalit et al., 2017)*

$$\mu\text{-}risk_a(f) = \int_{\mathcal{Y}\times\mathcal{X}} (y - f(x; A = a))^2 p(y; x = x \mid A = a)\, dydx$$

Applying Bayes rule, we can decompose the μ-risk on each intervention:

$$\mu\text{-risk}(f) = p_A\, \mu\text{-risk}_1(f) + (1 - p_A)\ \mu\text{-risk}_0(f) \text{with } p_A = \mathbb{P}(A = 1)$$

These definitions allows to state a intermediary result on each population:

Lemma 1 (Mean-variance decomposition) *We need a reweighted version of the classical mean-variance decomposition.*

For an outcome model $f : x \times A \to \mathcal{X}$. Let the inverse propensity weighting function $w(a; x) = ae(x)^{-1} + (1 - a)(1 - e(x))^{-1}$.

$$\int_{\mathcal{X}} (\mu_1(x) - f(x; 1))^2 p(x)dx = p_A \mu\text{-}risk_{IPW,1}(w, f) - \sigma^2_{Bayes}(1)$$

And

$$\int_{\mathcal{X}} (\mu_0(x) - f(x;0))^2 p(x) dx = (1 - p_A) \mu\text{-}risk_{IPW,0}(w, f) - \sigma^2_{Bayes}(0)$$

Proof 3

$$p_A \mu\text{-}risk_{IPW,1}(w, f) = \int_{\mathcal{X}\times\mathcal{Y}} \frac{1}{e(x)} (y - f(x;1))^2 p(y \mid x; A = 1) p(x; A = 1) dy dx$$
$$= \int_{\mathcal{X}\times\mathcal{Y}} (y - f(x;1))^2 p(y \mid x; A = 1) \frac{p(x; A = 1)}{p(x; A = 1)} p(x) dy dx$$
$$= \int_{\mathcal{X}\times\mathcal{Y}} \Big[(y - \mu_1(x))^2 + \big(\mu_1(x) - f(x;1)\big)^2 + 2\,(y - \mu_1(x))\,(\mu_1(x) - f(x,1)) \Big] p(y \mid x; A = 1) p(x) dy dx$$
$$= \int_{\mathcal{X}} \Big[\int_{\mathcal{Y}} (y - \mu_1(x))^2 p(y \mid x; A = 1) dy \Big] p(x) dx + \int_{\mathcal{X}\times\mathcal{Y}} \big(\mu_1(x) - f(x;1)\big)^2 p(x) p(y \mid x; A = 1) dx dy$$
$$+ \qquad 2 \int_{\mathcal{X}} \Big[\int_{\mathcal{Y}} (y - \mu_1(x))\, p(y \mid x; A = 1) dy \Big] (\mu_1(x) - f(x,1))\, p(x) dx$$
$$= \int_{\mathcal{X}} \sigma_y^2(x, 1) p(x) dx + \int_{\mathcal{X}} \big(\mu_1(x) - f(x;1)\big)^2 p(x) dx + 0$$

Proposition 1 (Upper bound with mu-IPW) *Let f be a given outcome model, let the weighting function w be the Inverse Propensity Weight $w(x;a) = \frac{a}{e(x)} + \frac{1-a}{1-e(x)}$. Then, under overlap (assumption 2),*

$$\tau\text{-}risk(f) \leq 2\, \mu\text{-}risk_{IPW}(w, f) \; - 2\, (\sigma^2_{Bayes}(1) + \sigma^2_{Bayes}(0))$$

Proof 4

$$\tau\text{-}risk(f) = \int_{\mathcal{X}} (\mu_1(x) - \mu_0(x) - (f(x;1) - f(x;0))^2 p(x) dx$$

By the triangle inequality $(u + v)^2 \leq 2(u^2 + v^2)$:

$$\tau\text{-}risk(f) \leq 2 \int_{\mathcal{X}} \Big[\big(\mu_1(x) - f(x;1)\big)^2 + \big(\mu_0(x) - f(x;0)\big)^2 \Big] p(x) dx$$

Applying Lemma 1,

$$\tau\text{-}risk(f) \leq 2 \Big[p_A \mu\text{-}risk_{IPW,1}(w, f) + (1 - p_A) \mu\text{-}risk_{IPW,0}(w, f)(w, f) \Big] - 2(\sigma^2_{Bayes}(0) + \sigma^2_{Bayes}(1))$$
$$= 2\mu\text{-}risk_{IPW}(w, f) - 2(\sigma^2_{Bayes}(0) + \sigma^2_{Bayes}(1)$$

E.2.2 Reformulation of the R-risk as reweighted τ-risk

Proposition 2 (R-risk as reweighted τ-risk) ***Proof 5*** *We consider the R-decomposition: (Robinson, 1988),*

$$y(a) = m(x) + \big(a - e(x)\big) \tau(x) + \varepsilon(x; a) \tag{E.1}$$

Where $\mathbb{E}[\varepsilon(X; A)|X, A] = 0$ We can use it as plug in the R-risk formula:

$$
\begin{aligned}
&R\text{-}risk(f) = \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{A}} [(y - m(x)) - \Big(a - e(x)\Big)\tau_f(x)]^2 p(y;x;a)dydxda \\
&= \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{A}} \Big[\Big(a - e(x)\Big)\tau(x) + \varepsilon(x;a) - \Big(a - e(x)\Big)\tau_f(x)\Big]^2 p(y;x;a)dydxda \\
&= \int_{\mathcal{X}\times\mathcal{A}} \Big(a - e(x)\Big)^2 \Big(\tau(x) - \tau_f(x)\Big)^2 p(x;a)dxda \\
&+ 2\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{A}} \Big(a - e(x)\Big)\Big(\tau(x) - \tau_f(x)\Big) \int_{\mathcal{Y}} \varepsilon(x;a)p(y \mid x;a)dyp(x;a)dxda \\
&+ \int_{\mathcal{X}\times\mathcal{A}} \int_{\mathcal{Y}} \varepsilon^2(x;a)p(y \mid x;a)dyp(x;a)dxda
\end{aligned}
$$

The first term can be decomposed on control and treated populations to force $e(x)$ to appear:

$$
\begin{aligned}
&\int_{\mathcal{X}} \Big(\tau(x) - \tau_f(x)\Big)^2 \Big[e(x)^2 p(x;0) + \Big(1 - e(x)\Big)^2 p(x;1)\Big] dx \\
&= \int_{\mathcal{X}} \Big(\tau(x) - \tau_f(x)\Big)^2 \Big[e(x)^2\Big(1 - e(x)\Big)p(x) + \Big(1 - e(x)\Big)^2 e(x)p(x)\Big] dx \\
&= \int_{\mathcal{X}} (\tau(x) - \tau_f(x))^2 (1 - e(x))e(x)[1 - e(x) + e(x)]p(x)dx \\
&= \int_{\mathcal{X}} (\tau(x) - \tau_f(x))^2 (1 - e(x))e(x)p(x)dx.
\end{aligned}
$$

The second term is null since, $\mathbb{E}[\varepsilon(x,a)|X,A] = 0$.
The third term corresponds to the modulated residuals 5.3 : $\tilde{\sigma}_B^2(0) + \tilde{\sigma}_B^2(1)$

E.3 Measuring overlap

Motivation of the Normalized Total Variation Computing overlap when working only on samples of the observed distribution, outside of simulation, requires a sophisticated estimator of discrepancy between distributions, as two data points never have the same exact set of features. Maximum Mean Discrepancy (Gretton et al., 2012) is typically used in the context of causal inference (Shalit et al., 2017; Johansson et al., 2022). However it needs a kernel, typically Gaussian, to extrapolate across neighboring observations. We prefer avoiding the need to specify such a kernel, as it must be adapted to the data which is tricky with categorical or non-Gaussian features, a common situation for medical data.

For simulated and some semi-simulated data, we have access to the probability of treatment for each data point, which sample both densities in the same data point. Thus, we can directly use distribution discrepancy measures and rely on the Normalized Total Variation (NTV) distance to measure the overlap between the treated and control propensities. This is the empirical measure of the total variation distance (Sriperumbudur et al., 2009) between the distributions, $TV(\mathbb{P}(X|A=1), \mathbb{P}(X|A=0))$. As we have both distribution sampled on the same points, we can rewrite it a sole function of the propensity score, a low dimensional score more tractable than the full distribution $\mathbb{P}(X|A)$:

$$
\widehat{NTV}(e, 1-e) = \frac{1}{2N} \sum_{i=1}^{N} \Big| \frac{e(x_i)}{p_A} - \frac{1 - e(x_i)}{1 - p_A} \Big| \tag{E.2}
$$

Formally, we can rewrite NTV as the Total Variation distance between the two population distributions. For a population $O = (Y(A), X, A) \sim \mathcal{D}$:

$$
\begin{aligned}
NTV(O) &= \frac{1}{2N}\sum_{i=1}^{N}\Big|\frac{e(x_i)}{p_A} - \frac{1-e(x_i)}{1-p_A}\Big| \\
&= \frac{1}{2N}\sum_{i=1}^{N}\Big|\frac{P(A=1|X=x_i)}{p_A} - \frac{P(A=0|X=x_i)}{1-p_A}\Big|
\end{aligned}
$$

Thus NTV approximates the following quantity in expectation over the data distribution $\mathcal{D}$:

$$
\begin{aligned}
NTV(\mathcal{D}) &= \int_{\mathcal{X}}\Big|\frac{p(A=1|X=x)}{p_A} - \frac{p(A=0|X=x)}{1-p_A}\Big|p(x)dx \\
&= \int_{\mathcal{X}}\Big|\frac{p(A=1,X=x)}{p_A} - \frac{p(A=0,X=x)}{1-p_A}\Big|dx \\
&= \int_{\mathcal{X}}\Big|p(X=x|A=1) - p(X=x|A=0)\Big|dx
\end{aligned}
$$

For countable sets, this expression corresponds to the Total Variation distance between treated and control populations covariate distributions : $TV(p_0(x), p_1(x))$.

Measuring overlap without the oracle propensity scores For ACIC 2018, or for non-simulated data, the true propensity scores are not known. To measure overlap, we rely on flexible estimations of the Normalized Total Variation, using gradient boosting trees to approximate the propensity score. Empirical arguments for this plug-in approach is given in Figure E.1.

Empirical arguments We show empirically that NTV is an appropriate measure of overlap by :

- Comparing the NTV distance with the MMD for Caussim which is gaussian distributed in Figure E.3,
- Verifying that setups with penalized overlap from ACIC 2016 have a higher total variation distance than unpenalized setups in Figure E.2.
- Verifying that the Inverse Propensity Weights extrema (the inverse of the ν overlap constant appearing in the overlap Assumption 2) positevely correlates with NTV for Caussim, ACIC 2016 and Twins in Figure E.4. Even if the same value of the maximum IPW could lead to different values of NTV, we expect both measures to be correlated : the higher the extrem propensity weights, the higher the NTV.

Estimating NTV in practice Finally, we verify that approximating the NTV distance with a learned plug-in estimates of e(x) is reasonnable. We used either a logistic regression or a gradient boosting classifier to learn the propensity models for the three datasets where we have access to the ground truth propensity scores: Caussim, Twins and ACIC 2016. We respectively sampled 1000, 1000 and 770 instances of these datasets with different seeds and overlap settings. We first run a hyperparameter search with cross-validation on the train set, then select the best estimator. We refit on the train set this estimator with or without calibration by cross validation and finally estimate the normalized TV with the obtained model. This training procedure reflects the one described in Algorithm 1 where nuisance models are fitted only on the train set.

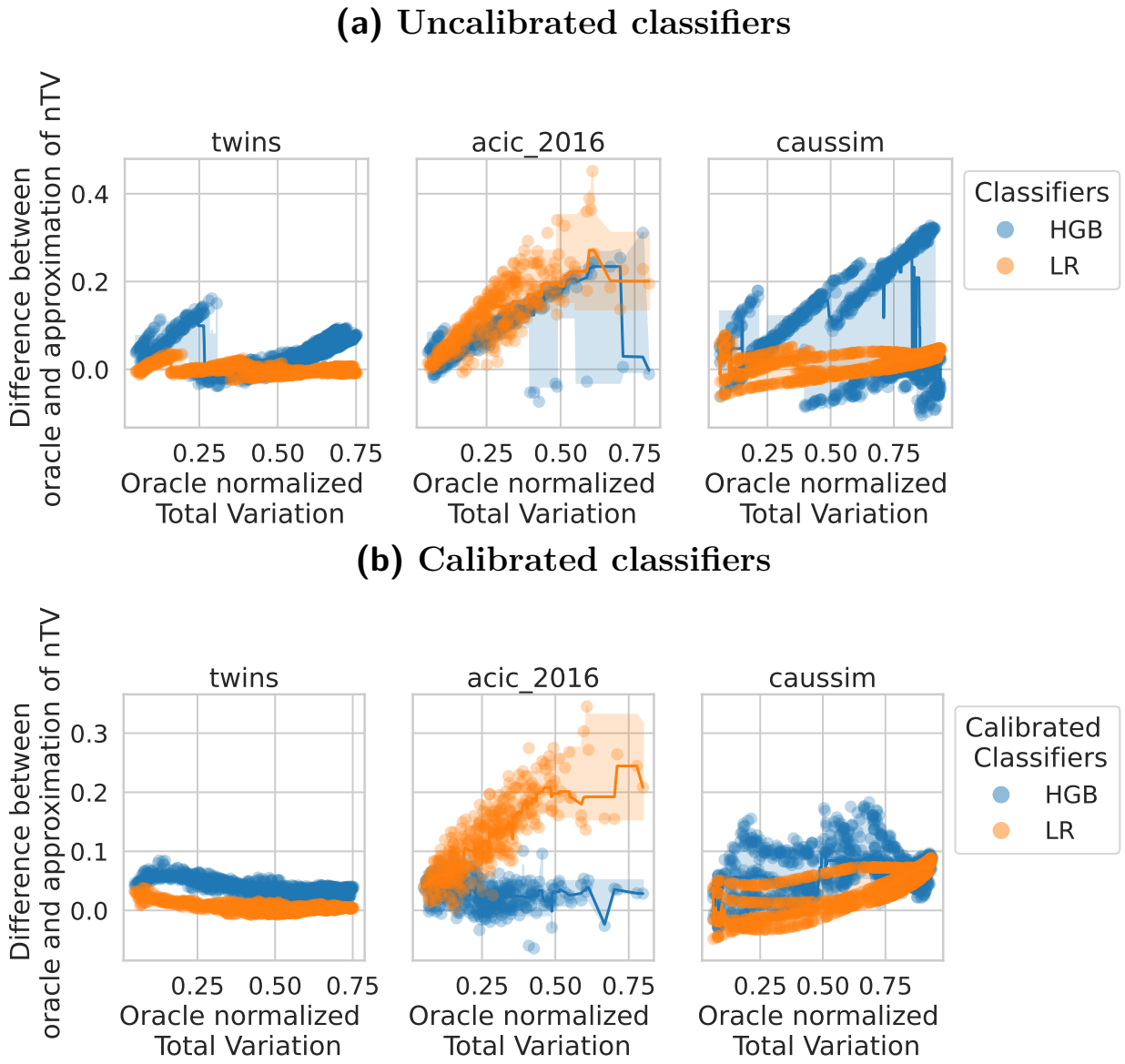


Fig. E.1. a) Without calibration, estimation of NTV is not trivial even for boosting models. b) Calibrated classifiers are able to recover the true Normalized Total Variation for all datasets where it is available.

The hyper parameters are : learning rate $\in [1e-3, 1e-2, 1e-1, 1]$, minimum samples leaf $\in [2, 10, 50, 100, 200]$ for boosting and L2 regularization $\in [1e-3, 1e-2, 1e-1, 1]$ for logistic regression.

Results in Figure E.1 comparing bias to the true normalized Total Variation of each dataset instances versus growing true NTV indicate that calibration of the propensity model is crucial to recover a good approximation of the NTV.

E.4 Experiments

E.4.1 Details on the data generation process

We use Gaussian-distributed covariates and random basis expansion based on Radial Basis Function kernels. A random basis of RBF kernel enables modeling non-linear and complex

Fig. E.2. NTV recovers well the overlap settings described in the ACIC paper (Dorie et al., 2019)

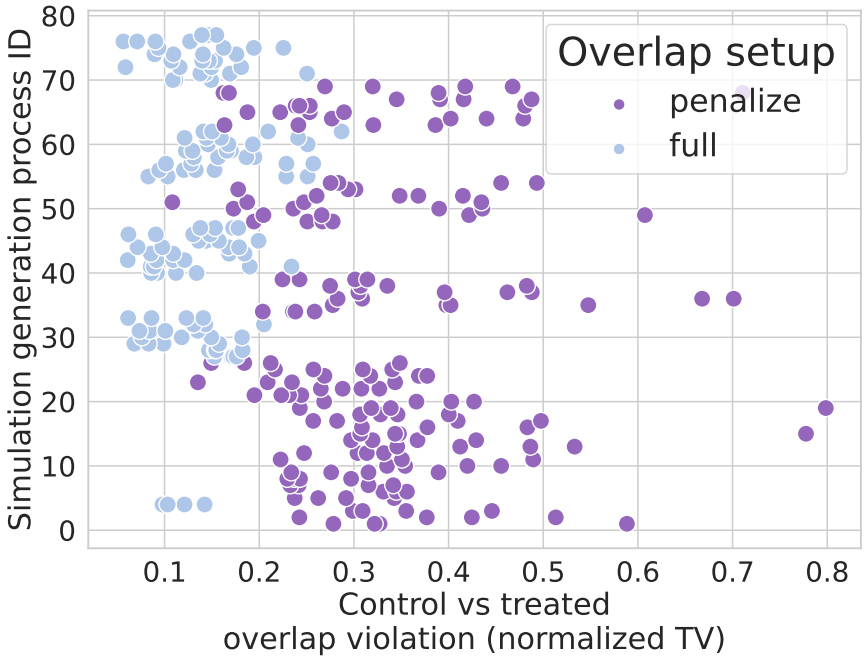
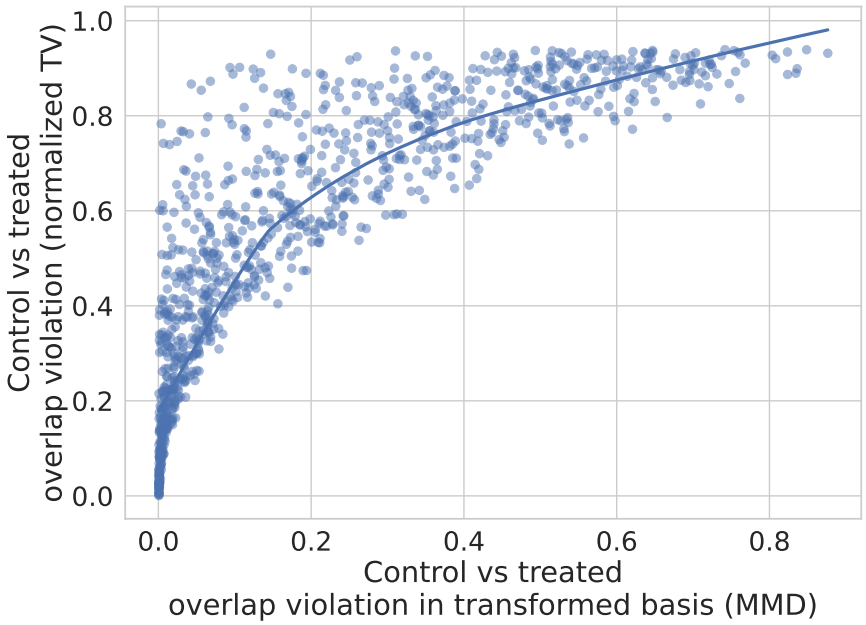


