Thèse de doctorat

NNT: 2020UPASA000

université PARIS-SACLAY

Université partenaire
cotutelle uniquement

Titre de la thèse en français (sur plusieurs lignes si nécessaire)

Representations and inference from time-varying routine care data

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 000, dénomination et sigle
Spécialité de doctorat: voir annexe
Unité de recherche: voir annexe
Référent: : voir annexe

Thèse présentée et soutenue à ....., le .... 202X, par

Prénom NOM

Composition du jury

Prénom Nom Titre, Affiliation	Président/e
Prénom Nom Titre, Affiliation	Rapportrice
Prénom Nom Titre, Affiliation	Rapporteur
Prénom Nom Titre, Affiliation	Examinatrice
Prénom Nom Titre, Affiliation	Examinateur
Prénom Nom Titre, Affiliationt	Examinateur

Direction de la thèse

Prénom Nom Titre, Affiliation	Directrice
Prénom Nom Titre, Affiliation	Codirecteur
Prénom Nom Titre, Affiliation	Coencadrante
Prénom Nom Titre, Affiliation	tuteur en entreprise

Representations and inference from time-varying routine care data

Matthieu Doutreligne

Real World Databases are increasingly accessible, exhaustive and with fine temporal details. Unlike traditional data used in clinical research, they describe the routine organization of care. These day-to-day records of patients care enable 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.

Chapter 1 uses a national case study of the 32 French regional and university hospitals to highlight key aspects of modern Clinical Data Warehouses (CDWs), the cornerstone infrastructure of precision medicine. From semi-structured interviews and an analysis of declared observational studies on CDWs in France, I highlight both the current potential and challenges of leveraging these routinely collected data for research purposes.

Contents

Chapter 1

Introduction: amazing opportunities of health data?

1.1 How I came into this landscape

I will present my progression among these opportunities.

1.1.1 Context: From a modern statistical formation to the first contact with epidemiology

Learning statistics during the Natural Language Processing revolution

Machine learning techniques, a subfield of artificial intelligence, became increasingly good as solving complex tasks where previous pattern recognition techniques were struggling. First, categorized into supervised and unsupervised approaches, the pre-training paradigm became increasingly popular in the 2010's: Halevy et al., 2009 proposed to take advantages of regularities present in large piles of data to design automatically interesting features for large application domains. The practical successes came in image processing (), then Natural Language Processing () and structural biology.

APHP: Removing identification elements from clinical notes

Continuous improvements in natural language processing led to the intuition that soon the information contained in clinical texts could be mined by appropriate to serve clinical research.

However, privacy concern: the access to very detailed elements of patient cares is no more restricted to the medical staff.

Need for pseudonymization : use newly to reach State of the Art pseudonymization (Dernoncourt et al., 2017; Paris et al., 2019).

Billing claims: an underused huge pile of data

French ministry department of statistics had billions of claims extracted in a dedicated server. But the amount of data made the job impossible.

I learned that distributed computing (aka appropriate tools), summary statistics and good documentation.

I began to wonder what were the appropriate methodological tools to extract relevant information from this data. Hammer that sees a nail : Can I use language modeling techniques to mine this data ?

First questions: how useful are such wealth of data ? Is unsupervised learning useful ? Is prediction a useful goal ? I was exactly in the atypical situation mentioned by Cox in his answer to Breiman's foundational paper on Machine Learning (Breiman, 2001) : *Data looking for a question are not unknown and raise puzzles but are, I believe, atypical in most contexts.*

Ironically, at the end of the day, I replicated unsupervised learning results from (Beam et al., 2019) but would not know how to use it. I moved towards more interpretable dictionary learning with control groups analyses to study polymedication.

Covid-19: National healthcare monitoring during the epidemics. Integrated data pipeline are very useful during crisis for monitoring: If you do not see all the data creating process, standard pipelines, standard tests and documentation : it is inherently a group effort.

1.1.2 Wrap-up: Increasing data collection and computing power

Data is collected at massive scale in healthcare

centralized into dedicated databases for analysis: eg. APHP clinical notes, French.

It is acknowledged that we entered since at least 2022 years in an era where we frenetically store data without having a dedicated question in mind. The vague concept of big data describes this new attitude towards data collection. I would propose another view on the current situation: data collected in huge quantity is part of the reality, but it does not preclude to assemble data suited for an appropriate question. In medical informatics, this is called computational phenotyping.

Models are scaling up

A whole generation of quantitative researchers has been trained with flexible data-hungry models. The unreasonable effectiveness of data (Halevy et al., 2009).

A recent work on the classification of data and models (AI) in clinical medicine suggests three main usages: AI in clinical practice, clinical research on AI/ML device and applications and AI/ML used to conduct clinical research (Haug; Drazen, 2023).

1.1.3 What pressing needs to use health data

Shortliffe, the inventer of MYCINE (Shortliffe, 1974), the first rule-based artificial intelligence expert systems described *medical practice, and biomedical research, [as] inherently information-management tasks* (Patel et al., 2009).

However, the incentives to collect and analyze data are not aligned between healthcare actors.

Healthcare practice

Automation of tedious tasks, decision support system

Research

Drug development, evidence based medicine, biomarkers, epidemiology ie. understanding of the mechanisms of disease *the study of the distribution and determinants of disease frequency* (MacMahon, Pugh, et al., 1970)

Public policies: monitoring and evaluation

Guidelines, quality of care, public health

Health Technology Agencies interest: organizational impact of health products, potential for real world efficiency (entire life cycle of the product), quality, security, pertinence and security of care,

Government bodies: health monitoring, health management, contextualization.

Marketing

Market access, tailored recommendations.

1.2 The data: Electronic Health Records

1.2.1 Various types of data: Real World Data

Real world data refer to routinely collected data

Gradient and distinction between research data and opportunistic data collection.

Traditional research data collections

Hand made data collection for one research question with a specific protocol, Specialized registry and cohorts

Claims

Billing system, eg. PMSI.

Advantages (space and time coverage, scale, structure), disadvantages (not clinic, no exam results, few measures, heterogeneity of collection)

EHRs and Hospital Information System

Increasing informatization EHR at the center Other applications part of HIS: give examples.

1.2.2 Interventional data vs observational data

Interventional data contains interaction with the patients / environment where the intervention probabilities are known. Eg. RCT : fixed probability for any patient (statistical unit): first RCT, RCTs as the basis of Evidence Based Medicine in the late 80s.

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 (K. J. Rothman, 2012)

Pbs of external validity raised in economics (Deaton, cf intro de la revue de Bénédicte). Focus is such problems is probably greater in economics because situations are not well controlled at all: very far from labs settings. Clinical situation is closer to lab (clinical epidemiology != social epidemiology). What about medico-economics ? I do not address this question in the present thesis, but it is a clear motivation.

Observational data cannot intervene in any way with the patients. It should rely on observation alone to estimate intervention probability.

Conditional probabilities are interventional as well: conditionally randomized experiments in epidemiology (Hernàn; Robins, 2020). This concept is known as reinforcement learning in the Machine Learning community (bareinboim2015bandits) where intervention probabilities are known Bareinboim et al., 2015. Decision making processes rely to the correct estimations of these conditional probabilities, hence the link between this work and precision medicine.

1.2.3 Focus of this thesis: real world data, observational

Inconvenient of interventional data : Increasing costs of RCTs,

Today's public health issues are closely linked to routine care, not ideal care: resource constraint, chronic disease. Not so new as it was already major issues mentioned in the late 60s Rutstein, 1967: *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 MD's; 4) increasing demands on the physician's time as increasing numbers of individuals began to demand quality medical care.*

Rise of digitalization and computing power suggesting new research methods

1.3 Two cultures of statistics for health

In 2001, Breiman, 2001 clearly separates two statistical cultures: a predominant community at the time focused on models, another one relying on predictive accuracy.

Epidemiological statistics tends to be closer to the first model-based culture whereas artificial intelligence in medicine embraced the second one. This cultural differences might be explained by different objectives. Clinical Epidemiology seeks to uncover nature mechanisms from data, to understand how nature works. AIM adopts a more pragmatic approach, trying to understand what deployed healthcare system is improving outcomes or not, without trying to uncover the full causal mechanism leading to observed results.