Fig. E.3. Good correlation between overlap measured as normalized Total Variation and Maximum Mean Discrepancy (200 sampled Caussim datasets)



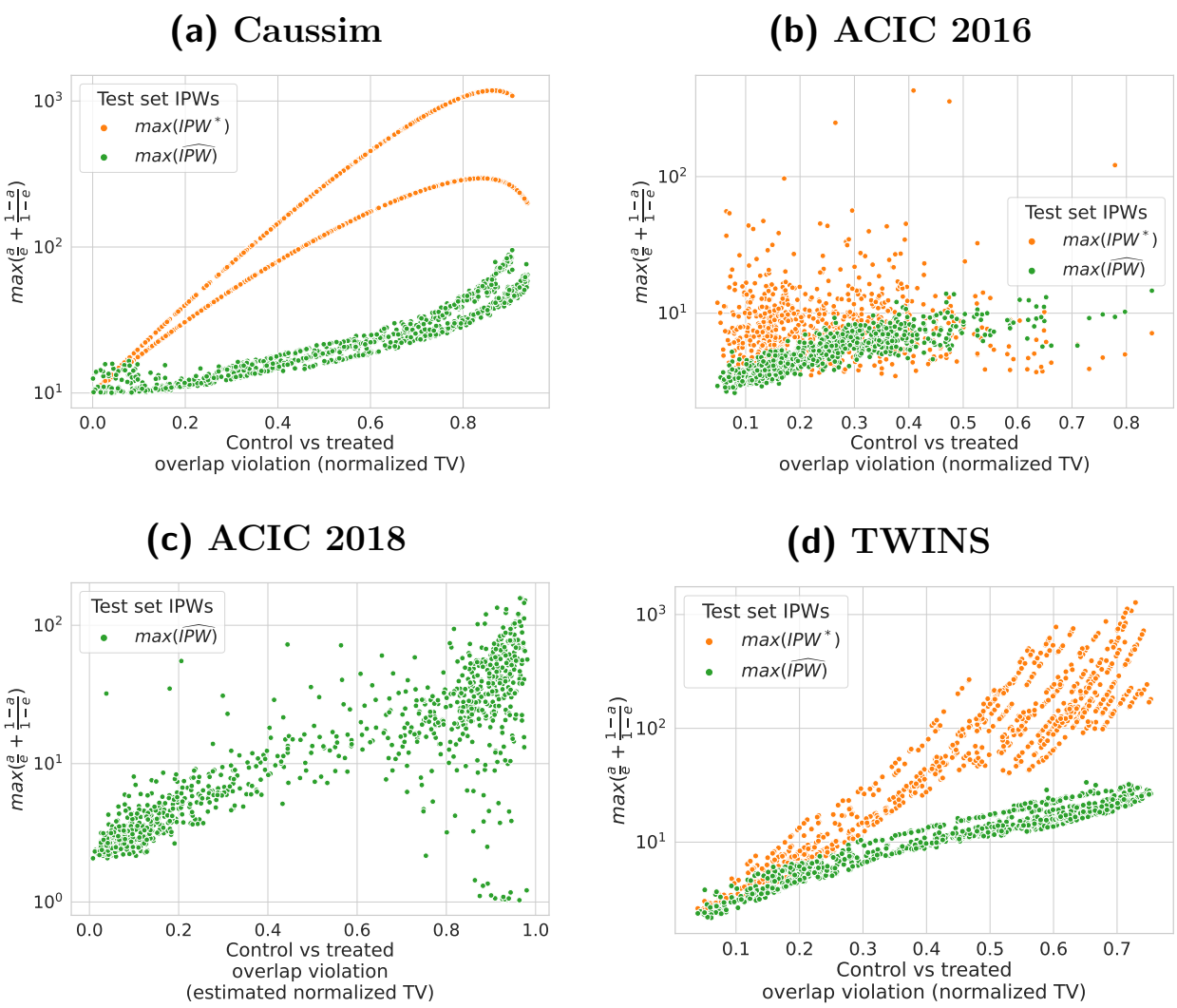


Fig. E.4. Maximal value of Inverse Propensity Weights increases exponentially with the overlap as measure by Normalized Total Variation.

relationships between covariates in a similar way to the well known spline expansion. The estimators of the response function are learned with a linear model on another random basis (which can be seen as a stochastic approximation of the full data kernel (Rahimi; Recht, 2008)). We carefully control the overlap between treated and control populations, a crucial assumption for causal inference.

- The raw features for both populations are drawn from a mixture of Gaussians: $\mathbb{P}(X) = p_A\mathbb{P}(X|A=1) + (1-p_A)\mathbb{P}(X|A=0)$ where $\mathbb{P}(x|A=a)$ is a rotated Gaussian:

$$\mathbb{P}(x|A=a) = W \cdot \mathcal{N}\left(\begin{bmatrix} (1-2a)\theta \\ 0 \end{bmatrix}; \begin{bmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{bmatrix} \right) \tag{E.3}$$

  with θ a parameter controlling overlap (bigger yields poorer overlap), W a random rotation matrix and $\sigma_0^2 = 2; \sigma_1^2 = 5$.

  This generation process allows to analytically compute the oracle propensity scores $e(x)$, to simply control for overlap with the parameter θ, the distance between the two Gaussian main axes and to visualize response surfaces.

- A basis expansion of the raw features increases the problem dimension. Using Radial Basis Function (RBF) Nystroem transformation [1], we expand the raw features into a transformed space. The basis expansion samples randomly a small number of representers in the raw data. Then, it computes an approximation of the full N-dimensional kernel with these basis components, yielding the transformed features $z(x)$.

  We generate the basis following the original data distribution, $[b_1..b_D] \sim \mathbb{P}(x)$, with D=2 in our simulations. Then, we compute an approximation of the full kernel of the data generation process $RBF(x, \cdot)$ *with* $x \sim \mathbb{P}(x)$ with these representers: $z(x) = [RBF_\gamma(x, b_d)]_{d=1..D} \cdot Z^T \in \mathbb{R}^D$ with RBF_γ being the Gaussian kernel $K(x, y) = exp(-\gamma||x - y||^2)$ and Z the normalization constant of the kernel basis, computed as the root inverse of the basis kernel $Z = [K(b_i, b_j)]^{-1/2}_{i,j \in 1..D}$
- Functions μ_0, τ are distinct linear functions of the transformed features:

$$\mu_0(x) = \Big[z(x); 1\Big] \cdot \beta_\mu^T$$

$$\tau(x) = \Big[z(x); 1\Big] \cdot \beta_\tau^T$$

- Adding a Gaussian noise, $\varepsilon \sim \mathcal{N}(0, \sigma(x; a))$, we construct the potential outcomes: $y(a) = \mu_0(x) + a\,\tau(x) + \varepsilon(x, a)$

We generated 1000 instances of this dataset with uniformly random overlap parameters $\theta \in [0, 2.5]$.

E.4.2 Model selection procedures

Nuisances estimation The nuisances are estimated with a stacked regressor inspired by the Super Learner framework, (Laan et al., 2007). The hyper-parameters are optimized with a random search with following search grid detailed in Table E.2. All implementations come from scikit-learn (Pedregosa et al., 2011).

Model	Estimator	Hyper-parameters grid
Outcome, m	StackedRegressor (HistGradientBoostingRegressor, ridge)	ridge regularization: [0.0001, 0.00 HistGradientBoostingRegressor le HistGradientBoostingRegressor m
Treatment, e	StackedClassifier (HistgradientBoostingClassifier, LogisticRegression)	LogisticRegression C: [0.0001, 0.0 HistGradientBoostingClassifier le HistGradientBoostingClassifier m

Table E.2. Hyper-parameters grid used for nuisance models

E.4.3 Additional Results

Definition of the Kendall's tau, κ The Kendall's tau is a widely used statistics to measure the rank correlation between two set of observations. It measures the number of

[1] We use the Sklearn implementation, (Pedregosa et al., 2011)

Metric	Dataset	Strong Overlap Median	Strong Overlap IQR	Weak Overlap Median	Weak Overlap IQR
$\widehat{\mu-\text{risk}}$	Twins (N= 11 984)	-0.32	0.12	-0.19	0.12
	ACIC 2016 (N=4 802)	-0.03	0.13	0.11	0.19
	Caussim (N=5 000)	-0.40	0.55	-0.16	0.31
	ACIC 2018 (N=5 000)	0.00	0.30	0.01	0.40
$\widehat{\mu-\text{risk}}_{IPW}$	Twins (N= 11 984)	-0.31	0.13	-0.17	0.12
	ACIC 2016 (N=4 802)	-0.02	0.13	0.11	0.19
	Caussim (N=5 000)	-0.34	0.50	0.09	0.31
	ACIC 2018 (N=5 000)	0.00	0.30	-0.01	0.43
$\widehat{\mu-\text{risk}}^{*}_{IPW}$	Twins (N= 11 984)	-0.32	0.13	-0.17	0.13
	ACIC 2016 (N=4 802)	-0.02	0.13	0.11	0.21
	Caussim (N=5 000)	-0.33	0.54	0.26	0.27
$\widehat{\tau-\text{risk}}_{IPW}$	Twins (N= 11 984)	0.13	0.12	0.27	0.12
	ACIC 2016 (N=4 802)	-0.07	0.18	0.05	0.31
	Caussim (N=5 000)	-0.19	0.43	-0.14	0.18
	ACIC 2018 (N=5 000)	-0.16	0.40	-0.11	0.66
$\widehat{\tau-\text{risk}}^{*}_{IPW}$	Twins (N= 11 984)	0.12	0.14	0.20	0.16
	ACIC 2016 (N=4 802)	-0.03	0.16	-0.09	0.43
	Caussim (N=5 000)	-0.15	0.46	-0.17	0.19
$\widehat{\text{U}-\text{risk}}$	Twins (N= 11 984)	0.13	0.12	0.02	0.25
	ACIC 2016 (N=4 802)	0.04	0.11	0.11	0.26
	Caussim (N=5 000)	0.04	0.43	-0.04	0.17
	ACIC 2018 (N=5 000)	0.12	0.26	-0.02	0.50
$\widehat{\text{U}-\text{risk}}^{*}$	Twins (N= 11 984)	0.25	0.08	-0.41	0.45
	ACIC 2016 (N=4 802)	0.08	0.13	-0.59	0.57
	Caussim (N=5 000)	0.46	0.12	0.02	0.44
$\widehat{\text{R}-\text{risk}}$	Twins (N= 11 984)	0.15	0.10	0.25	0.18
	ACIC 2016 (N=4 802)	0.07	0.12	0.22	0.15
	Caussim (N=5 000)	0.34	0.26	0.13	0.21
	ACIC 2018 (N=5 000)	0.13	0.27	0.21	0.47
$\widehat{\text{R}-\text{risk}}^{*}$	Twins (N= 11 984)	0.25	0.10	0.32	0.15
	ACIC 2016 (N=4 802)	0.12	0.12	0.25	0.15
	Caussim (N=5 000)	0.47	0.11	0.16	0.14

Table E.3. Values of relative $\kappa(\ell, \tau-\text{risk})$ compared to the mean over all metrics Kendall's as shown in the boxplots of Figure 5.5

concordant pairs minus the discordant pairs normalized by the total number of pairs. It takes values in the $[-1, 1]$ range.

$$\kappa = \frac{(\text{number of concordant pairs }) - (\text{number of discordant pairs})}{(\text{number of pairs})} \tag{E.4}$$

Values of relative $\kappa(\ell, \tau-\text{risk})$ Values of relative $\kappa(\ell, \tau-\text{risk})$ compared to the mean over all metrics Kendall's as shown in the boxplots of Figure 5.5.

Figure E.5 – Results measured in relative Kendall's for feasible and semi-oracle risks Because of extreme propensity scores in the denominator and bayes error residuals in the numerator, the semi-oracle U-risk has poor performance at bad overlap. Estimating these propensity scores in the is feasible U-risk reduces the variance since clipping is performed.

Figure E.6 – Results measured in absolute Kendall's

Figure E.7 – Results measured as distance to the oracle tau-risk To see practical gain in term of τ-risk, we plot the results as the normalized distance between the estimator selected by the oracle τ-risk and the estimator selected by each causal metric.

Then, $\widehat{R\text{-risk}}^*$ is more efficient than all other metrics. The gain are substantial for every datasets.

Figure E.8 – Stacked models for the nuisances is more efficient For each metrics the benefit of using a stacked model of linear and boosting estimators for nuisances compared to a linear model. The evaluation measure is Kendall's tau relative to the oracle R-risk* to have a stable reference between exepriments. Thus, we do not include in this analysis the ACIC 2018 dataset since R-risk* is not available due to the lack of the true propensity score.

Figure E.9 – Low population overlap hinders model selection for all metrics

Figure E.10 – Stacked models for the nuisances is more efficient For each metrics the benefit of using a stacked model of linear and boosting estimators for nuisances compared to a linear model. The evaluation measure is Kendall's tau relative to the oracle R-risk* to have a stable reference between exepriments. Thus, we do not include in this analysis the ACIC 2018 dataset since R-risk* is not available due to the lack of the true propensity score.

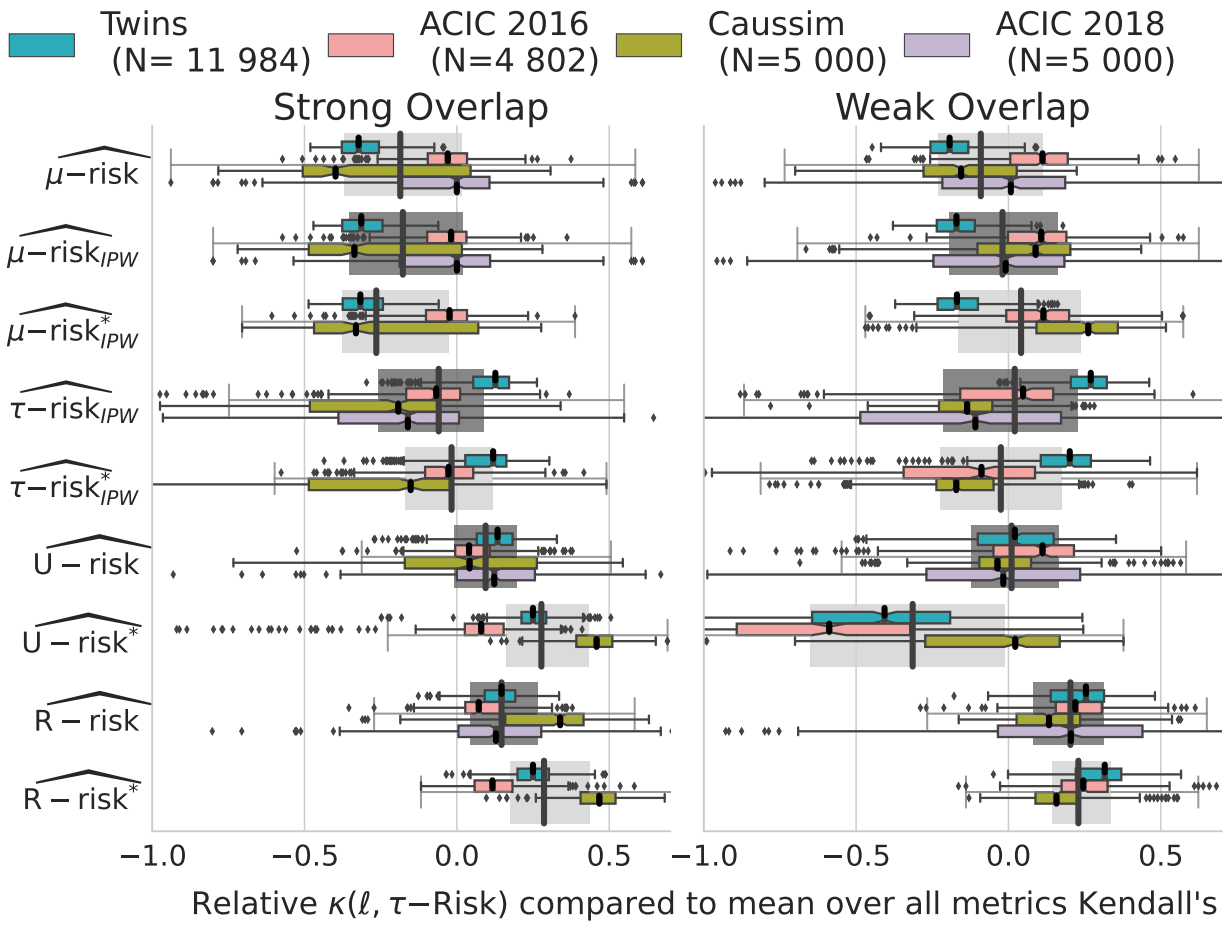


Fig. E.5. The R-risk is the best metric: Relative Kendall's τ agreement with τ-risk. Strong and weak overlap correspond to the first and last tertiles of the overlap distribution measured with Normalized Total Variation eq. E.2.

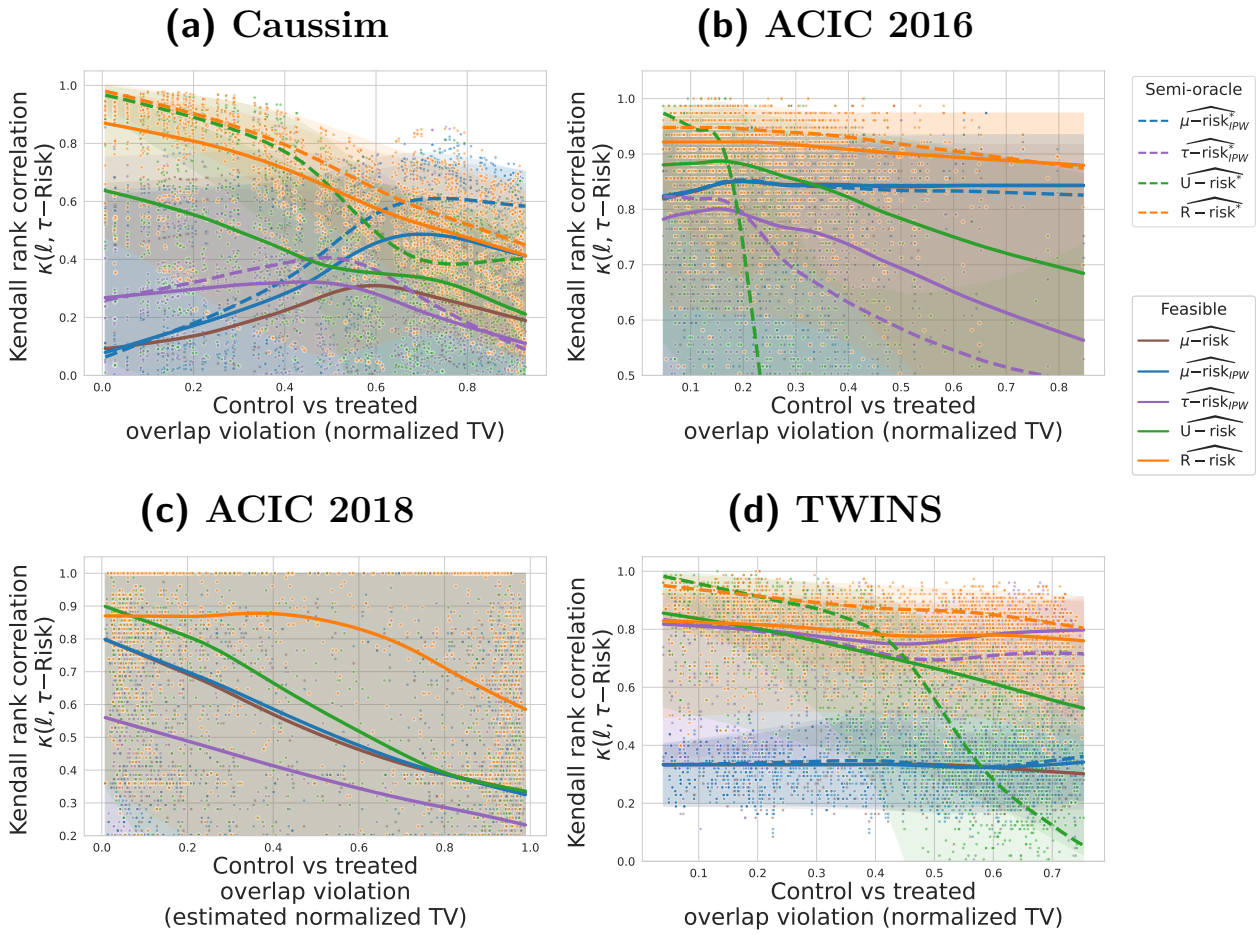


Fig. E.6. Agreement with τ-risk ranking of methods function of overlap violation. The lines represent medians, estimated with a lowess. The transparent bands denote the 5% and 95% confidence intervals.