1.3.1 Machine Learning Framework

Departing from carefully designing functional forms for statistical models, seeking algorithm efficiency and out-of sample predictive accuracy.

Rashomon effect: multiple models can equally well model a given dataset

Model selection

Hyper-parameters selection, cross-validation (stone, 1974) and Textbook figure

1.3.2 Biostatistics Framework

In medical journals, Cox model for survival data and logistic regression for binary outcomes have been the standard for publication.

Focus on variable selection: root in hand collected variables in carefully designed subpopulation ?

Focus on simple models for juridical and credibility reasons but also for practical reasons: the input data should be readily available to question one model individual prediction (Wyatt; Altman, 1995).

The divide between modeling and empirical algorithmic approaches is best illustrated in the commentary to Breimann by David Cox or Bad Efron in:

Often the prediction is under quite different conditions from the data; [...] what would be the effect on annual incidence of cancer in the United States of reducing by 10% the medical use of X-rays. This example is a pure causal inference question.

1.3.3 One choice of perspective: Recent statistical learning

This perspective reflects more naturally my progression:

Learning statistics in the age of deep learning pushes towards adopting a view on statistics that consider mechanisms as generally unknown, and predictive accuracy as the hallmark of validity (Breiman, 2001).

Recent discourse and successes are heavily influenced by this line of though:

Machines read medical images faster and more efficiently than most practitioners (Zhou et al., 2021). Structured data from Electronic Health Records (Rajkomar et al., 2018) 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., 2020). Recently, large language models (LLMs)

leveraged clinical notes from several hospitals for length of stay prediction (Jiang et al., 2023). Hope is high that LLMs models will soon be able to help practitioners during consultation (Lee et al., 2023).

1.4 Notions of causality

1.4.1 Association is not causation

Observational data can have different causal interpretations

One statistical model yields multiple causal models, only one of which correctly reflects the reality : find a good medical example (otherwise look into Peter jonas's book or K. P. Murphy, 2022, chapter 36).

Causality in epidemiology

Opening an epidemiological book, confounding is the very first explained concept (K. J. Rothman, 2012).

Causes are mainly presented as binary variables absent or present in (). Their strengh is only defined at a population level, making them relative to a given population (K. J. Rothman, 2012, chapter 3). This point of view is rather aligned with statistical modeling related to variable significance, not with model performances.

Interesting concept of promoter –the last component cause– that

Importance of validity by generalization of theory rather than statistical generalization: established epidemiological knowledge should *tell us what to expect in people or settings that were not studied.* Statistical generalization is considered by K. J. Rothman, 2012 as applied epidemiology, focused on specific cases and not as the core of epidemiological science.

Causality in Machine Learning

Relation to dataset shift (Subbaswamy; Saria, 2020)

Distinction between association and causation (ladder of causation): first insight that different tools than statistics are needed.

1.4.2 Neyman-rubin causal framework

I quickly introduce the statistical framework for causality, and will progressively introduce supplementary concepts as needed in Chapters 4 and ??.

Background Information

Background Information

The Neyman-Rubin Potential Outcomes framework

(Naimi; Whitcomb, 2023; Imbens; Rubin, 2015) enables statistical reasoning on causal treatment effects: Given an outcome $Y \in \mathbb{R}$ (eg. mortality risk or hospitalization length), function of a binary treatment $A \in \mathcal{A} = \{0, 1\}$ (eg. 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 (unobserved 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

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

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

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

Randomized Control Trials, the gold standard for evidence based medicine

- Send back to appendix the Difference in Means (in randomized experiments)

1.5 Overview and contributions

Chapter 2

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

2.1 Abstract

This chapter has been published in Plos Digital Health as:
Doutreligne, M., Degremont, A., Jachiet, P. A., Lamer, A., & Tannier, X. (2023). Good practices for clinical data warehouse implementation: A case study in France. PLOS Digital Health, 2(7), e0000298.
As the first author I formulated the research question, designed the study, led the interviews and wrote the manuscript.

2.2 Motivation and background: A changing world

2.2.1 Data for primary or secondary usages?

Primary data usages

Primary usages directly serve one patient care.

Secondary data usages

Secondary usages do not concern directly the care and support of one patient: research, quality or management indicators, billings.

Towards mix usages: the learning health system paradigm

The notion of learning health system (McGinnis et al., 2013).
Mix usages such as learning algorithms

2.2.2 Healthcare data collection is heavily influenced from the local healthcare organization

Centralized data collections: ex. Israël, HIRA

More heterogeneous data collections structured into networks

ex. from
US, Korea, Germany,

The case of France

centralized national insurer but scattered hospital ehrs. Projects in other areas: eg. cancer (unicancer), gp (cnge project, darmoni). National initiative to develop and structure hospital cdws.

2.2.3 The multiplication of Clinical Data Warehouses

An infrastructure is needed to pool data from one or more medical information systems. Definition of Clinical Data Warehouse: Health data warehouses (HDWs) refer to the pooling of data from one or more medical information systems, in a homogeneous format for reuse in management, research or care.

The 4 phases of data flow from the various sources that make up the HIS (figure) Focus of this thesis: Clinical Data Warehouse and EHR, though most of my work should apply to claims

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

2.3.1 Interviews and study coverage

Semi-structured interviews

Topics, questions, link to full data and questionnaires in appendix.

Regional and university hospitals in France: different levels of maturity

Scope of 32 CHUs, out of the 3000 care sites in France to yield exhaustive conclusion on a restricted scope. Drive most specialized care, research in their core mission. Date of interview

Focus on the 18 CDWs with highest level of maturity

The denominator for the quantitative results is the 18 CDWs in production

2.3.2 A classification of observational studies

Contrast with classical epidemiology study types and the notion of experiment (K. J. Rothman, 2012).

Outcome frequency

Population characterization

Risk factors

Treatment effect

Development of diagnostic or prognostic algorithms

Medical informatics

2.4 Observations from a rapidly evolving and heterogeneous ecosystem

2.4.1 Governance: CDW are federating multiple teams in the hospital

Initiation and actors

Figure temporality of CDW Federating potential Multiple departments involved Multiple skills involved, strong ties with the academics In-house solution development vs industrialization

Management of studies

Scientific committee and project follow-up platform

2.4.2 Uneven transparency of ongoing studies

Uneven public reference on hospital websites of ongoing studies.

2.4.3 Triple usage of data: Research, management, clinic

Strong dependance to the HIS

Data collected reflect data collection Exemples, AP-HP, HCLs

Categories of Data

Main functionalities of HIS are the same. Common base Details: Table Takeaway: most of the current accessible data are billing, administrative and text, importantly, low access to temporality.

Data reuse: Research

Details of current study types. Details of specialty of the principal investigator Interest for research data network but lack of resources

Data reuse: CDW are used for monitoring and management

Initialization for billing Potential for professional feedbacks Pharmacovigilance

Data reuse: Strong interest for CDW in the context of care

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). There is a growing interest in such computational phenotyping tools to support the development of digital health solutions (Wen et al., 2023). 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.

2.4.4 A multi-layered technical architecture

Three layer : Data preprocessing (acquisition and normalization), storage, exposure Datalab: a crucial technological brick

2.4.5 Too little data quality and too many data formats

Quality tools

Automatic tooling for acquisition First development for automatic data checks

Standard format

No single standard data model, Omop eHop

Documentation

Half of the CDWs have put in place documentation accessible within the organization No documentation is public

2.5 Recommendations: How to consolidate EHRs and expand usages

2.5.1 Governance: CDWs are infrastructures

CDW becomes an essential component of data management in the hospital Resources specific to the warehouse are rare and only project-based. Should promote long-middle term teams (eg. inria ?) Multi-layered governance

2.5.2 Transparency: Keep the bar high

Public registration of comparative observational study protocols for research Patient information stays limited

2.5.3 Data: Complex data collection requires a variety of expertise

study design : change of focus from data collection to data preprocessing -> other complementary skills needed Link with the HIS, lack of standard at the HIS level, lack of sharing of data schema Data reuse oriented towards primary care is still rare and rarely supported by appropriate funding.