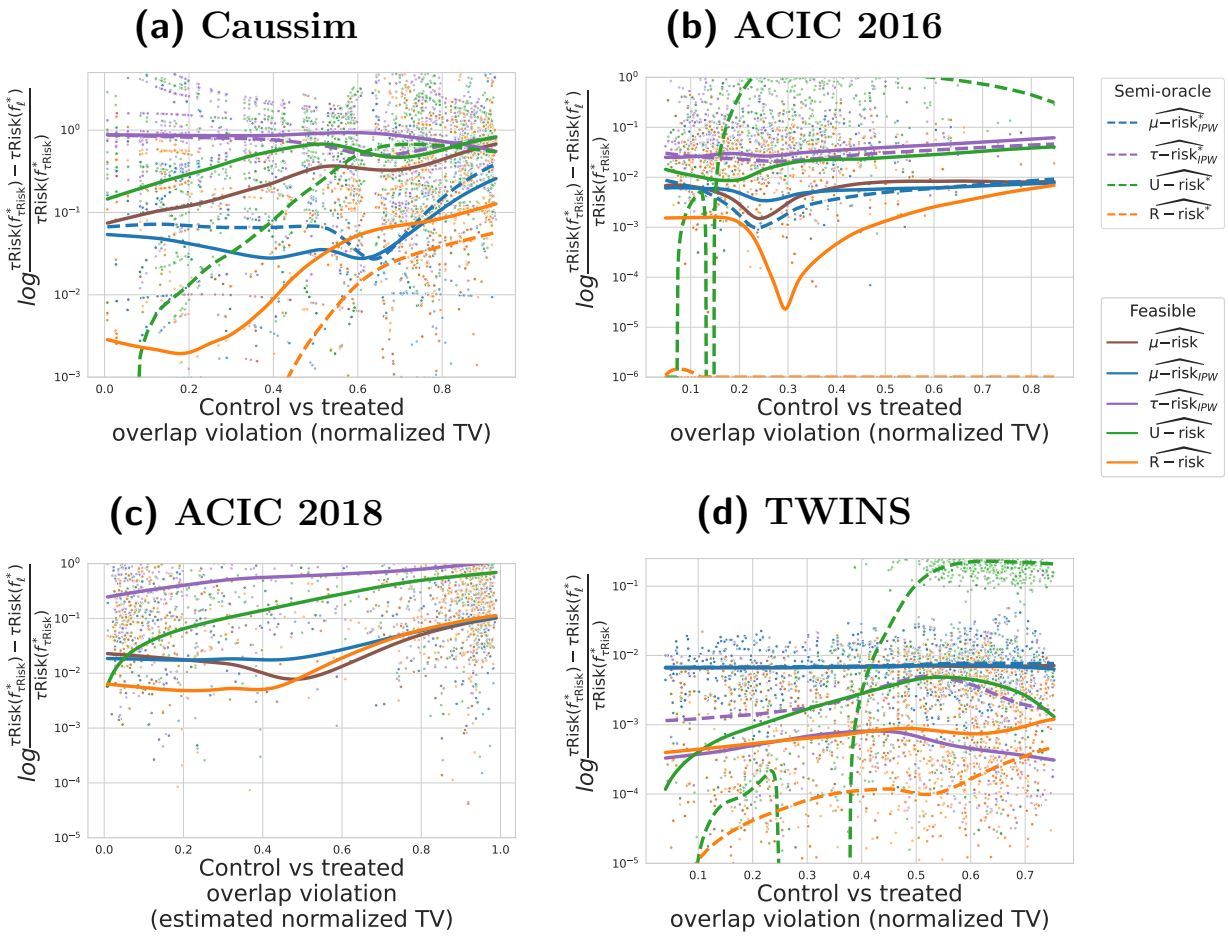


Fig. E.7. Metric performance by normalized tau-risk distance to the best method selected with τ-risk. All nuisances are learned with the same estimator stacking gradient boosting and ridge regression. Doted and plain lines corresponds to 60% lowess quantile estimates. This choice of quantile allows to see better the oracle metrics lines for which outliers with a value of 0 distord the curves.

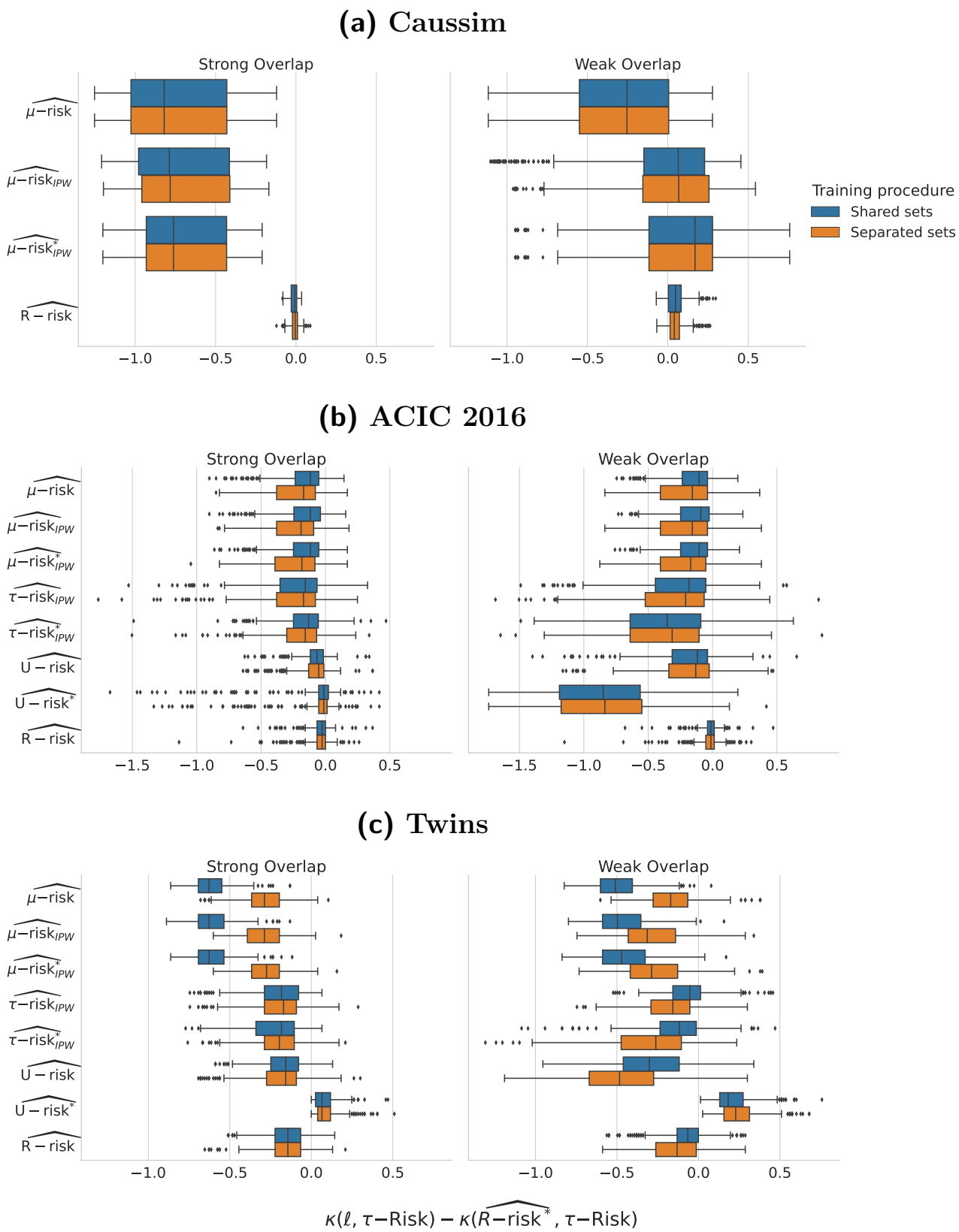


Fig. E.8. Results are similar between the Shared nuisances/candidate set and the Separated nuisances set procedure. The experience has not been run on the full metrics for Caussim due to computation costs.

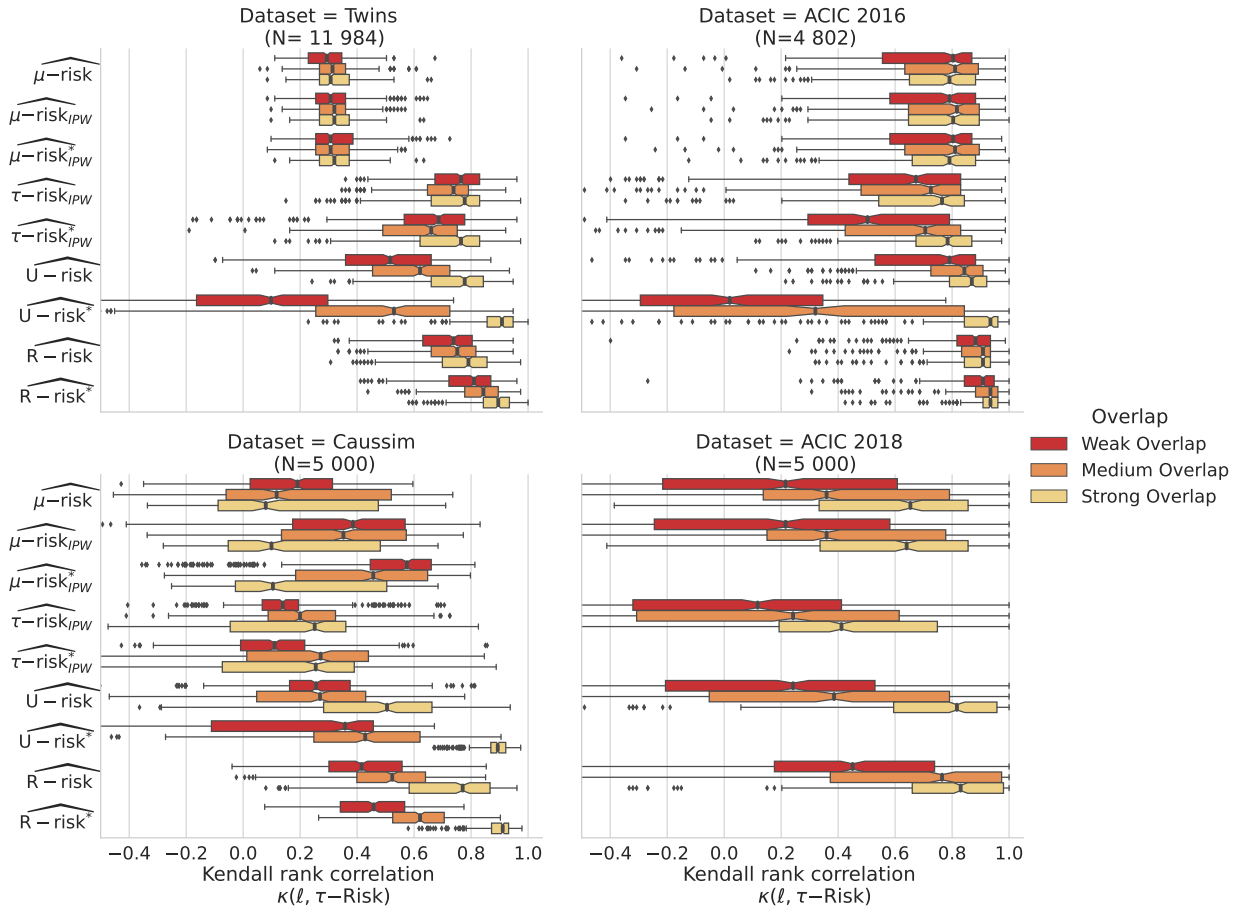


Fig. E.9. Low population overlap hinders causal model selection for all metrics: Kendall's τ agreement with τ-risk. Strong, medium and weak overlap correspond to the tertiles of the overlap distribution measured with Normalized Total Varation eq. E.2.

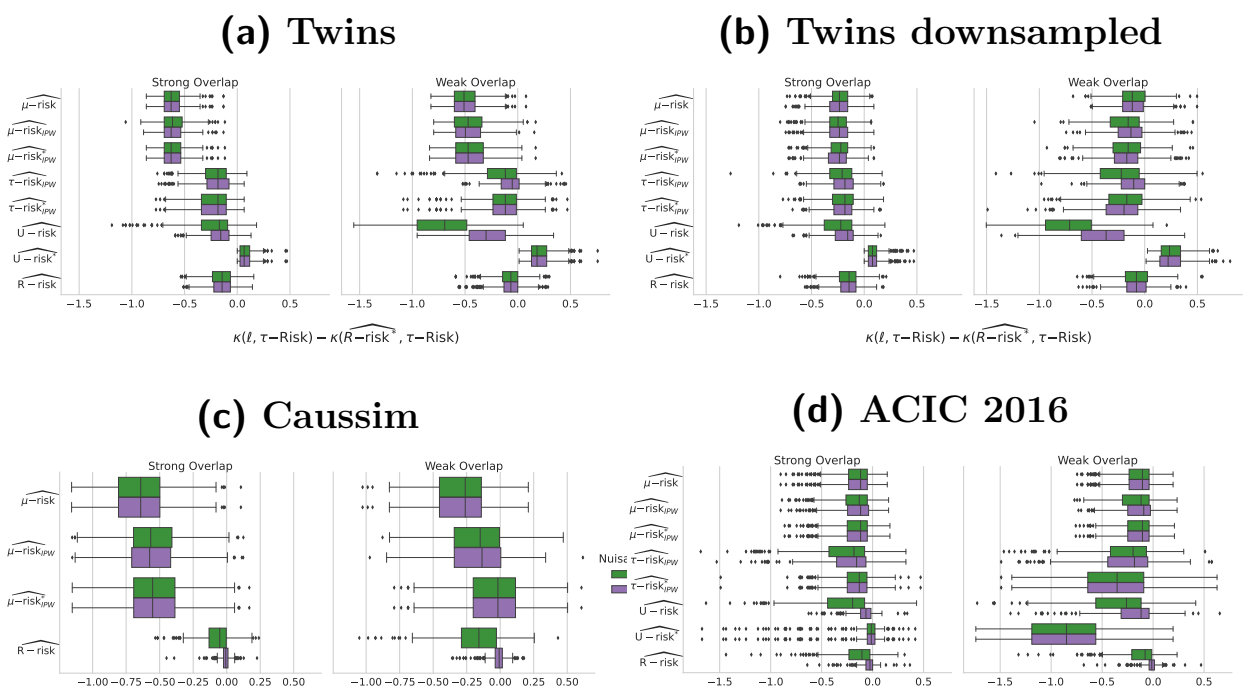


Fig. E.10. Learning the nuisances with stacked models (linear and gradient boosting) is important for successful model selection with R-risk. For Twins dataset, there is no improvement for stacked models compared to linear models because of the linearity of the propensity model.

Figure E.11 – Flexible models are performant in recovering nuisances even in linear setups

Selecting different seeds and parameters is crucial to draw conclusions One strength of our study is the various number of different simulated and semi-simulated datasets. We are convinced that the usual practice of using only a small number of generation processes does not allow to draw statistically significant conclusions.

Figure E.12 illustrate the dependence of the results on the generation process for caussim simulations. We highlighted the different trajectories induced by three different seeds for data generation and three different treatment ratio instead of 1000 different seeds. The result curves are relatively stable from one setup to another for $R{-}risk$, but vary strongly for μ-risk and μ-risk$_{IPW}$.

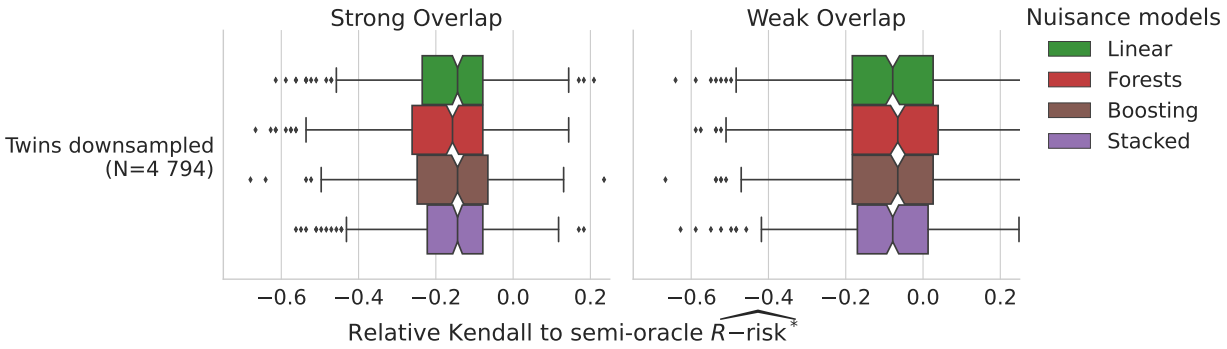
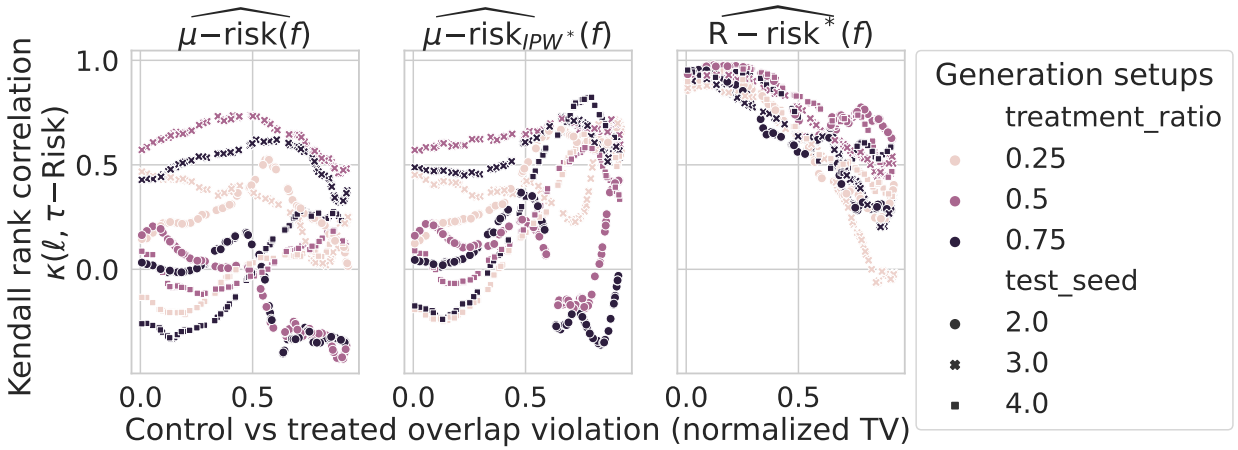


Fig. E.11. Flexible models are performant in recovering nuisances in the downsampled Twins dataset. The propensity score is linear in this setup, making it particularly challenging for flexible models compared to linear methods.