2.5.4 Technical architecture: Towards more harmonization and open source ?

Lack of harmonization: focus on fewer solutions Commercial solutions emerging Case for open source: transparency, less technological lock-in, mutualization, favor modularity, help build consensus Opportunity for open source solutions Data quality, standard formats Quality is not sufficiently considered as a relevant scientific topic itself. Tooling: Link with devops/ automated CI in industrial data science

there is a need for open-source publication of research code to ensure quality retrospective research

2.5.5 Data quality an document: more incentives needed

Quality is not sufficiently considered as a relevant scientific topic itself. However, it is the backbone of all research done within a 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 [52,54–56]. These studies should contribute to a reflection on methodologies and standard tools for data quality, such as those developed by the OHDSI research network [41].

Finally, there is a need for open-source publication of research code to ensure quality retrospective research [55,57]. Recent research in data analysis has shown that innumerable biases can lurk in training data sets [58,59]. Open publication of data schemas is considered an indispensable prerequisite for all data science and artificial intelligence uses [58]. Inspired by data set cards [58] and data set publication guides, it would be interesting to define a standard CDW card documenting the main data flows.

2.6 Conclusion

The French CDW ecosystem is beginning to take shape

The priority is the creation and perpetuation of multidisciplinary warehouse teams

Constitution of a multilevel collaboration network is another priority.

Common data model should be encouraged

The question of expanding the scope of the data beyond the purely hospital domain must be asked.

Heterogeneity of data collection calls for distributed information sharing models.

Chapter 3

Exploring a complexity gradient in representation and predictive algorithms for EHRs

3.1 Abstract

This chapter has not been submitted to any journals or conference until now.
XXX
As the first author I formulated the research question, designed the study, performed the experiments and wrote the manuscript.

3.2 Predictive models in healthcare: Less fascination, more practical utility

3.2.1 Why such interest in predictive models in healthcare

Interest in the clinic

Being able to predict key outcomes could, theoretically, make the use of hospital palliative care resources more efficient and precise (Topol, 2019)

The Framingham risk score, one of the earliest predictive model 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).

Harrell et al., 2001 mention diagnosis, prognosis, confounder adjustment and heterogeneous treatment effects (mainly for cost-effectiveness analyses) as the main uses of predictive models in healthcare. Ten years later, Ewout W Steyerberg; E. Steyerberg, 2009 also mentions diagnosis, prognosis and therapy (He has a great figure from pubmed). I should do the same for 2023.

Some are used in clinical practice every day : the Glasgow coma scale (Teasdale; Jennett, 1974), the APACHE III score (Knaus et al., 1991). But a very small part of the published models are used in practice (Wyatt; Altman, 1995). Reasons for this poor adoption are lack of evidences for credibility, generalizability or clinical effectiveness.

Prevention : *Early detection and appropriate treatment of sepsis have been associated with a significant mortality benefit in hospitalized patients* (Wong et al., 2021)

Alert systems (Yu et al., 2018).

Patient deterioration detection (M. J. Rothman et al., 2013): *Rather than attempting to forecast a particular adverse event, we argue that intervention during early deterioration can help prevent such an adverse event from occurring*

Risk stratification

risk stratify (Tang et al., 2007)

Long term prevention

Planning and piloting

Rejoin the logistic and administrative help from Topol, 2019.

The LOS task: planning the number of beds and members of staff required, identifying individual outliers and case mix correction for benchmarking (Verburg et al., 2017)

Exploring french CDWs:

23 % of studies, just after population definitions).

3.2.2 Data sources fueling predictive models are increasingly complex

Why is EHR so complex ? How ? time, high cardinality, multi-modality

In the literature

: early article in Wu et al., 2010, literature review from 2017 (Goldstein et al., 2017), literature review more focused on models and task from soda prez.

Within french CDW(link with previous chapter)

3.2.3 Low prevalence and local practice: Headaches for machine learning

For logistic models, a requirement is that the test set should have at least 10 cases per feature with a positive event (eg. mortality) (Harrell Jr et al., 1985; Wyatt; Altman, 1995).

3.2.4 What makes a healthcare predictive model useful?

What are the most impactful algorithms for predictive tasks from structured EHR ? A wealth of methods, but a lack of insights on the advantages and inconvenience for specific problems and resources.

We build upon the criteria for adoption from Wyatt; Altman, 1995.

More modern sources eg. Subbaswamy; Saria, 2020.

We can share it easily:

Hence, people can use it AND contribute to external validity and continuous improvements.

Performances

linked to accuracy, but questions on how to measure it.

Insertion in the care workflow

Arguments for ease of deployment.

3.3 From basic to complex: four increasingly sophisticated predictive pipelines

3.3.1 Demographic features

3.3.2 Count encoding with event features

3.3.3 Static Embeddings of event features

3.3.4 Transformer based

3.4 Empirical Study

3.4.1 Evaluation pipeline

3.4.2 Three clinical and operational tasks

Length Of Stay: Plan

Diagnosis Prediction: Benchmark

Major Adverse Cardiovascular Events: Prevent

3.4.3 Results: Performance-sample trade-offs

3.5 Conclusion

Chapter 4

Prediction is not all we need: Causal analysis of EHRs to ground decision making

4.1 Abstract

> Nature marks each growth ... according to its curative benefit.
>
> – Paracelsus

This chapter has been submitted to Nature Digital Medicine as:
XXX
As the first author I formulated the research question, designed the study, performed the experiments and wrote the manuscript.

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

Even 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, the use-case of heart failure prediction is motivated by agressive 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).*. It is almost always the case that prognosis models are motivated by decision making processes. It is clearly described by Ewout W Steyerberg; E. 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.*

4.2.1 Predictive medicine currently suffers from biases

(shortcuts, population shifts). Racial, gender and under-served population biases raise concern on fairness.

4.2.2 A key ingredient to ground data-driven decision making is causal thinking

4.2.3 The need for synthetic materials for practitioners

The relevant concepts for causal inference are scattered in different literatures. A dedicated exposition to time-varying data in EHRs would help practitioners and data scientists that study them.

4.2.4 Motivating example

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

4.3.1 Robust study design to avoid biases: Framing the question

PICOT Selection Bias Immortal time bias

4.3.2 Is the dataset sufficient to inform on the intervention: identification

Causal graph Computing the causal effect of interest: Estimation Confounders extractions Confounders aggregation Causal estimators Nuisance estimators

4.3.3 Assessing the robustness of the hypothesis: Vibration or sensitivity analysis

Sensitivity vs robustness

4.3.4 Heterogeneity of treatment

Interest of HTE How to do HTE ? Final regression analysis.

4.4 Application on MIMIC-IV

Emulated trial: Effect of albumin in combination with crystalloids compared to crystalloids alone on 28-day mortality in sepsis patients Choice of the trial Known effects

4.4.1 Framing the question

4.4.2 Identification

Covariates and dag

4.4.3 Estimation

Confounders extractions Confounders aggregation Causal estimators Nuisance estimators

4.4.4 Vibration analysis

Varying estimation choices: Varying inclusion criteria: illustration of immortal time bias

4.4.5 Heterogeneity of treatment

4.5 Discussion

Chapter 5

Conclusion

Modern healthcare burdens and costs are driven by chronic diseases where death is not the only outcome of interest. Focus on smaller rewards, on which experiments are easier to conduct and where error is possible: it would allow better learning for decision making since we can repeat more experiments (limit: signals could be highly delayed).

Text is pervasive, we are using it to communicate and to log most of our information. We can rely pretraining models outside the healthcare domain. We should leverage it more in the context of care (limit: temporality is hard to capture but is a key aspect of causal inference).

The unreasonable effectiveness of healthcare data is yet out of reach due to hugely difficult transfer of models or administrative barriers to access data (due to multiplicity of the involved actors). This forces us to rely on efficient techniques that make the best of medium-sized data or rely on sharable sources of knowledge (aggregated statistics, federated learning approaches, ontologies, ...).

Bibliography

Bareinboim, Elias, Andrew Forney; Judea Pearl (2015): "Bandits with unobserved confounders: A causal approach". In: *Advances in Neural Information Processing Systems* 28.

Beam, Andrew L, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai; Isaac 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.

Beaulieu-Jones, Brett K, William Yuan, Gabriel A Brat, Andrew L Beam, Griffin Weber, Marshall Ruffin; Isaac 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.

Brand, Richard J, Ray H Rosenman, Robert I Sholtz; Meyer 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.