Fig. E.12. Kendall correlation coefficients for each causal metric. Each (color, shape) pair indicates a different (treatment ratio, seed) of the generation process.



E.5 Heterogeneity in practices for data split

Splitting the data is common when using machine learning for causal inference, but practices vary widely in terms of the fraction of data to allocate to train models, outcomes and nuisances, and to evaluate them.

Before even model selection, data splitting is often required for estimation of the treatment effect, ATE or CATE, for instance to compute the nuisances required to optimize the outcome model (as the R-risk, definition 6). The most frequent choice is use 80% of the data to fit the models, and 20% to evaluate them. For instance, for CATE estimation, the R-learner has been introduced using K-folds with K = 5 and K = 10: 80% of the data (4 folds) to train the nuisances and the remaining fold to minimize the corresponding R-loss (Nie; Wager, 2017). Yet, it has been implemented with K=5 in causallib (Shimoni et al., 2019) or K=3 in econML (Battocchi et al., 2019). Likewise, for ATE estimation, Chernozhukov et al., 2018a introduce doubly-robust machine learning, recommending K=5 based on an empirical comparison K=2. However, subsequent works use doubly robust ML with varying choices of K: Loiseau et al., 2022 use K=3, Gao et al., 2021 use K=2. In the econML implementation, K is set to 3 (Battocchi et al., 2019). Naimi et al., 2021 evaluate various machine-learning approaches –including R-learners– using K=5 and 10, drawing inspiration from the TMLE literature which sets K=5 in the TMLE package (Gruber; van der Laan, 2012).

Causal model selection has been much less discussed. The only study that we are aware of, Schuler et al. (2018), use a different data split: a 2-folds train/test procedure, training the nuisances on the first half of the data, and using the second half to estimate the R-risk and select the best treatment effect model.

Bibliography

AAMC (2021): *The Complexities of Physician Supply and Demand: Projections From 2019 to 2034*. Association of American Medical Colleges. URL: https://www.aamc.org/media/54681/download (cited p. 10).

A. Abadie; G. W. Imbens (2008): "On the failure of the bootstrap for matching estimators". In: *Econometrica* 76.6, pp. 1537–1557 (cited p. 124).

A. Achille; S. Soatto (2018): "Emergence of invariance and disentanglement in deep representations". In: *The Journal of Machine Learning Research* 19.1, pp. 1947–1980 (cited p. 8).

H. Adam, M. Y. Yang, K. Cato, I. Baldini, C. Senteio, L. A. Celi, J. Zeng, M. Singh; M. Ghassemi (2022): "Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations". In: *AAAI/ACM Conference on AI, Ethics, and Society* (cited p. 60).

M. Adibuzzaman, Y. Jung, E. Bareinboim, P. Griffin, S. Kethireddy, M. Bikak; R. Kaplan (2019): "Methods for quantifying efficacy-effectiveness gap of randomized controlled trials: examples in ards". In: *Critical Care Medicine* 47.1, p. 143 (cited p. 121).

J. Adler-Milstein, A. J. Holmgren, P. Kralovec, C. Worzala, T. Searcy; V. Patel (2017): "Electronic health record adoption in US hospitals: the emergence of a digital advanced use divide". In: *Journal of the American Medical Informatics Association* 24.6, pp. 1142–1148 (cited p. 4).

R. Aggarwal, V. Sounderajah, G. Martin, D. S. Ting, A. Karthikesalingam, D. King, H. Ashrafian; A. Darzi (2021): "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis". In: *NPJ digital medicine* 4.1, p. 65 (cited p. 48).

D. Agniel, I. S. Kohane; G. M. Weber (2018): "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study". In: *Bmj* 361 (cited p. 37).

A. Alaa; M. V. D. Schaar (2019): "Validating Causal Inference Models via Influence Functions". In: *International Conference on Machine Learning*, pp. 191–201 (cited p. 67, 73).

D. Almond, K. Y. Chay; D. S. Lee (2005): "The Costs of Low Birth Weight". In: *The Quarterly Journal of Economics* 120.3 (cited p. 75).

D. G. Altman, Y. Vergouwe, P. Royston; K. G. Moons (2009): "Prognosis and prognostic research: validating a prognostic model". In: *Bmj* 338 (cited p. 64).

M. Anguis, M. Bergeat, J. Pisarik, N. Vergier, H. Chaput, M. Monziols, et al. (2021): "Quelle démographie récente et à venir pour les professions médicales et pharmaceutique". In: *N* 76, p. 74 (cited p. 10).

D. Annane, S. Siami, S. Jaber, C. Martin, S. Elatrous, A. D. Declere, J. C. Preiser, H. Outin, G. Troche, C. Charpentier, et al. (2013): "Effects of fluid resuscitation with colloids vs crystalloids on mortality in critically ill patients presenting with hypovolemic shock: the CRISTAL randomized trial". In: *Jama* 310.17, pp. 1809–1817 (cited p. 55, 60).

N. C. Apathy, A. J. Holmgren; J. Adler-Milstein (2021): "A decade post-HITECH: Critical access hospitals have electronic health records but struggle to keep up with other advanced functions". In: *Journal of the American Medical Informatics Association* 28.9, pp. 1947–1954 (cited p. 3).

S. Artemova, P.-E. Madiot, A. Caporossi, PREDIMED group, P. Mossuz; A. Moreau-Gaudry (2019): "PREDIMED: Clinical Data Warehouse of Grenoble Alpes University Hospital". In: *Studies in Health Technology and Informatics* 264, pp. 1421–1422 (cited p. 23).

S. Athey; G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy*

of Sciences 113.27, pp. 7353–7360 (cited p. 70, 76).

S. Athey; G. W. Imbens (2006): "Identification and inference in nonlinear difference-in-differences models". In: *Econometrica* 74.2, pp. 431–497 (cited p. 82).

S. Athey, J. Tibshirani; S. Wager (2019): "Generalized random forests". In: *Annals of Statistics* 47.2, pp. 1148–1178 (cited p. 65).

S. Athey; S. Wager (2021): "Policy learning with observational data". In: *Econometrica* 89.1, pp. 133–161 (cited p. 82).

J. Attia, E. Holliday; C. Oldmeadow (2022): *A proposal for capturing interaction and effect modification using DAGs* (cited p. 53).

J.-M. Aubert, S. Billet, C. Colin, E. Fery Lemonnier, D. Guidoni, E. Hatton, J. Hubert, N. Lemaire, C. Marty-Chastan, M. Doutreligne; D. Claire-Lise (2019): *Réformes des modes de financement et de régulation*. Ministère des solidarités et de la Santé. URL: https://sante.gouv.fr/IMG/pdf/dicom_rapport_final_vdef_2901.pdf (cited p. 2, 17).

P. C. Austin; E. A. Stuart (2015): "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies". In: *Statistics in medicine* 34.28, pp. 3661–3679 (cited p. 54, 64, 76, 124, 130).

— (2017): "Estimating the effect of treatment on binary outcomes using full matching on the propensity score". In: *Statistical methods in medical research* 26.6, pp. 2505–2525 (cited p. 80).

P. C. Austin (2011): "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". In: *Multivariate Behavioral Research* 3, pp. 399–424 (cited p. 76).

A. J. Averitt, C. Weng, P. Ryan; A. Perotte (2020): "Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations". In: *NPJ digital medicine* 3.1, p. 67 (cited p. 48).

E. Bacry, S. Gaiffas, F. Leroy, M. Morel, D.-P. Nguyen, Y. Sebiat; D. Sun (2020): "SCALPEL3: a scalable open-source library for healthcare claims databases". In: *International Journal of Medical Informatics* 141, p. 104203 (cited p. 2, 17, 36, 39).

N. D. Bankhead C Aronson JK (2017): *Attrition bias, Catalogue of Bias Collaboration*. URL: https://catalogofbias.org/biases/attrition-bias/ (cited p. 51).

E. Bareinboim, A. Forney; J. Pearl (2015): "Bandits with unobserved confounders: A causal approach". In: *Advances in Neural Information Processing Systems* 28 (cited p. 6).

K. Battocchi, E. Dillon, M. Hei, G. Lewis, P. Oka, M. Oprescu; V. Syrgkanis (2019): *EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation*. Version 0.14.0. URL: %5Curl%7Bhttps://github.com/py-why/EconML%7D (cited p. 57, 153).

A. L. Beam; I. S. Kohane (2018): "Big data and machine learning in health care". In: *Jama* 319.13, pp. 1317–1318 (cited p. 11, 40, 64).

A. L. Beam, J. Drazen, I. S. Kohane, T.-Y. Leong, A. K. Manrai; E. J. Rubin (2023): "Artificial Intelligence in Medicine". In: *New England Journal of Medicine* 388.13, pp. 1220–1221 (cited p. 9).

A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai; I. S. Kohane (2019): "Clinical concept embeddings learned from massive sources of multimodal medical data". In: *Pacific Symposium on Biocomputing 2020*. World Scientific, pp. 295–306 (cited p. 2, 17, 39, 94, 97, 102, 103).

B. K. Beaulieu-Jones, W. Yuan, G. A. Brat, A. L. Beam, G. Weber, M. Ruffin; I. S. Kohane (2021): "Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?" In: *NPJ digital medicine* 4.1, p. 62 (cited p. 10, 48, 94, 97).

R. Bellman (1957): *Dynamic Programming*. Princeton University Press. ISBN: 069107951X (cited p. 2, 17).

A. Belloni, V. Chernozhukov; C. Hansen (2014): "High-dimensional methods and inference on structural and treatment effects". In: *Journal of Economic Perspectives* 28.2, pp. 29–50 (cited p. 49).

Y. Bengio, A. Courville; P. Vincent (2013): "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828 (cited p. 8).

C. C. Bennett; K. Hauser (2013): "Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach". In: *Artificial intelligence in medicine* 57.1, pp. 9–19 (cited p. 6).

M. L. Berger, H. Sox, R. J. Willke, D. L. Brixner, H.-G. Eichler, W. Goettsch, D. Madigan, A. Makady, S. Schneeweiss, R. Tarricone, et al. (2017): "Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making". In: *Value in Health* 20.8, pp. 1003–1008 (cited p. 30).

N. Black (1996): "Why we need observational studies to evaluate the effectiveness of health care". In: *Bmj* 312.7040, pp. 1215–1218 (cited p. 64).

T. Blakely, J. Lynch, K. Simons, R. Bentley; S. Rose (2020): "Reflection on modern methods: when worlds collide–prediction, machine learning and causal inference". In: *International journal of epidemiology* 49.6, pp. 2058–2064 (cited p. 64).

P. Bojanowski, E. Grave, A. Joulin; T. Mikolov (2017): "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* 5, pp. 135–146 (cited p. 103).

J. Bor, E. Moscoe, P. Mutevedzi, M.-L. Newell; T. Bärnighausen (2014): "Regression discontinuity designs in epidemiology: causal inference without randomized trials". In: *Epidemiology (Cambridge, Mass.)* 25.5, p. 729 (cited p. 82).

E. Bosco, L. Hsueh, K. W. McConeghy, S. Gravenstein; E. Saade (2021): "Major adverse cardiovascular event definitions used in observational analysis of administrative databases: a systematic review". In: *BMC Medical Research Methodology* 21.1, pp. 1–18 (cited p. 100).

R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, et al. (2020): "Variability in the analysis of a single neuroimaging dataset by many teams". In: *Nature* 582.7810, pp. 84–88 (cited p. 7).

X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, et al. (2021a): "Accounting for variance in machine learning benchmarks". In: *Proceedings of Machine Learning and Systems* 3, pp. 747–769 (cited p. 128).

X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, S. Ebrahimi Kahou, V. Michalski, T. Arbel, C. Pal, G. Varoquaux; P. Vincent (2021b): "Accounting for Variance in Machine Learning Benchmarks". In: *Proceedings of Machine Learning and Systems* 3, pp. 747–769 (cited p. 137).

C. M. Boyd, J. Darer, C. Boult, L. P. Fried, L. Boult; A. W. Wu (2005): "Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance". In: *Jama* 294.6, pp. 716–724 (cited p. 11).

R. J. Brand, R. H. Rosenman, R. I. Sholtz; M. Friedman (1976): "Multivariate prediction of coronary heart disease in the Western Collaborative Group Study compared to the findings of the Framingham study." In: *Circulation* 53.2, pp. 348–355 (cited p. 11, 36, 93).

M. L. Braunstein (2019): "Health Care in the Age of Interoperability Part 6: The Future of FHIR". In: *IEEE Pulse* 4, pp. 25–27 (cited p. 29).

L. Breiman (2001a): "Random forests". In: *Machine learning* 45, pp. 5–32 (cited p. 86).

L. Breiman (2001b): “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”. In: *Statistical science* 16.3, pp. 199–231 (cited p. 1, 2, 6, 16, 17).

N. Breznau, E. M. Rinke, A. Wuttke, H. H. Nguyen, M. Adem, J. Adriaans, A. Alvarez-Benjumea, H. K. Andersen, D. Auer, F. Azevedo, et al. (2022): “Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty”. In: *Proceedings of the National Academy of Sciences* 119.44, e2203150119 (cited p. 59).

A. J. Butte; I. S. Kohane (2006): “Creation and implications of a phenome-genome network”. In: *Nature biotechnology* 24.1, pp. 55–62 (cited p. 4).

N. Caetano, R. M. Laureano; P. Cortez (2014): “A data-driven approach to predict hospital length of stay-a portuguese case study”. In: *International Conference on Enterprise Information Systems*. Vol. 2, pp. 407–414 (cited p. 99).

X. Cai, J. Gao, K. Y. Ngiam, B. C. Ooi, Y. Zhang; X. Yuan (2018): “Medical concept embedding with time-aware attention”. In: *arXiv preprint arXiv:1806.02873* (cited p. 95).

P. Caironi, G. Tognoni, S. Masson, R. Fumagalli, A. Pesenti, M. Romero, C. Fanizza, L. Caspani, S. Faenza, G. Grasselli, et al. (2014): “Albumin replacement in patients with severe sepsis or septic shock”. In: *New England Journal of Medicine* 370.15, pp. 1412–1421 (cited p. 55, 59, 60, 121, 122, 133).

C. .-. Caisse Nationale d’Assurance Maladie (n.d.): *Méthodologie médicale de la cartographie des pathologies et des dépenses, version G9*. URL: https://assurance-maladie.ameli.fr/sites/default/files/2022_methode-reperage-pathologies_cartographie_0.pdf (cited p. 100).

D. T. Campbell (1957): “Factors relevant to the validity of experiments in social settings”. In: *Sociological methods*, pp. 243–263 (cited p. 5).

E. A. Campbell, M. G. Maltenfort, J. Shults, C. B. Forrest; A. J. Masino (2022): “Characterizing clinical pediatric obesity subtypes using electronic health record data”. In: *PLOS Digital Health* 1 (cited p. 10).

Canadian Medical Association (2015): “Appropriateness in Health Care”. In: *Canadian Medical Association Policy* (cited p. 2, 10, 13, 17, 48).

A. Caruana, M. Bandara, K. Musial, D. Catchpoole; P. J. Kennedy (2023): “Machine learning for administrative health records: A systematic review of techniques and applications”. In: *Artificial Intelligence in Medicine*, p. 102642 (cited p. 82).

J. A. Casey, B. S. Schwartz, W. F. Stewart; N. E. Adler (2016): “Using electronic health records for population health research: a review of methods and applications”. In: *Annual review of public health* 37, pp. 61–81 (cited p. 4).

A. L. Celi, J. Ken, G. Marzyeh, C. Guzman, U. Shalit; D. Sontag (2016): *An Open Benchmark for Causal Inference Using the MIMIC-III Dataset* (cited p. 122).

M. E. Charlson, P. Pompei, K. L. Ales; C. MacKenzie (1987): “A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation”. In: *Journal of Chronic Diseases* 40.5, pp. 373–383 (cited p. 4, 66).

E. Chazard, P. Balaye, T. Balcaen, M. Genin, M. Cuggia, G. Bouzille; A. Lamer (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 (cited p. 39).

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey; J. Robins (2018a): “Double/Debiased Machine Learning for Treatment and Structural Parameters”. In: *The Econometrics Journal*, p. 71 (cited p. 65, 68, 69, 80, 153).

— (2018b): *Double/debiased machine learning for treatment and structural parameters* (cited p. 49, 54, 126).

V. Chernozhukov, W. Newey, V. M. Quintas-Martınez; V. Syrgkanis (2022): "Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests". In: *International Conference on Machine Learning*. PMLR, pp. 3901–3914 (cited p. 82).

E. E. Chinaeke, B. L. Love, J. Magagnoli, I. Yunusa; G. Reeder (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 (cited p. 121).

E. Choi, A. Schuetz, W. F. Stewart; J. Sun (2017): "Using recurrent neural network models for early detection of heart failure onset". In: *Journal of the American Medical Informatics Association* 24.2, pp. 361–370 (cited p. 94).

CHoRUS (2023): *A Patient-Focused CHoRUS for Equitable AI*. URL: https://chorus4ai.org/ (cited p. 22).

C. G. Chute, S. A. Beck, T. B. Fisk; D. N. Mohr (2010): "The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data". In: *Journal of the American Medical Informatics Association* 17, pp. 131–135 (cited p. 23).

Clalit (2023): *Clalit Research Institute*. URL: http://clalitresearch.org/about-us/our-data/ (cited p. 22).

CMS for Medicare (2019): *Readmissions-Reduction-Program*. URL: https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-%20reduction-program.html (cited p. 94).

CNAM (2023): *Data pathologies*. URL: https://data.ameli.fr/pages/data-pathologies/ (cited p. 37).

A. L. Cochrane (1972): "Effectiveness and efficiency: random reflections on health services". In: (cited p. 6).

B. Colnet, J. Josse, G. Varoquaux; E. Scornet (2023): "Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize?" In: *arXiv preprint arXiv:2303.16008* (cited p. 53, 80).

B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse; S. Yang (2020): "Causal inference methods for combining randomized trials and observational studies: a review". In: *arXiv preprint arXiv:2011.08047* (cited p. 5).

Y. Conan, J. Herbert, C. Salpetrier, L. Godillon, F. Fourquet, T. Dhalluin, E. Laurent; L. Grammatico-Guillon (2021): "Les entrepôts de données cliniques : un outil d'aide au pilotage de crise". In: *Infectious Diseases Now* 51.5, S56 (cited p. 23).

J. Concato, N. Shah; R. I. Horwitz (2000): "Randomized, controlled trials, observational studies, and the hierarchy of research designs". In: *New England journal of medicine* 342.25, pp. 1887–1892 (cited p. 5).

K. A. Corl, M. Prodromou, R. C. Merchant, I. Gareen, S. Marks, D. Banerjee, T. Amass, A. Abbasi, C. Delcompare, A. Palmisciano, et al. (2019): "The Restrictive Intravenous Fluid Trial in Severe Sepsis and Septic Shock (RIFTS): a Randomized Pilot Study". In: *Critical care medicine* 47.7, p. 951 (cited p. 121).

A. Coronato, M. Naeem, G. De Pietro; G. Paragliola (2020): "Reinforcement learning for intelligent healthcare applications: A survey". In: *Artificial Intelligence in Medicine* 109, p. 101964 (cited p. 6).

M. R. Cowie, J. I. Blomster, L. H. Curtis, S. Duclaux, I. Ford, F. Fritz, S. Goldman, S. Janmohamed, J. Kreuzer, M. Leenay, et al. (2017): "Electronic health records to facilitate clinical research". In: *Clinical Research in Cardiology* 106, pp. 1–9 (cited p. 4).

D. R. Cox (2001): "statistical modeling: The two cultures: Comment". In: *Statistical science* 16.3, pp. 199–231 (cited p. vii, 1, 6, 7, 16).

— (2006): *Principles of statistical inference*. Cambridge university press (cited p. 7).

M. Cuggia, N. Garcelon, B. Campillo-Gimenez, T. Bernicot, J.-F. Laurent, E.

Garin, A. Happe; R. Duvauferrier (2011): "Roogle: an information retrieval engine for clinical data warehouse". In: *Studies in Health Technology and Informatics* 169, pp. 584–588 (cited p. 23).

A. Curth, D. Svensson; J. Weatherall (2021): "Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation". In: *Neurips Process 2021*, p. 14 (cited p. 75).

A. D'Amour, P. Ding, A. Feller, L. Lei; J. Sekhon (2021): "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* 221.2, pp. 644–654 (cited p. 52, 67, 76).

C. Daniel, P. Serre, N. Orlova, S. Bréant, N. Paris; N. Griffon (2018): "Initializing a hospital-wide data quality program. The AP-HP experience." In: *Computer Methods and Programs in Biomedicine* (cited p. 23).

R. M. Daniel (2018): "Double Robustness". In: *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, pp. 1–14 (cited p. 80).

C. P. R. Datalink (2022): *Questions and answers - EU Health: European Health Data Space (EHDS)*. URL: https://ec.europa.eu/commission/presscorner/detail/en/qanda_22_2712 (cited p. 9).

A. Deaton (2020): *Randomization in the tropics revisited: a theme and eleven variations*. Tech. rep. National Bureau of Economic Research (cited p. 6).

A. J. DeGrave, J. D. Janizek; S.-I. Lee (2021): "AI for radiographic COVID-19 detection selects shortcuts over signal". In: *Nature Machine Intelligence* 3.7, pp. 610–619 (cited p. 48).

R. J. Desai, M. E. Matheny, K. Johnson, K. Marsolo, L. H. Curtis, J. C. Nelson, P. J. Heagerty, J. Maro, J. Brown, S. Toh, et al. (2021): "Broadening the reach of the FDA Sentinel System: a roadmap for integrating electronic health record data in a causal analysis framework". In: *NPJ digital medicine* 4.1, p. 170 (cited p. 48).

R. J. Desai, S. V. Wang, M. Vaduganathan, T. Evers; S. Schneeweiss (2020): "Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes". In: *JAMA network open* 3.1, e1918962–e1918962 (cited p. 64).

J. Devlin, M.-W. Chang, K. Lee; K. Toutanova (2018): "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (cited p. 2, 16, 98).

V. Dorie, J. Hill, U. Shalit, M. Scott; D. Cervone (2019): "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition". In: *Statistical Science* 34.1, pp. 43–68 (cited p. 54, 64, 65, 67, 75, 142).

M. Doutreligne, A. Leduc, D.-P. Nguyen; A. Vuagnat (2021): "Representations of medical concepts learned from 3 millions patients in the French National Health Insurance Database, SNDS (2008-2016)". In: *Data Science Summer School* (cited p. 40).

M. Doutreligne; G. Varoquaux (2023): "How to select predictive models for causal inference?" In: *arXiv preprint arXiv:2302.00370* (cited p. 54, 131).

V. J. Dzau (2023): "Anticipating the Future of Health and Medicine—The National Academy of Medicine Prepares for Its Next 50 Years". In: *JAMA* 329.17, pp. 1445–1446 (cited p. 9).

EC (2022): *European Health Data Space*. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en (cited p. 32).

D. E. Ehrmann, S. Joshi, S. D. Goodfellow, M. L. Mazwi; D. Eytan (2023): "Making machine learning matter to clinicians: model actionability in medical decision-making". In: *NPJ Digital Medicine* 6.1, p. 7 (cited p. 60).

EMA (2023): *Real-world evidence framework to support EU regulatory decision-making*. Tech. rep. European Medicines Agency. URL: https://www.ema.europa.eu/

en/news/use-real-world-evidence-regulatory-decision-making-ema-publishes-review-its-studies (cited p. 9).

M. Esdar, J. Hüsers, J.-P. Weiß, J. Rauch; U. Hübner (2019): "Diffusion dynamics of electronic health records: A longitudinal observational study comparing data from hospitals in Germany and the United States". In: *International Journal of Medical Informatics* 131 (cited p. 3).

A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean; R. Socher (2021): "Deep learning-enabled medical computer vision". In: *NPJ digital medicine* 4.1, p. 5 (cited p. 48).

FDA (2018): *Real World Evidence Program*. FDA. URL: https://www.fda.gov/media/120060/download (cited p. 9, 10).

— (2021a): *Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products*. FDA, p. 39 (cited p. 3, 10, 30).

— (2021b): *Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials*. Tech. rep. FDA (cited p. 54).

A. R. Feinstein; R. I. Horwitz (1997): "Problems in the "evidence" of "evidence-based medicine"". In: *The American journal of medicine* 103.6, pp. 529–535 (cited p. 5).

M. Feng, J. I. McSparron, D. T. Kien, D. J. Stone, D. H. Roberts, R. M. Schwartzstein, A. Vieillard-Baron; L. A. Celi (2018): "Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database". In: *Intensive care medicine* 44, pp. 884–892 (cited p. 120).

S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane; S. Saria (2021): "The clinician and dataset shift in artificial intelligence". In: *The New England journal of medicine* 385.3, p. 283 (cited p. 37).

M. A. Fontana, S. Lyman, G. K. Sarker, D. E. Padgett; C. H. MacLean (2019): "Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty?" In: *Clinical orthopaedics and related research* 477.6, p. 1267 (cited p. 64).

D. J. Foster; V. Syrgkanis (2019): "Orthogonal statistical learning". In: *arXiv preprint arXiv:1901.09036* (cited p. 126).

Y. Freund; R. E. Schapire (1995): "A desicion-theoretic generalization of on-line learning and an application to boosting". In: *European conference on computational learning theory*. Springer, pp. 23–37 (cited p. 87).

J. H. Friedman (2001): "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232 (cited p. 87).

E. L. Fu, M. Evans, J.-J. Carrero, H. Putter, C. M. Clase, F. J. Caskey, M. Szymczak, C. Torino, N. C. Chesnaye, K. J. Jager, et al. (2021): "Timing of dialysis initiation to reduce mortality and cardiovascular events in advanced chronic kidney disease: nationwide cohort study". In: *bmj* 375 (cited p. 51).

J. Futoma, M. Simons, T. Panch, F. Doshi-Velez; L. A. Celi (2020): "The myth of generalisability in clinical research and machine learning in health care". In: *The Lancet Digital Health* 2.9, e489–e492 (cited p. 37).

M. O. Gani, S. Kethireddy, R. Adib, U. Hasan, P. Griffin; M. Adibuzzaman (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 (cited p. 120).

Z. Gao, T. Hastie; R. Tibshirani (2021): "Assessment of heterogeneous treatment effect estimation accuracy via matching". In: *Statistics in Medicine* 17 (cited p. 153).

N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, S. Kracker, F. Suarez, N. Bahi-Buisson, S. Hadj-Rabia, A. Fischer, A. Munnich; A. Burgun (2017): "Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the nee-

dle stack". In: *Journal of Biomedical Informatics* 73, pp. 51–61 (cited p. 23, 29).
T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii; K. Crawford (2021): "Datasheets for datasets". In: *Communications of the ACM* 64, pp. 86–92 (cited p. 32).
S. Gehring; R. Eulenfeld (2018): "German Medical Informatics Initiative: Unlocking Data for Research and Health Care". In: *Methods of information in medicine* 57 (S 01), e46–e49 (cited p. 22).
R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge; F. A. Wichmann (2020): "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11, pp. 665–673 (cited p. 48).
M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen; R. Ranganath (2020): "A review of challenges and opportunities in machine learning for health". In: *AMIA Summits on Translational Science Proceedings* 2020, p. 191 (cited p. 48).
M. A. Gianfrancesco; N. D. Goldstein (2021): "A narrative review on the validity of electronic health record-based research in epidemiology". In: *BMC medical research methodology* 21.1, pp. 1–10 (cited p. 4).
M. A. Gianfrancesco, S. Tamang, J. Yazdany; G. Schmajuk (2018): "Potential biases in machine learning algorithms using electronic health record data". In: *JAMA internal medicine* 178.11, pp. 1544–1547 (cited p. 37).
J. W. Gichoya, I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, et al. (2022): "AI recognition of patient race in medical imaging: a modelling study". In: *The Lancet Digital Health* 4.6, e406–e414 (cited p. 48).
B. Goldacre, J. Morley; N. Hamilton (2022): *Better, Broader, Safer: Using Health Data for Research and Analysis*. Secretary of State for Health and Social Care. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067058/summary-goldacre-review-using-health-data-for-research-and-analysis.pdf (cited p. 27, 30, 32).
B. A. Goldstein, A. M. Navar, M. J. Pencina; J. P. Ioannidis (2017): "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review". In: *Journal of the American Medical Informatics Association: JAMIA* 24.1, p. 198 (cited p. 37, 94).
Google (2023): *Tensorflow Responsible AI*. URL: https://www.tensorflow.org/responsible_ai (cited p. 60).
S. Greenland (2000): "An introduction to instrumental variables for epidemiologists". In: *International journal of epidemiology* 29.4, pp. 722–729 (cited p. 82).
S. Greenland, J. Pearl; J. M. Robins (1999): "Causal diagrams for epidemiologic research". In: *Epidemiology*, pp. 37–48 (cited p. 53).
A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf; A. Smola (2012): "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1, pp. 723–773 (cited p. 139).
L. Grinsztajn, E. Oyallon; G. Varoquaux (2022): "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Advances in Neural Information Processing Systems* 35, pp. 507–520 (cited p. 81).
E. Grose, S. Wilson, J. Barkun, K. Bertens, G. Martel, F. Balaa; J. A. Khalil (2020): "Use of Propensity Score Methodology in Contemporary High-Impact Surgical Literature". In: *Journal of the American College of Surgeons* 230.1, 101–112.e2 (cited p. 64).
S. Gruber; M. J. van der Laan (2012): "tmle: An R Package for Targeted Maximum Likelihood Estimation". In: *Journal of Statistical Software* 51.13, pp. 1–35 (cited p. 153).
A. Guez, R. D. Vincent, M. Avoli; J. Pineau (2008): "Adaptive Treatment of Epilepsy via Batch-mode Reinforcement Learning." In: *AAAI*. Vol. 8, pp. 1671–1678 (cited p. 6).

T. D. Gunter; N. P. Terry (2005): "The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions". In: *Journal of medical Internet research* 7.1, e383 (cited p. 4).

P. Gutierrez; J.-Y. Gerardy (2016): "Causal Inference and Uplift Modeling A review of the literature". In: *Proceedings of The 3rd International Conference on Predictive Applications and APIs*. Proceedings of Machine Learning Research 67, p. 14 (cited p. 67, 70, 76).

G. H. Guyatt, D. L. Sackett, J. C. Sinclair, R. Hayward, D. J. Cook, R. J. Cook, E. Bass, H. Gerstein, B. Haynes, A. Holbrook, et al. (1995): "Users' guides to the medical literature: IX. A method for grading health care recommendations". In: *Jama* 274.22, pp. 1800–1804 (cited p. 3).

A. Halevy, P. Norvig; F. Pereira (2009): "The unreasonable effectiveness of data". In: *IEEE intelligent systems* 24.2, pp. 8–12 (cited p. 1, 9, 16).

F. E. Harrell et al. (2001): *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Vol. 608. Springer (cited p. 93).

H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg; A. Galstyan (2019): "Multitask learning and benchmarking with clinical time series data". In: *Scientific data* 6.1, p. 96 (cited p. 94).

HAS (2020): *Guide méthodologique impacts organisationnels*. Haute Autorité de Santé. URL: https://www.has-sante.fr/upload/docs/application/pdf/2020-12/guide_methodologique_impacts_organisationnels.pdf (cited p. 10).

— (2021): *Real-world studies for the assessment of medicinal products and medical devices*. Haute Autorité de Santé, p. 50. URL: https://www.has-sante.fr/upload/docs/application/pdf/2021-06/real-world_studies_for_the_assessment_of_medicinal_products_and_medical_devices.pdf (cited p. 3, 10, 30).

T. Hastie, R. Tibshirani, J. H. Friedman; J. H. Friedman (2009): *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer (cited p. 85).

HDH (2023a): *Doucmentations collaboratives du SNDS*. URL: https://documentation-snds.health-data-hub.fr/ (cited p. 2, 17).

— (2023b): *Repertoire public des projets du Health Data Hub*. URL: https://www.health-data-hub.fr/projets (cited p. 27).

J. M. Hendriksen, G.-J. Geersing, K. G. Moons; J. A. de Groot (2013): "Diagnostic and prognostic prediction models". In: *Journal of Thrombosis and Haemostasis* 11, pp. 129–141 (cited p. 80).

M. A. Hernan (2021): "Methods of public health research–strengthening causal inference from observational data". In: *New England Journal of Medicine* 385.15, pp. 1345–1348 (cited p. 50).

M. A. Hernan, J. Hsu; B. Healy (2019): "A second chance to get causal inference right: a classification of data science tasks". In: *Chance* 32.1, pp. 42–49 (cited p. 49, 60).

M. A. Hernan, B. C. Sauer, S. Hernandez-Diaz, R. Platt; I. Shrier (2016): "Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses". In: *Journal of clinical epidemiology* 79, pp. 70–75 (cited p. 51).

M. Hernán; J. Robins (2020): *Causal Inference: What If*. CRC Boca Raton, FL (cited p. 52, 64).

M. A. Hernàn; J. M. Robins (2020): *Causal inference: What If*. (Cited p. 6, 12, 49–51, 54).

M. A. Hernán (2021): "Methods of Public Health Research — Strengthening Causal Inference from Observational Data". In: *New England Journal of Medicine* 385.15, pp. 1345–1348 (cited p. 25, 64).

M. A. Hernán; J. M. Robins (2016): "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available". In: *American Journal of Epidemiology* 183.8 (cited p. 9, 50, 59).

G. Hernandez, C. Vaquero, L. Colinas, R. Cuena, P. Gonzalez, A. Canabal, S.

Sanchez, M. L. Rodriguez, A. Villasclaras; R. Fernandez (2016): "Effect of postextubation high-flow nasal cannula vs noninvasive ventilation on reintubation and postextubation respiratory failure in high-risk patients: a randomized clinical trial". In: *Jama* 316.15, pp. 1565–1574 (cited p. 119).

A. B. Hill (1965): "The environment and disease: association or causation?" In: *Proceedings of the Royal Society of Medicine* (cited p. 12).

J. L. Hill (2011): "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 1, pp. 217–240 (cited p. 65, 69, 70).

J. Hippisley-Cox, C. Coupland; P. Brindle (2017): "Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study". In: *bmj* 357 (cited p. 38).

A. T. N. Ho, S. Patolia; C. Guervilly (2020): "Neuromuscular blockade in acute respiratory distress syndrome: a systematic review and meta-analysis of randomized controlled trials". In: *Journal of intensive care* 8, pp. 1–11 (cited p. 122).

A. Hoerbst; E. Ammenwerth (2010): "Electronic health records". In: *Methods of information in medicine* 49.04, pp. 320–336 (cited p. 4).

R. Hofmann, S. K. James, T. Jernberg, B. Lindahl, D. Erlinge, N. Witt, G. Arefalk, M. Frick, J. Alfredsson, L. Nilsson, et al. (2017): "Oxygen therapy in suspected acute myocardial infarction". In: *New England Journal of Medicine* 377.13, pp. 1240–1249 (cited p. 122).

A. A. de Hond, A. M. Leeuwenberg, L. Hooft, I. M. Kant, S. W. Nijman, H. J. van Os, J. J. Aardoom, T. P. Debray, E. Schuit, M. van Smeden, et al. (2022): "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review". In: *NPJ digital medicine* 5.1, p. 2 (cited p. 36).

C. Hong, E. Rush, M. Liu, D. Zhou, J. Sun, A. Sonabend, V. M. Castro, P. Schubert, V. A. Panickan, T. Cai, et al. (2021): "Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data". In: *NPJ digital medicine* 4.1, pp. 1–11 (cited p. 97, 103).

J. Hoogland, J. IntHout, M. Belias, M. M. Rovers, R. D. Riley, F. E. Harrell Jr, K. G. Moons, T. P. Debray; J. B. Reitsma (2021): "A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint". In: *Statistics in medicine* 40.26, pp. 5961–5981 (cited p. 64).

S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro; L. A. Nathanson (2017): "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning". In: *PloS one* 12.4, e0174708 (cited p. 54, 64).

C. J. Howe, S. R. Cole, D. J. Westreich, S. Greenland, S. Napravnik; J. J. Eron (2011): "Splines for trend analysis and continuous confounder control". In: *Epidemiology (Cambridge, Mass.)* 22.6, pp. 874–875 (cited p. 73).

G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, et al. (2015a): "Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers". In: *MEDINFO 2015: eHealth-enabled Health*. IOS Press, pp. 574–578 (cited p. 36, 82, 99).

G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Norén, Y.-C. Li, P. E. Stang, D. Madigan; P. B. Ryan (2015b): "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers". In: *Studies in health technology and informatics* 216, pp. 574–578 (cited p. 28, 29).

D. J. Hsu, M. Feng, R. Kothari, H. Zhou, K. P. Chen; L. A. Celi (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 (cited p. 120).

Y. Huang, J. Lee, S. Wang, J. Sun, H. Liu, X. Jiang, et al. (2018): "Privacy-preserving predictive modeling: Harmonization of contextual embeddings from different sources". In: *JMIR medical informatics* 6.2, e9455 (cited p. 103).

Hugo (2022): *Ouest Data Hub*. URL: https://www.chu-hugo.fr/accueil/wp-content/uploads/sites/2/2022/02/CP-Autorisations-CNIL-projets-ODH.pdf (cited p. 24).

G. W. Imbens; J. M. Wooldridge (2009): "Recent developments in the econometrics of program evaluation". In: *Journal of economic literature* 47.1, pp. 5–86 (cited p. 12, 49).

G. W. Imbens (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review". In: *The Review of Economics and Statistics* 86.1, pp. 4–29 (cited p. 53).

G. W. Imbens; D. B. Rubin (2015): *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press (cited p. 12).

S. S. Investigators (2007): "Saline or albumin for fluid resuscitation in patients with traumatic brain injury". In: *New England Journal of Medicine* 357.9, pp. 874–884 (cited p. 134).

J. P. Ioannidis (2005): "Why most published research findings are false". In: *PLoS medicine* 2.8, e124 (cited p. 59).

E. L. Ionides (2008): "Truncated Importance Sampling". In: *Journal of Computational and Graphical Statistics* 17.2, pp. 295–311 (cited p. 71, 76).

IQVIA (2023): *Harness the power of Real World Data*. URL: https://www.iqvia.com/solutions/real-world-evidence/real-world-data-and-insights (cited p. 9).

ISIS-1 Collaborative Group (1986): "Randomised trial of intravenous atenolol among 16 027 cases of suspected acute myocardial infarction: isis-1". In: *The Lancet* (cited p. 3).

A.-S. Jannot, E. Zapletal, P. Avillach, M.-F. Mamzer, A. Burgun; P. Degoulet (2017): "The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience". In: *International Journal of Medical Informatics* 102, pp. 21–28 (cited p. 23).

A. Jesson, S. Mindermann, U. Shalit; Y. Gal (2020): "Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models". In: *Advances in Neural Information Processing Systems* 33, pp. 11637–11649 (cited p. 52).

A. K. Jha, C. M. DesRoches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris, A. Shields, S. Rosenbaum; D. Blumenthal (2009): "Use of electronic health records in U.S. hospitals". In: *The New England Journal of Medicine* 360, pp. 1628–1638 (cited p. 3).

L. Y. Jiang, X. C. Liu, N. P. Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, K. Eaton, H. A. Riina, I. Laufer, P. Punjabi, et al. (2023): "Health system-scale language models are all-purpose prediction engines". In: *Nature*, pp. 1–6 (cited p. 10, 54, 64, 82, 97, 99).

F. D. Johansson, U. Shalit, N. Kallus; D. Sontag (2022): "Generalization bounds and representation learning for estimation of potential outcomes and causal effects". In: *The Journal of Machine Learning Research* 23.1, pp. 7489–7538 (cited p. 68, 72, 76, 82, 139).

A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi; R. Mark (2020): "Mimic-iv". In: *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)* (cited p. 49, 55).

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, et al. (2021): "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589 (cited p. 2, 16).

P.-A. Juven (2013): "Codage de la performance ou performance du codage: Mise en chiffre et optimisation de l'information médicale". In: *Journal de gestion et d'économie médicales* 31.2, pp. 75–91 (cited p. 4, 37).

T. Kanakubo; H. Kharrazi (2019): "Comparing the Trends of Electronic Health Record Adoption Among Hospitals of the United States and Japan". In: *Journal of Medical Systems* 43, p. 224 (cited p. 3).

C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado; D. King (2019): "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC medicine* 17, pp. 1–9 (cited p. 38, 81).

M. G. Kendall (1938): "A new measure of rank correlation". In: *Biometrika* 30.1-2, pp. 81–93 (cited p. 76).

E. H. Kennedy (2020): "Optimal doubly robust estimation of heterogeneous causal effects". In: *arXiv preprint arXiv:2004.14497* (cited p. 80).

T. Kennedy-Martin, S. Curtis, D. Faries, S. Robinson; J. Johnston (2015): "A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results". In: *Trials* 16, pp. 1–14 (cited p. 60).

D. M. Kent, E. Steyerberg; D. van Klaveren (2018): "Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects". In: *Bmj* 363 (cited p. 60).

S. Kent, L. Kincaid, S. Manuj, S. Rowark, S. Duffield, V. Ayyar Gupta; P. Jonsson (2022): *NICE real-world evidence framework*. National Institute for Health and Care Excellence, p. 95. URL: https://www.nice.org.uk/corporate/ecd9/resources/nice-realworld-evidence-framework-pdf-1124020816837 (cited p. 3, 10).

M. Khojaste-Sarakhsi, S. S. Haghighi, S. F. Ghomi; E. Marchiori (2022): "Deep learning for Alzheimer's disease diagnosis: A survey". In: *Artificial Intelligence in Medicine*, p. 102332 (cited p. 64).

Y.-G. Kim, K. Jung, Y.-T. Park, D. Shin, S. Y. Cho, D. Yoon; R. W. Park (2017): "Rate of electronic health record adoption in South Korea: A nation-wide survey". In: *International Journal of Medical Informatics* 101, pp. 100–107 (cited p. 3).

W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano, et al. (1991): "The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults". In: *Chest* 100.6, pp. 1619–1636 (cited p. 38).

I. S. Kohane, B. J. Aronow, P. Avillach, B. K. Beaulieu-Jones, R. Bellazzi, R. L. Bradford, G. A. Brat, M. Cannataro, J. J. Cimino, N. García-Barrio, N. Gehlenborg, M. Ghassemi, A. Gutiérrez-Sacristán, D. A. Hanauer, J. H. Holmes, C. Hong, J. G. Klann, N. H. W. Loh, Y. Luo, K. D. Mandl, M. Daniar, J. H. Moore, S. N. Murphy, A. Neuraz, K. Y. Ngiam, G. S. Omenn, N. Palmer, L. P. Patel, M. Pedrera-Jiménez, P. Sliz, A. M. South, A. L. M. Tan, D. M. Taylor, B. W. Taylor, C. Torti, A. K. Vallejos, K. B. Wagholikar, T. C. F. C. C. O. C.-1. B. Ehr (4ce), G. M. Weber; T. Cai (2021): "What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask". In: *Journal of Medical Internet Research* 23 (cited p. 31, 32).

M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon; A. A. Faisal (2018): "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care". In: *Nature medicine* 24.11, pp. 1716–1720 (cited p. 6).

K. Kreis, S. Neubauer, M. Klora, A. Lange; J. Zeidler (2016): "Status and perspectives of claims data analyses in Germany—a systematic review". In: *Health policy* 120.2, pp. 213–226 (cited p. 22).

S. R. Kristensen, M. Bech; W. Quentin (2015): "A roadmap for comparing readmission policies with application to Denmark, England, Germany and the United States".

In: *Health policy* 119.3, pp. 264–273 (cited p. 94).

A. Krizhevsky, I. Sutskever; G. E. Hinton (2012): "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (cited p. 2, 8, 16).

S. R. Künzel, J. S. Sekhon, P. J. Bickel; B. Yu (2019): "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the National Academy of Sciences* 116.10, pp. 4156–4165 (cited p. 65, 70, 74, 75).

D.-S. Kyoung; H.-S. Kim (2022): "Understanding and utilizing claim data from the Korean National Health Insurance Service (NHIS) and Health Insurance Review & Assessment (HIRA) database for research". In: *Journal of Lipid and Atherosclerosis* 11.2, p. 103 (cited p. 22).

T. Kyu Oh, I.-A. Song, J. H. Lee, C. Lim, Y.-T. Jeon, H.-J. Bae, Y. H. Jo; H.-J. Jee (2019): "Preadmission statin use and 90-day mortality in the critically ill: a retrospective association study". In: *Anesthesiology* 131.2, pp. 315–327 (cited p. 121).

M. J. van der Laan, M. Laan; J. Robins (2003): *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media (cited p. 69, 70).

M. J. v. d. Laan, E. C. Polley; A. E. Hubbard (2007): "Super Learner". In: *Statistical Applications in Genetics and Molecular Biology* 6 (cited p. 144).

M. J. v. d. Laan; S. Rose (2011): *Targeted Learning*. Springer Series in Statistics (cited p. 65, 68).

A. Lamer, M. Moussa, R. Marcilly, R. Logier, B. Vallet; B. Tavernier (2022): "Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project". In: *Journal of Clinical Monitoring and Computing* (cited p. 23).

A. Lamont, M. D. Lyons, T. Jaki, E. Stuart, D. J. Feaster, K. Tharmaratnam, D. Oberski, H. Ishwaran, D. K. Wilson; M. L. Van Horn (2018): "Identification of predicted individual treatment effects in randomized clinical trials". In: *Statistical methods in medical research* 27.1, pp. 142–157 (cited p. 64).

H. Lee; D. Nunan (2020): *Immortal time bias, Catalogue of Bias Collaboration*. URL: https://catalogofbias.org/biases/immortaltimebias/ (cited p. 51, 52).

M. Lee, C. Lee, C. Lai, T. Hsu, L. Porta, M. Lee, S. Chang, K. Chien, Y. Chen, N. T. U. H. H. Economics; O. R. Group (2017): "Preadmission statin use improves the outcome of less severe sepsis patients-a population-based propensity score matched cohort study". In: *BJA: British Journal of Anaesthesia* 119.4, pp. 645–654 (cited p. 121).

P. Lee, S. Bubeck; J. Petro (2023): "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine". In: *New England Journal of Medicine* 388.13, pp. 1233–1239 (cited p. 10).

R. Lelong, L. F. Soualmia, J. Grosjean, M. Taalba; S. J. Darmoni (2019): "Building a Semantic Health Data Warehouse in the Context of Clinical Trials: Development and Usability Study". In: *JMIR Medical Informatics* 7.4, e13917 (cited p. 23).

M.-C. Lenormand; D. Panteli (2021): *What can be learned from funding programmes that support the devolpment and testing of new care and payment models?* European Observatory on Health Systems and Policies. URL: https://sante.gouv.fr/IMG/pdf/cnam_obs_innov_funds_report_july2021_final.pdf (cited p. 82).

O. Levy; Y. Goldberg (2014): "Neural word embedding as implicit matrix factorization". In: *Advances in neural information processing systems* 27 (cited p. 103).

B. Li, H. Zhao, J. Zhang, Q. Yan, T. Li; L. Liu (2020a): "Resuscitation fluids in septic shock: a network meta-analysis of randomized controlled trials". In: *Shock* 53.6, pp. 679–685 (cited p. 55, 122).

Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi; G. Salimi-Khorshidi (2020b): "BEHRT: transformer for electronic health records". In: *Scientific reports* 10.1, pp. 1–12 (cited p. 10, 37, 41, 48, 95–97, 102).

J. Liang, Y. Li, Z. Zhang, D. Shen, J. Xu, X. Zheng, T. Wang, B. Tang, J. Lei; J. Zhang (2021): "Adoption of Electronic Health Records (EHRs) in China During the Past 10 Years: Consecutive Survey Data Analysis and Comparison of Sino-American Challenges and Experiences". In: *Journal of Medical Internet Research* 23.2, e24813 (cited p. 3).

Z. C. Lipton (2018): "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57 (cited p. 7, 9).

Z. C. Lipton, D. C. Kale, C. Elkan; R. Wetzel (2016): "Learning to diagnose with LSTM recurrent neural networks". In: *ICLR* (cited p. 37, 94).

T. Liu, Q. Zhao; B. 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 (cited p. 119, 121).

X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, et al. (2019): "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis". In: *The lancet digital health* 1.6, e271–e297 (cited p. 48).

N. Loiseau, P. Trichelair, M. He, M. Andreux, M. Zaslavskiy, G. Wainrib; M. G. B. Blum (2022): "External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning". In: *BMC Medical Research Methodology* 22 (cited p. 153).

V. Looten, L. Kong Win Chang, A. Neuraz, M.-A. Landau-Loriot, B. Vedie, J.-L. Paul, L. Mauge, N. Rivet, A. Bonifati, G. Chatellier, A. Burgun; B. Rance (2019): "What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse". In: *Computer Methods and Programs in Biomedicine* 181, p. 104825 (cited p. 32).

C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel; M. Welling (2017): "Causal Effect Inference with Deep Latent-Variable Models". In: *Advances in neural information processing systems* (cited p. 75).

M. Lu, S. Sadiq, D. J. Feaster; H. Ishwaran (2018): "Estimating individual treatment effect in observational data using random forest methods". In: *Journal of Computational and Graphical Statistics* 27.1, pp. 209–219 (cited p. 54).

M. F. MacDorman; J. O. Atkinson (1998): "Infant mortality statistics from the linked birth/infant death data set–1995 period data". In: *Monthly Vital Statistics Report* 46.6 Suppl 2 (cited p. 75).

D. Mackle, R. Bellomo, M. Bailey, R. Beasley, A. Deane, G. Eastwood, S. Finfer, R. Freebairn, V. King, N. Linke, et al. (2019): "Conservative oxygen therapy during mechanical ventilation in the ICU." In: *The New England journal of medicine* 382.11, pp. 989–998 (cited p. 120).

B. MacMahon, T. F. Pugh, et al. (1970): "Epidemiology: principles and methods." In: *Epidemiology: principles and methods.* (cited p. 10).

J. Madec, G. Bouzillé, C. Riou, P. Van Hille, C. Merour, M.-L. Artigny, D. Delamarre, V. Raimbert, P. Lemordant; M. Cuggia (2019): "eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network". In: *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pp. 1536–1537 (cited p. 28).

N. Malafaye, D. Demoulin, P. Mailhe, M. Morell, D. Pellecuer; C. Dunoyer (2018): "Mise en place et exploitation d'un entrepôt de données au département d'information médicale du CHU de Montpellier, France". In: *Revue d'Épidémiologie et de Santé Publique*. Colloque Adelf-Emois - Montpellier, 29 et 30 mars 2018 66, S26 (cited p. 23).

M. L. Malbrain, P. E. Marik, I. Witters, C. Cordemans, A. W. Kirkpatrick, D. J.

Roberts; N. Van Regenmortel (2014): "Fluid overload, de-resuscitation, and outcomes in critically ill or injured patients: a systematic review with suggestions for clinical practice". In: *Anaesthesiology intensive therapy* 46.5, pp. 361–380 (cited p. 121).

J. Mandel; P. M. Palevsky (2023): "Treatment of severe hypovolemia or hypovolemic shock in adults". In: *UpToDate* (cited p. 55).

J. M. McGinnis, L. Stuckhardt, R. Saunders, M. Smith, et al. (2013): "Best care at lower cost: the path to continuously learning health care in America". In: (cited p. 2, 9–11, 17).

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman; A. Galstyan (2021): "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6, 115:1–115:35 (cited p. 32).

T. S. Meyhoff, M. H. Moller, P. B. Hjortrup, M. Cronhjort, A. Perner; J. Wetterslev (2020): "Lower vs higher fluid volumes during initial management of sepsis: a systematic review with meta-analysis and trial sequential analysis". In: *Chest* 157.6, pp. 1478–1496 (cited p. 121).

Microsoft (2023): *Responsible AI toolbox*. URL: https://responsibleaitoolbox.ai (cited p. 60).

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado; J. Dean (2013): "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26 (cited p. 102).

M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran; M. Lucic (2021): "Revisiting the Calibration of Modern Neural Networks". In: *Advances in Neural Information Processing Systems* 34, pp. 15682–15694 (cited p. 80).

N. Mitra, J. Roy; D. Small (2022): "The Future of Causal Inference". In: *American Journal of Epidemiology* 191.10, pp. 1671–1676 (cited p. 60).

M. M. A. Monshi, J. Poon; V. Chung (2020): "Deep learning in generating radiology reports: A survey". In: *Artificial Intelligence in Medicine* 106, p. 101878 (cited p. 65).

S. Mooney, D. Westreich; A. El-Sayed (2015): "Epidemiology in the era of big data". In: *Epidemiology (Cambridge, Mass.)* 26.3, p. 390 (cited p. 9).

S. Mooney; V. Pejaver (2018): "Big data in public health: terminology, machine learning, and privacy". In: *Annual review of public health* 39, p. 95 (cited p. 64).

R. Moraffah, P. Sheth, M. Karami, A. Bhattacharya, Q. Wang, A. Tahir, A. Raglin; H. Liu (2021): "Causal inference for time series analysis: Problems, methods and evaluation". In: *Knowledge and Information Systems* 63, pp. 3041–3085 (cited p. 49).

L. Munshi, L. Del Sorbo, N. K. Adhikari, C. L. Hodgson, H. Wunsch, M. O. Meade, E. Uleryk, J. Mancebo, A. Pesenti, V. M. Ranieri, et al. (2017): "Prone position for acute respiratory distress syndrome. A systematic review and meta-analysis". In: *Annals of the American Thoracic Society* 14.Supplement 4, S280–S288 (cited p. 122).

A. I. Naimi, A. E. Mishler; E. H. Kennedy (2021): "Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms". In: *American Journal of Epidemiology* (cited p. 80, 153).

A. I. Naimi; B. W. Whitcomb (2023): "Defining and Identifying Average Treatment Effects". In: *American Journal of Epidemiology* (cited p. 52).

A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar; O. Elgendy (2022): "Breast cancer detection using artificial intelligence techniques: A systematic literature review". In: *Artificial Intelligence in Medicine*, p. 102276 (cited p. 64).

L. National Heart; B. I. A. C. T. Network (2004): "Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome". In: *New England Journal of Medicine* 351.4, pp. 327–336 (cited p. 121).

— (2014): "Rosuvastatin for sepsis-associated acute respiratory distress syndrome". In:

New England Journal of Medicine 370.23, pp. 2191–2200 (cited p. 121).

E. Neumayer; T. Plümper (2017): *Robustness tests for quantitative research*. Cambridge University Press (cited p. 54).

P. Nguyen, T. Tran, N. Wickramasinghe; S. Venkatesh (2016): "Deepr: a convolutional net for medical records". In: *IEEE journal of biomedical and health informatics* 21.1, pp. 22–30 (cited p. 94).

A. Niculescu-Mizil; R. Caruana (2005): "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd international conference on Machine learning - ICML '05*. ACM Press, pp. 625–632 (cited p. 80).

X. Nie; S. Wager (2017): "Quasi-Oracle Estimation of Heterogeneous Treatment Effects". In: *Biometrika* 108.2, pp. 299–319 (cited p. 65, 67, 69, 70, 75, 77, 80, 153).

— (2021): "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2, pp. 299–319 (cited p. 126).

NIH (2023): *2023 NIH Data Management and Sharing Policy*. URL: https://www.oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy (cited p. 31).

K. R. Niswander; U. S. N. I. o. N. D. a. Stroke (1972): *The Women and Their Pregnancies: The Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*. Google-Books-ID: A0bdVhlhDQkC. National Institute of Health. 562 pp. (cited p. 75).

OECD (2023): *OECD Health spending*. URL: https://data.oecd.org/healthres/health-spending.htm (cited p. 10).

OHDSI (2021): *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI. URL: https://ohdsi.github.io/TheBookOfOhdsi/ (cited p. 41, 51).

J. Oke, T. Fanshawe; D. Nunan (2021): *Lead time bias, Catalogue of Bias Collaboration*. URL: https://catalogofbias.org/biases/lead-time-bias/ (cited p. 51).

V. Omachonu, S. Suthummanon, M. Akcin; S. Asfour (2004): "Predicting length of stay for Medicare patients at a teaching hospital". In: *Health Services Management Research* 17.1, pp. 1–12 (cited p. 99).

OpenSAFELY (2022): *OpenSAFELY, Secure analytics platform for NHS electronic health records*. URL: https://www.opensafely.org/ (cited p. 22).

— (2023): *OpenSAFELY-TPP Database Schema*. URL: https://reports.opensafely.org/reports/opensafely-tpp-database-schema/ (cited p. 22).

C. Pang, X. Jiang, K. S. Kalluri, M. Spotnitz, R. Chen, A. Perotte; K. Natarajan (2021): "CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks". In: *Machine Learning for Health*. PMLR, pp. 239–260 (cited p. 41, 95–98, 102).

R. Panwar, M. Hardie, R. Bellomo, L. Barrot, G. M. Eastwood, P. J. Young, G. Capellier, P. W. Harrigan; M. Bailey (2016): "Conservative versus liberal oxygenation targets for mechanically ventilated patients. A pilot multicenter randomized controlled trial". In: *American journal of respiratory and critical care medicine* 193.1, pp. 43–51 (cited p. 120).

L. Papazian, J.-M. Forel, A. Gacouin, C. Penot-Ragon, G. Perrin, A. Loundou, S. Jaber, J.-M. Arnal, D. Perez, J.-M. Seghboyan, et al. (2010): "Neuromuscular blockers in early acute respiratory distress syndrome". In: *New England Journal of Medicine* 363.12, pp. 1107–1116 (cited p. 121, 122).

A. K. Parekh; M. B. Barton (2010): "The challenge of multiple comorbidity for the US health care system". In: *Jama* 303.13, pp. 1303–1304 (cited p. 11).

J. Pasco, B. Campillo-Gimenez, L. Guillon; M. Cuggia (2019): "Pré-screening et études de faisabilité : l'apport des entrepôts de données de cliniques". In: *Revue d'Épidémiologie et de Santé Publique* 67, S96 (cited p. 25).

C. J. Patel, B. Burford; J. P. Ioannidis (2015): "Assessment of vibration of effects due to

model specification can demonstrate the instability of observational associations". In: *Journal of clinical epidemiology* 68.9, pp. 1046–1058 (cited p. 54).

V. L. Patel, E. H. Shortliffe, M. Stefanelli, P. Szolovits, M. R. Berthold, R. Bellazzi; A. Abu-Hanna (2009): "The coming of age of artificial intelligence in medicine". In: *Artificial intelligence in medicine* 46.1, pp. 5–17 (cited p. 1, 4, 11, 16).

E. Pavlenko, D. Strech; H. Langhof (2020): "Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies". In: *BMC medical informatics and decision making* 20.1, p. 157 (cited p. 23, 31).

J. M. Pearce (2012): "The case for open source appropriate technology". In: *Environment, Development and Sustainability* 14, pp. 425–431 (cited p. 38).

J. Pearl; D. Mackenzie (2018): *The book of why: the new science of cause and effect*. Basic books (cited p. 12, 51).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot; É. Duchesnay (2011): "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830 (cited p. 57, 137, 144).

A. Perez-Lebel, M. L. Morvan; G. Varoquaux (2022): "Beyond calibration: estimating the grouping loss of modern neural networks". In: *arXiv preprint arXiv:2210.16315* (cited p. 80).

A. Perperoglou, W. Sauerbrei, M. Abrahamowicz; M. Schmid (2019): "A review of spline function procedures in R". In: *BMC Medical Research Methodology* 19.1, p. 46 (cited p. 73).

Pfizer (2019): *Collection and Use of Real-World Data Continues to Grow Around the World*. URL: https://www.pfizer.com/news/articles/collection_and_use_of_real_world_data_continues_to_grow_around_the_world (cited p. 9).

G. Plamondon, Y. Auclair, P. Dufort, S. Beha, C. Gonthier, M. Benigeri, Q. Nha-Hong, G. Boily, F. Kuzminski, I. Boisvert, E. Strumpf, J. Boulanger; M.-È. Tremblay (2022): *Intégration des données et des preuves du contexte réel dans les évaluations en appui à la prise de décision dans le secteur des médicaments*. Institut national d'excellence en santé et en services sociaux. URL: https://www.inesss.qc.ca/fileadmin/doc/INESSS/Rapports/Medicaments/INESSS_Donnees_preuves_contexte_reel_EC.pdf (cited p. 10).

J. C. Platt; J. C. Platt (1999): "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Advances in Large Margin Classifiers*, pp. 61–74 (cited p. 80).

D. Plecko; E. Bareinboim (2022): "Causal fairness analysis". In: *arXiv preprint arXiv:2207.11385* (cited p. 12, 48, 60).

PLOS Medicine Editors (2014): "Observational studies: getting clear about transparency". In: *PLoS medicine* 11.8, e1001711 (cited p. 31).

R. A. Poldrack, G. Huckins; G. Varoquaux (2020): "Establishment of best practices for evidence for prediction: a review". In: *JAMA psychiatry* 77.5, pp. 534–540 (cited p. 64, 65).

M. Porta (2014): *A dictionary of epidemiology*. Oxford university press (cited p. 3).

S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie; R. Tibshirani (2018): "Some methods for heterogeneous treatment effect estimation in high dimensions". In: *Statistics in Medicine* 37.11, pp. 1767–1787 (cited p. 65, 67).

M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan; J. Bian (2020): "Causal inference and counterfactual prediction in machine learning for actionable healthcare". In: *Nature Machine Intelligence* 2.7, pp. 369–375 (cited p. 48, 60).

PwC (2023): *Responsible AI Toolkit*. URL: https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-

intelligence/what-is-responsible-ai.html (cited p. 60).

H. I. Quick-Stat (2023): *Office-based Physician Electronic Health Record Adoption, Health IT Quick-Stat #50*. Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption (cited p. 4).

W. Raghupathi; V. Raghupathi (2014): "Big data analytics in healthcare: promise and potential." In: *Health information science and systems* (cited p. 36).

A. Rahimi; B. Recht (2008): "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Vol. 20 (cited p. 73, 143).

I. D. Raji, E. M. Bender, A. Paullada, E. Denton; A. Hanna (2021): "AI and the everything in the whole wide world benchmark". In: *35th Conference on Neural Information Processing Systems (NeurIPS 2021)* (cited p. 8).

A. Rajkomar, J. Dean; I. Kohane (2019): "Machine learning in medicine". In: *New England Journal of Medicine* 380.14, pp. 1347–1358 (cited p. 1, 16, 64).

A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado; M. H. Chin (2018a): "Ensuring fairness in machine learning to advance health equity". In: *Annals of internal medicine* 169.12, pp. 866–872 (cited p. 48).

A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. (2018b): "Scalable and accurate deep learning with electronic health records". In: *NPJ digital medicine* 1.1, p. 18 (cited p. 10, 48).

L. Rasmy, Y. Xiang, Z. Xie, C. Tao; D. Zhi (2021): "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction". In: *NPJ digital medicine* 4.1, p. 86 (cited p. 41, 95–97, 102).

A. Rekkas, D. van Klaveren, P. B. Ryan, E. W. Steyerberg, D. M. Kent; P. R. Rijnbeek (2023): "A standardized framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases". In: *npj Digital Medicine* 6.1, p. 58 (cited p. 48).

W. S. Richardson, M. C. Wilson, J. Nishikawa, R. S. Hayward, et al. (1995): "The well-built clinical question: a key to evidence-based decisions". In: *Acp j club* 123.3, A12–A13 (cited p. 50).

J. G. Richens, C. M. Lee; S. Johri (2020): "Improving the accuracy of medical diagnosis with causal machine learning". In: *Nature communications* 11.1, p. 3923 (cited p. 60).

R. L. Richesson, W. E. Hammond, M. Nahm, D. Wixted, G. E. Simon, J. G. Robinson, A. E. Bauck, D. Cifelli, M. M. Smerek, J. Dickerson, R. L. Laws, R. A. Madigan, S. A. Rusincovitch, C. Kluchar; R. M. Califf (2013): "Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory". In: *Journal of the American Medical Informatics Association* 20, e226–e231 (cited p. 10).

N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al. (2020): "The future of digital health with federated learning". In: *NPJ digital medicine* 3.1, p. 119 (cited p. 82).

J. J. Riva, K. M. Malik, S. J. Burnie, A. R. Endicott; J. W. Busse (2012): "What is your research question? An introduction to the PICOT format for clinicians". In: *The Journal of the Canadian Chiropractic Association* 56.3, p. 167 (cited p. 50).

S. E. Robertson, A. Leith, C. H. Schmid; I. J. Dahabreh (2021): "Assessing heterogeneity of treatment effects in observational studies". In: *American Journal of Epidemiology* 190.6, pp. 1088–1100 (cited p. 55).

J. Robins (1986): "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7.9, pp. 1393–1512 (cited p. 68).

J. M. Robins, A. Rotnitzky; L. P. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed". In: *Journal of the American statistical Association* 89.427, pp. 846–866 (cited p. 54, 126).

J. M. Robins; S. Greenland (1986): "The role of model selection in causal inference from non experimental data". In: *American Journal of Epidemiology* 123.3, pp. 392–402 (cited p. 54, 64, 123, 124).

P. M. Robinson (1988): "Root-N-Consistent Semiparametric Regression". In: *Econometrica* 4, pp. 931–954 (cited p. 69, 126, 138).

C. A. Rolling; Y. Yang (2014): "Model selection for estimating treatment effects". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.4, pp. 749–769 (cited p. 67).

E. Röösli, S. Bozkurt; T. Hernandez-Boussard (2022): "Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model". In: *Scientific Data* 9.1, p. 24 (cited p. 48).

S. Rose (2018): "Machine Learning for Prediction in Electronic Health Data". In: *JAMA Network Open* 1.4 (cited p. 37).

P. R. Rosenbaum; D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70, pp. 41–55 (cited p. 52, 64, 69, 124, 125).

K. J. Rothman (2012): *Epidemiology: an introduction*. Oxford university press (cited p. 5, 6).

K. Rothman, S. Greenland; T. Lash (2008): "Case-control studies, chapter 8". In: *Modern epidemiology*, pp. 111–127 (cited p. 80).

M. J. Rothman, S. I. Rothman; J. Beals IV (2013): "Development and validation of a continuous measure of patient condition using the electronic medical record". In: *Journal of biomedical informatics* 46.5, pp. 837–848 (cited p. 36, 93).

P. M. Rothwell (2005): "External validity of randomised controlled trials:"to whom do the results of this trial apply?"". In: *The Lancet* 365.9453, pp. 82–93 (cited p. 5).

— (2006): "Factors that can affect the external validity of randomised controlled trials". In: *PLoS clinical trials* 1.1, e9 (cited p. 60).

D. B. Rubin (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5, p. 688 (cited p. 12).

— (2005): "Causal Inference Using Potential Outcomes". In: *Journal of the American Statistical Association* 100.469, pp. 322–331 (cited p. 52, 68).

L. Rushton (2011): "Should protocols for observational research be registered?" In: *Occupational and environmental medicine* 68.2, pp. 84–86 (cited p. 31).

D. D. Rutstein (1967): "The coming revolution in medicine". In: (cited p. 10).

SAFE Study Investigators (2011): "Impact of albumin compared to saline on organ function and mortality of patients with severe sepsis". In: *Intensive care medicine* 37, pp. 86–96 (cited p. 60).

C. Safran, M. Bloomrosen, W. E. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang; D. E. Detmer (2007): "Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper". In: *Journal of the American Medical Informatics Association : JAMIA* 14.1, pp. 1–9 (cited p. 9, 10).

Y. Saito; S. Yasui (2020): "Counterfactual Cross-Validation: Stable Model Selection Procedure for Causal Inference Models". In: *International Conference on Machine Learning*. PMLR, pp. 8398–8407 (cited p. 67).

S. Salloum, R. Dautov, X. Chen, P. X. Peng; J. Z. Huang (2016): "Big data analytics on Apache Spark". In: *International Journal of Data Science and Analytics* 1.3, pp. 145–164 (cited p. 103).

A. J. Schaefer, M. D. Bailey, S. M. Shechter; M. S. Roberts (2004): "Modeling medical treatment using Markov decision pro-

cesses". In: *Operations research and health care: A handbook of methods and applications*, pp. 593–612 (cited p. 6).

S. Schneeweiss (2006): "Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics". In: *Pharmacoepidemiology and drug safety* 15.5, pp. 291–303 (cited p. 54).

S. Schneeweiss; E. Patorno (2021): "Conducting real-world evidence studies on the clinical outcomes of diabetes treatments". In: *Endocrine Reviews* 42.5, pp. 658–690 (cited p. 49).

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal; Y. Bengio (2021): "Toward causal representation learning". In: *Proceedings of the IEEE* 109.5, pp. 612–634 (cited p. 60).

N. J. Schork (2015): "Personalized medicine: time for one-person trials". In: *Nature* 520.7549, pp. 609–611 (cited p. 6, 10).

M. Schuemie (2021): *The Book of OHDSI*. URL: https://ohdsi.github.io/TheBookOfOhdsi/ (cited p. 24, 31, 32).

P. Schulam; S. Saria (2017): "Reliable Decision Support using Counterfactual Models". In: *Advances in neural information processing systems* 30 (cited p. 70).

A. Schuler, M. Baiocchi, R. Tibshirani; N. Shah (2018): "A comparison of methods for model selection when estimating individual treatment effects". In: *arXiv:1804.05146 [cs, stat]* (cited p. 67–70, 73, 153).

M. S. Schuler; S. Rose (2017): "Targeted maximum likelihood estimation for causal inference in observational studies". In: *American journal of epidemiology* 185.1, pp. 65–73 (cited p. 54, 65, 68, 76).

N. Schwalbe; B. Wahl (2020): "Artificial intelligence and the future of global health". In: *The Lancet* 395.10236, pp. 1579–1586 (cited p. 9).

W. B. Schwartz, R. S. Patil; P. Szolovits (1987): *Artificial intelligence in medicine* (cited p. 9, 10).

M. Schweinsberg, M. Feldman, N. Staub, O. R. van den Akker, R. C. van Aert, M. A. Van Assen, Y. Liu, T. Althoff, J. Heer, A. Kale, et al. (2021): "Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis". In: *Organizational Behavior and Human Decision Processes* 165, pp. 228–249 (cited p. 7).

K. P. Seastedt, P. Schwab, Z. O'Brien, E. Wakida, K. Herrera, P. G. F. Marcelo, L. Agha-Mir-Salim, X. B. Frigola, E. B. Ndulue, A. Marcelo; L. A. Celi (2022): "Global healthcare fairness: We should be sharing more, not less, data". In: *PLOS Digital Health* 1.10. Publisher: Public Library of Science (cited p. 32).

L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen; M. Ghassemi (2021): "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations". In: *Nature medicine* 27.12, pp. 2176–2182 (cited p. 48).

Z. Shahn, N. I. Shapiro, P. D. Tyler, D. Talmor; L.-w. H. Lehman (2020): "Fluid-limiting treatment strategies among sepsis patients in the ICU: a retrospective causal analysis". In: *Critical Care* 24.1, pp. 1–9 (cited p. 121).

U. Shalit, F. D. Johansson; D. Sontag (2017): "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085 (cited p. 70, 71, 137, 139).

U. Shalit; D. Sontag (2016): *Causal Inference for Observational studies: Tutorial*. URL: https://docplayer.net/64797211-Causal-inference-for-observational-studies.html (cited p. 49).

N. Shang, C. Weng; G. Hripcsak (2018): "A conceptual framework for evaluating data suitability for observational studies". In: *Journal of the American Medical Informatics Association: JAMIA* 25.3, pp. 248–258 (cited p. 32).

A. Sharma (2018): *Tutorial on causal inference and counterfactual reasoning*. URL:

https://causalinference.gitlab.io/kdd-tutorial/ (cited p. 49, 57).

A. Sheikh, A. Jha, K. Cresswell, F. Greaves; D. W. Bates (2014): "Adoption of electronic health records in UK hospitals: lessons from the USA". In: *Lancet (London, England)* 384.9937, pp. 8–9 (cited p. 3).

D. Shen, G. Wu; H.-I. Suk (2017): "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19, pp. 221–248 (cited p. 65).

L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride; W. Sieh (2019): "Deep learning to improve breast cancer detection on screening mammography". In: *Scientific reports* 9.1, p. 12495 (cited p. 64).

L. Shen, G. Geleijnse; M. Kaptein (2023): "RCTrep: An R Package for the Validation of Estimates of Average Treatment Effects". In: *Journal of Statistical Software* (cited p. 74).

B. Shickel, P. J. Tighe, A. Bihorac; P. Rashidi (2017): "Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis". In: *IEEE journal of biomedical and health informatics* 22.5, pp. 1589–1604 (cited p. 37, 94).

Y. Shimoni, E. Karavani, S. Ravid, P. Bak, T. H. Ng, S. H. Alford, D. Meade; Y. Goldschmidt (2019): "An Evaluation Toolkit to Guide Model Selection and Cohort Definition in Causal Inference". In: *arXiv preprint arXiv:1906.00442* (cited p. 153).

Y. Shimoni, C. Yanover, E. Karavani; Y. Goldschmnidt (2018): "Benchmarking Framework for Performance-Evaluation of Causal Inference Analysis". In: *arXiv:1802.05046 [cs, stat]* (cited p. 75).

E. H. Shortliffe (1993): "The adolescence of AI in medicine: will the field come of age in the'90s?" In: *Artificial intelligence in medicine* 5.2, pp. 93–106 (cited p. 4).

E. SHORTLIFFE (1976): "Computer Based Medical Consultations: MYCIN". In: *Elsevier* (cited p. 93).

G. E. Simon, E. Johnson, J. M. Lawrence, R. C. Rossom, B. Ahmedani, F. L. Lynch, A. Beck, B. Waitzfelder, R. Ziebell, R. B. Penfold, et al. (2018): "Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records". In: *American Journal of Psychiatry* 175.10, pp. 951–960 (cited p. 64).

H. Singh, V. Mhasawade; R. Chunara (2022): "Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database". In: *PLOS Digital Health* 1.4, e0000023 (cited p. 48).

R. K. Singh, V. Agarwal, A. K. Baronia, S. Kumar, B. Poddar; A. Azim (2017): "The effects of atorvastatin on inflammatory responses and mortality in septic shock: a single-center, randomized controlled trial". In: *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine* 21.10, p. 646 (cited p. 121).

S. T. Skou, F. S. Mair, M. Fortin, B. Guthrie, B. P. Nunes, J. J. Miranda, C. M. Boyd, S. Pati, S. Mtenga; S. M. Smith (2022): "Multimorbidity". In: *Nature Reviews Disease Primers* 8.1, p. 48 (cited p. 11).

J. M. Snowden, S. Rose; K. M. Mortimer (2011): "Implementation of G-computation on a simulated data set: demonstration of a causal inference technique". In: *American journal of epidemiology* 173.7, pp. 731–738 (cited p. 64, 68, 123).

O. Sofrygin, Z. Zhu, J. A. Schmittdiel, A. S. Adams, R. W. Grant, M. J. van der Laan; R. Neugebauer (2019): "Targeted learning with daily EHR data". In: *Statistics in medicine* 38.16, pp. 3073–3090 (cited p. 53).

J. R. A. Solares, Y. Zhu, A. Hassaine, S. Rao, Y. Li, M. Mamouei, D. Canoy, K. Rahimi; G. Salimi-Khorshidi (2021): "Transfer Learning in Electronic Health Records through Clinical Concept Embedding". In: *arXiv preprint arXiv:2107.12919* (cited p. 95).

I. Spasic, G. Nenadic, et al. (2020): "Clinical text data in machine learning: systematic review". In: *JMIR medical informatics* 8.3, e17984 (cited p. 64).

J. Splawa-Neyman, D. M. Dabrowska; T. P. Speed (1990): "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." In: *Statistical Science*, pp. 465–472 (cited p. 122).

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf; G. R. G. Lanckriet (2009): "On integral probability metrics, \phi-divergences and binary classification". In: *arXiv:0901.2698 [cs, math]* (cited p. 139).

F. Stéphan, B. Barrucand, P. Petit, S. Rézaiguia-Delclaux, A. Médard, B. Delannoy, B. Cosserant, G. Flicoteaux, A. Imbert, C. Pilorge, et al. (2015): "High-flow nasal oxygen vs noninvasive positive airway pressure in hypoxemic patients after cardiothoracic surgery: a randomized clinical trial". In: *Jama* 313.23, pp. 2331–2339 (cited p. 119, 122).

R. A. Stewart, P. Jones, B. Dicker, Y. Jiang, T. Smith, A. Swain, A. Kerr, T. Scott, D. Smyth, A. Ranchord, et al. (2021): "High flow oxygen and risk of mortality in patients with a suspected acute coronary syndrome: pragmatic, cluster randomised, crossover trial". In: *bmj* 372 (cited p. 122).

E. W. Steyerberg (2009): *Applications of prediction models*. Springer (cited p. 11, 93).

M. Stone (1974): "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the royal statistical society: Series B (Methodological)* 36.2, pp. 111–133 (cited p. 7).

E. A. Stuart (2010): "Matching methods for causal inference: A review and a look forward". In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 25.1, p. 1 (cited p. 124).

X. Su, A. T. Peña, L. Liu; R. A. Levine (2018): "Random forests of interaction trees for estimating individualized treatment effects in randomized trials". In: *Statistics in medicine* 37.17, pp. 2547–2560 (cited p. 64).

A. Subbaswamy; S. Saria (2020): "From development to deployment: dataset shift, causality, and shift-stable models in health AI". In: *Biostatistics* 21.2, pp. 345–352 (cited p. 12).

S. Suissa (2008): "Immortal time bias in pharmacoepidemiology". In: *American journal of epidemiology* 167.4, pp. 492–499 (cited p. 49, 51).

A. Swaminathan; T. Joachims (2015): "Counterfactual risk minimization: Learning from logged bandit feedback". In: *International Conference on Machine Learning*. PMLR, pp. 814–823 (cited p. 71, 76).

P. Szolovits (1982): *Artificial Intelligence and Medicine*. Westview Press (cited p. 36, 93).

E. W. Tang, C.-K. Wong; P. Herbison (2007): "Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long-term mortality post acute coronary syndrome". In: *American heart journal* 153.1, pp. 29–35 (cited p. 36, 93).

G. Teasdale; B. Jennett (1974): "Assessment of coma and impaired consciousness: a practical scale". In: *The Lancet* 304.7872, pp. 81–84 (cited p. 38).

J. Textor, J. Hardt; S. Knüppel (2011): "DAGitty: a graphical tool for analyzing causal diagrams". In: *Epidemiology* 22.5, p. 745. URL: http://dagitty.net/ (cited p. 53).

L. Thabane, L. Mbuagbaw, S. Zhang, Z. Samaan, M. Marcucci, C. Ye, M. Thabane, L. Giangregorio, B. Dennis, D. Kosa, et al. (2013): "A tutorial on sensitivity analyses in clinical trials: the what, why, when and how". In: *BMC medical research methodology* 13.1, pp. 1–12 (cited p. 54).

N. Tomašev, N. Harris, S. Baur, A. Mottram, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, V. Magliulo, et al. (2021): "Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records". In: *Nature Protocols* 16.6, pp. 2765–2787 (cited p. 41).

E. J. Topol (2019): "High-performance medicine: the convergence of human and

artificial intelligence". In: *Nature medicine* 25.1, pp. 44–56 (cited p. 1, 6, 10, 16, 36, 93, 94).

T. Tran, T. D. Nguyen, D. Phung; S. Venkatesh (2015): "Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)". In: *Journal of biomedical informatics* 54, pp. 96–105 (cited p. 94).

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* 62.3, pp. 219–223 (cited p. 6, 48, 60).

C.-H. Tseng, T.-T. Chen, M.-Y. Wu, M.-C. Chan, M.-C. Shih; Y.-K. Tu (2020): "Resuscitation fluid types in sepsis, surgical, and trauma patients: a systematic review and sequential network meta-analyses". In: *Critical Care* 24.1, pp. 1–12 (cited p. 122).

P. Tuppin, J. Rudant, P. Constantinou, C. Gastaldi-Ménager, A. Rachas, L. de Roquefeuil, G. Maura, H. Caillol, A. Tajahmady, J. Coste, C. Gissot, A. Weill; A. Fagot-Campagna (2017): "Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France". In: *Revue d'Épidémiologie et de Santé Publique*. Réseau REDSIAM 65, S149–S167 (cited p. 10, 22).

M. J. Van der Laan, E. C. Polley; A. E. Hubbard (2007): "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (cited p. 54).

L. Van der Maaten; G. Hinton (2008): "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (cited p. 103).

T. J. VanderWeele (2019): "Principles of confounder selection". In: *European journal of epidemiology* 34, pp. 211–219 (cited p. 53, 64).

A. Vanier, J. Fernandez, S. Kelley, L. Alter, P. Semenzato, C. Alberti, S. Chevret, D. Costagliola, M. Cucherat, B. Falissard, et al. (2023): "Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health". In: *BMJ Evidence-Based Medicine* (cited p. 10).

V. Vapnik (1999): *The nature of statistical learning theory*. Springer science & business media (cited p. 7).

G. Varoquaux; O. Colliot (2022): *Evaluating machine learning models and their diagnostic value* (cited p. 38, 64, 65).

G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz; B. Thirion (2017): "Assessing and tuning brain decoders: cross-validation, caveats, and guidelines". In: *NeuroImage* 145, pp. 166–179 (cited p. 7, 8).

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser; I. Polosukhin (2017): "Attention is all you need". In: *Advances in neural information processing systems* 30 (cited p. 41, 95).

V. Veitch, A. D'Amour; K. P. Murphy (2022): "Probabilistic machine learning: Advanced topics". In: MIT press. Chap. 36: Causality (cited p. 12).

I. W. M. Verburg, A. Atashi, S. Eslami, R. Holman, A. Abu-Hanna, E. de Jonge, N. Peek; N. F. de Keizer (2017): "Which models can I use to predict adult ICU length of stay? A systematic review". In: *Critical care medicine* 45.2, e222–e231 (cited p. 94).

M. Wack (2017): "Installation d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy : déploiement technique, intégration et gouvernance des données". Université de Lorraine. URL: https://hal.univ-lorraine.fr/hal-01931928 (cited p. 23).

S. Wager (2020a): "Cross-validation, risk estimation, and model selection: Comment on a paper by Rosset and Tibshirani". In: *Journal of the American Statistical Association* 115.529, pp. 157–160 (cited p. 7).

S. Wager (2020b): *Stats 361: Causal inference* (cited p. 54, 126, 127).

S. Wager; S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242 (cited p. 65, 69, 70).

Y.-D. Wan, T.-W. Sun, Q.-C. Kan, F.-X. Guan; S.-G. Zhang (2014): "Effect of statin therapy on mortality from infection and sepsis: a meta-analysis of randomized and observational studies". In: *Critical care* 18, pp. 1–13 (cited p. 121).

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy; S. R. Bowman (2018): "GLUE: A multi-task benchmark and analysis platform for natural language understanding". In: *arXiv preprint arXiv:1804.07461* (cited p. 8).

H. Wang, Y. Li, S. A. Khan; Y. Luo (2020): "Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network". In: *Artificial intelligence in medicine* 110, p. 101977 (cited p. 64).

J. Wang, L. Jiang, S. Ding, S.-Y. He, S.-B. Liu, Z.-J. Lu, Y.-Z. Liu, L.-W. Hou, B.-S. Wang; J.-B. Zhang (2023a): "Early Enteral Nutrition and Sepsis-Associated Acute Kidney Injury: A Propensity Score Matched Cohort Study Based on the MIMIC-III Database". In: *Yonsei Medical Journal* 64.4, pp. 259–268 (cited p. 121).

S. V. Wang, S. Schneeweiss, J. M. Franklin, R. J. Desai, W. Feldman, E. M. Garry, R. J. Glynn, K. J. Lin, J. Paik, E. Patorno, et al. (2023b): "Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials". In: *JAMA* 329.16, pp. 1376–1385 (cited p. 10, 49, 50, 60).

S. V. Wang, S. K. Sreedhara, L. G. Bessette; S. Schneeweiss (2022): "Understanding variation in the results of real-world evidence studies that seem to address the same question". In: *Journal of Clinical Epidemiology* 151, pp. 161–170 (cited p. 51).

N. G. Weiskopf, D. A. Dorr, C. Jackson, H. P. Lehmann; C. A. Thompson (2023): "Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse". In: *Journal of the American Medical Informatics Association*, ocad013 (cited p. 51).

T. Wendling, K. Jung, A. Callahan, A. Schuler, N. H. Shah; B. Gallego (2018a): "Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases". In: *Statistics in Medicine* 23, pp. 3309–3324 (cited p. 64).

T. Wendling, K. Jung, A. Callahan, A. Schuler, N. H. Shah; B. Gallego (2018b): "Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases". In: *Statistics in medicine* 37.23, pp. 3309–3324 (cited p. 54, 123).

K. White, T. Williams; B. Greenberg (1961): "The ecology of medical care". In: *The New England Journal of Medicine* 265, pp. 885–892 (cited p. 82).

B. Wieseler, M. Neyt, T. Kaiser, F. Hulstaert; J. Windeler (2023): "Replacing RCTs with real world data for regulatory decision making: a self-fulfilling prophecy?" In: *bmj* 380 (cited p. 10).

J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. 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* 155.10, pp. 1135–1141 (cited p. 48).

M. F. Wisniewski, P. Kieszkowski, B. M. Zagorski, W. E. Trick, M. Sommers, R. A. Weinstein; for the Chicago Antimicrobial Resistance Project (2003): "Development of a Clinical Data Warehouse for Hospital Infection Control". In: *Journal of the American Medical Informatics Association* 10, pp. 454–462 (cited p. 10).

A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J.

Pestrue, M. Phillips, J. Konye, C. Penoza, et al. (2021): "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients". In: *JAMA Internal Medicine* 181.8, pp. 1065–1070 (cited p. 36, 37, 93).

M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries; N. H. Shah (2023): "The shaky foundations of large language models and foundation models for electronic health records". In: *npj Digital Medicine* 6.1, p. 135 (cited p. 36, 38, 95).

J. Wu, J. Roy; W. F. Stewart (2010): "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches". In: *Medical care*, S106–S113 (cited p. 11, 94).

J. C. Wyatt; D. G. Altman (1995): "Commentary: Prognostic models: clinically useful or quickly forgotten?" In: *Bmj* 311.7019, pp. 1539–1541 (cited p. 38).

Y. Xiang, J. Xu, Y. Si, Z. Li, L. Rasmy, Y. Zhou, F. Tiryaki, F. Li, Y. Zhang, Y. Wu, et al. (2019): "Time-sensitive clinical concept embeddings learned from large electronic health records". In: *BMC medical informatics and decision making* 19, pp. 139–148 (cited p. 103).

J.-Y. Xu, Q.-H. Chen, J.-F. Xie, C. Pan, S.-Q. Liu, L.-W. Huang, C.-S. Yang, L. Liu, Y.-Z. Huang, F.-M. Guo, et al. (2014): "Comparison of the effects of albumin and crystalloid on mortality in adult patients with severe sepsis and septic shock: a meta-analysis of randomized clinical trials". In: *Critical Care* 18.6, pp. 1–8 (cited p. 60, 121).

A. Yala, C. Lehman, T. Schuster, T. Portnoi; R. Barzilay (2019): "A deep learning mammography-based model for improved breast cancer risk prediction". In: *Radiology* 292.1, pp. 60–66 (cited p. 64).

R. Yamamoto, I. Nahara, M. Toyosaki, T. Fukuda, Y. Masuda; S. Fujishima (2020): "Hydrocortisone with fludrocortisone for septic shock: a systematic review and meta-analysis". In: *Acute medicine & surgery* 7.1, e563 (cited p. 122).

X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, et al. (2022): "Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records". In: *arXiv preprint arXiv:2203.03540* (cited p. 97).

C. J. Yarnell, F. Angriman, B. L. Ferreyro, K. Liu, H. J. De Grooth, L. Burry, L. Munshi, S. Mehta, L. Celi, P. Elbers, et al. (2023): "Oxygenation thresholds for invasive ventilation in hypoxemic respiratory failure: a target trial emulation in two cohorts". In: *Critical Care* 27.1, pp. 1–13 (cited p. 119).

K.-H. Yu, A. Beam; I. Kohane (2018): "Artificial intelligence in healthcare". In: *Nature biomedical engineering* 2.10, pp. 719–731 (cited p. 4, 9, 93).

B. Zadrozny; C. Elkan (2001): "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers". In: p. 8 (cited p. 80).

J. Zeng, M. F. Gensheimer, D. L. Rubin, S. Athey; R. D. Shachter (2022): "Uncovering interpretable potential confounders in electronic medical records". In: *Nature Communications* 13.1, p. 1014 (cited p. 49, 54).

J. Zhang, J. Symons, P. Agapow, J. T. Teo, C. A. Paxton, J. Abdi, H. Mattie, C. Davie, A. Z. Torres, A. Folarin, H. Sood, L. A. Celi, J. Halamka, S. Eapen; S. Budhdeo (2022): "Best practices in the real-world data life cycle". In: *PLOS Digital Health* 1 (cited p. 31, 32).

Z. Zhang; E. Sejdić (2019): "Radiological images and machine learning: trends, perspectives, and prospects". In: *Computers in biology and medicine* 108, pp. 354–370 (cited p. 64).

S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert; R. M. Summers (2021a): "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises".

In: *Proceedings of the IEEE* 109.5, pp. 820–838 (cited p. 10).

S. Zhou, Z. Zeng, H. Wei, T. Sha; S. An (2021b): "Early combination of albumin with crystalloids administration might be beneficial for the survival of septic patients: a retrospective analysis from MIMIC-IV database". In: *Annals of intensive care* 11, pp. 1–10 (cited p. 55, 56, 59, 121, 134).

J. Ziegler, B. N. Rush, E. R. Gottlieb, L. A. Celi; M. Á. Armengol de la Hoz (2022): "High resolution data modifies intensive care unit dialysis outcome predictions as compared with low resolution administrative data set". In: *PLOS Digital Health* (cited p. 4).

G. A. Zielhuis; L. A. Kiemeney (2001): "Social epidemiology? No way". In: *International journal of epidemiology* 30.1, pp. 43–44 (cited p. 6).

P. N. Zivich; A. Breskin (2021): "Machine learning for causal inference: on the use of cross-fit estimators". In: *Epidemiology (Cambridge, Mass.)* 32.3, p. 393 (cited p. 80).

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Représentations et inférence à partir de données de santé temporelles collectées en routine
Mots clés: Apprentissage automatique, Dossiers de santé électroniques, Causalité, Epidémiologie, Santé publique

Résumé: Les bases de données de vie réelle sont de plus en plus accessibles, exhaustives, avec des détails temporels précis. Contrairement aux données utilisées dans la recherche clinique traditionnelle, elles capturent l'organisation routinière des soins. Ces données de soins quotidiens ouvrent la porte à de nouvelles questions de recherche, notamment en ce qui concerne la qualité des soins, l'efficacité des interventions après leur mise sur le marché, l'hétérogénéité de leurs bénéfices dans les populations mal desservies ou le développement de traitements personnalisés. D'un autre côté, la complexité et la nature à grande échelle de ces bases de données posent un certain nombre de défis pour une utilisation efficace. Pour remédier à ces problèmes, les économètres et les épidémiologistes ont récemment proposé l'utilisation de modèles flexibles combinant l'inférence causale et l'apprentissage automatique en grande dimension.

Dans un premier temps, nous illustrons par trois exemples la tension actuelle entre ces nouvelles sources de données, l'apprentissage automatique et des problématiques modernes de santé publique. Ces exemples motivent notre principale question de recherche : Comment des modèles flexibles peuvent-ils aider à fournir un traitement approprié à chaque patient afin d'améliorer sa santé ? Afin de mieux comprendre les infrastructures modernes de collecte et d'analyse des dossiers patients informatisés (DPI), nous faisons la synthèse d'entretiens semi-structurés menés dans le cadre d'une étude de cas nationale portant sur les entrepôts de données cliniques des 32 hôpitaux régionaux et universitaires français. Reconnaissant la difficulté d'accéder à des échantillons de grande taille et à la puissance de calcul pour développer des modèles prédictifs généralisables, nous étudions un gradient de complexité dans les représentations et les algorithmes prédictifs sur DPI. En se tournant vers le cadre causal, nous détaillons ensuite les éléments clés nécessaires pour estimer de manière robuste l'effet du traitement à partir de données de DPI variant dans le temps. Nous documentons l'impact de différents choix méthodologiques pour l'étude de l'effet de l'albumine sur la mortalité dans des cas de septicémie avec la base de données MIMIC-IV (Medical Information Mart for Intensive Care). Les DPIs sont des bases de données à grandes dimensions. Pour de tels problèmes, la sélection d'hyperparamètres pour les modèles causaux est cruciale pour éviter le sous-apprentissage ou le surapprentissage. Pour une simulation et trois ensembles de données semi-simulées, nous montrons que le risque usuel en apprentissage statistique n'est pas adapté au cadre causal et que le risque R doublement robuste surpasse d'autres risques causaux existants.

Title: Representations and inference from time-varying routine care data
Keywords: Machine learning, Electronic Health Records, Causality, Epidemiology, Public health

Abstract: Real World Databases are increasingly accessible, exhaustive and with fine temporal details. Unlike traditional data used in clinical research, they capture the routine organization of care. These day-to-day records of patients care open the door to new research questions, notably concerning the efficiency of interventions after market access, the heterogeneity of their benefits in under-served populations or the development of personalized medicine. On the other hand, the complexity and large-scale nature of these databases pose a number of challenges for effectively answering these questions. To remedy these problems, econometricians and epidemiologists have recently proposed the use of flexible models combining causal inference with high-dimensional machine learning.

We first illustrate with three examples the current tension between these new sources of data, machine learning and modern public health issues. These examples motivate the main research question of this work: How flexible models can help delivering appropriate treatment to each and every patient to improve her health? In order to gain a better understanding of the modern infrastructures for collecting and analyzing Electronic Health Records (EHRs), we summarize semi-structured interviews conducted as part of a national case study of the clinical data warehouses (CDWs) of the 32 French regional and university hospitals. Acknowledging the difficulty to access large sample sizes and computational power to develop generalizable predictive models, we explore a complexity gradient in representation and predictive algorithms for EHRs. We then turn to causal thinking, detailing key elements necessary to robustly estimate treatment effect from time-varying EHR data. We illustrate the impact of methodological choices in studying the effect of albumin on sepsis mortality in the Medical Information Mart for Intensive Care database (MIMIC-IV). EHRs are high-dimensional databases. For such settings, the selection of hyper-parameters for the causal model is crucial to avoid under- or over-learning. In a simulation and three semi-simulated datasets, we show that the usual machine learning risk are not adapted to the causal setting and that the doubly robust R-risk outperforms other existing causal risks.

Maison du doctorat de Université Paris-Saclay
2e étage, aile ouest, École normale supérieure Paris-Saclay
4 avenue des Sciencs
91190 Gif-sur-Yvette, France