Breiman, Leo (2001): "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical science* 16.3, pp. 199–231.

Dernoncourt, Franck, Ji Young Lee, Ozlem Uzuner; Peter Szolovits (2017): "De-identification of patient notes with recurrent neural networks". In: *Journal of the American Medical Informatics Association* 24.3, pp. 596–606.

Garcelon, Nicolas, Antoine Neuraz, Vincent Benoit, Rémi Salomon, Sven Kracker, Felipe Suarez, Nadia Bahi-Buisson, Smail Hadj-Rabia, Alain Fischer, Arnold Munnich, et al. (2017): "Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack". In: *Journal of biomedical informatics* 73, pp. 51–61.

Goldstein, Benjamin A, Ann Marie Navar, Michael J Pencina; John PA 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.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf; Alexander Smola (2012): "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1, pp. 723–773.

Halevy, Alon, Peter Norvig; Fernando Pereira (2009): "The unreasonable

effectiveness of data". In: *IEEE intelligent systems* 24.2, pp. 8–12.

Harrell, Frank E et al. (2001): *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Vol. 608. Springer.

Harrell Jr, Frank E, Kerry L Lee, David B Matchar; Thomas A Reichert (1985): "Regression models for prognostic prediction: advantages, problems, and suggested solutions." In: *Cancer treatment reports* 69.10, pp. 1071–1077.

Haug, Charlotte J; Jeffrey M Drazen (2023): "Artificial intelligence and machine learning in clinical medicine, 2023". In: *New England Journal of Medicine* 388.13, pp. 1201–1208.

Hernàn, Miguel A; James M Robins (2020): *Causal inference: What If*.

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

Jiang, Lavender Yao, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. (2023): "Health system-scale language models are all-purpose prediction engines". In: *Nature*, pp. 1–6.

Johansson, Fredrik D, Uri Shalit, Nathan Kallus; David Sontag (2022): *Generalization bounds and representation learning for estimation of potential outcomes and causal effects*.

Knaus, William A, Douglas P Wagner, Elizabeth A Draper, Jack E Zimmerman, Marilyn Bergner, Paulo G Bastos, Carl A Sirio, Donald J Murphy, Ted Lotring, Anne 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.

Lee, Peter, Sebastien Bubeck; Joseph 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.

Li, Yikuan, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi; Gholamreza Salimi-Khorshidi (2020): "BEHRT: transformer for electronic health records". In: *Scientific reports* 10.1, pp. 1–12.

MacMahon, Brian, Thomas F Pugh, et al. (1970): "Epidemiology: principles and methods." In: *Epidemiology: principles and methods*.

McGinnis, J Michael, Leigh Stuckhardt, Robert Saunders, Mark Smith, et al. (2013): "Best care at lower cost: the path to continuously learning health care in America". In.

Murphy, Kevin P (2022): *Probabilistic machine learning: Advanced topics*. MIT press.

Naimi, Ashley I; Brian W Whitcomb (2023): "Defining and Identifying Average Treatment Effects". In: *American Journal of Epidemiology*.

Paris, Nicolas, Matthieu Doutreligne, Adrien Parrot; Xavier Tannier (2019): "Désidentification de comptes-rendus hospitaliers dans une base de données OMOP". In: *TALMED 2019: Symposium satellite francophone sur le traitement automatique des langues dans le domaine biomédical*.

Patel, Vimla L, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi; Ameen Abu-Hanna (2009): "The coming of age of artificial intelligence in medicine". In: *Artificial intelligence in medicine* 46.1, pp. 5–17.

Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. (2018): "Scalable and accurate deep learning with electronic health records". In: *NPJ digital medicine* 1.1, p. 18.

Rothman, Kenneth J (2012): *Epidemiology: an introduction*. Oxford university press.

Rothman, Michael J, Steven I Rothman; Joseph 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.

Rutstein, David D (1967): "The coming revolution in medicine". In.

Shalit, Uri, Fredrik D Johansson; David Sontag (2017): "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085.

Shortliffe, Edward Hance (1974): "MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection". PhD thesis. Stanford University Ph. D. dissertation.

Sriperumbudur, Bharath K., Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf; Gert R. G. Lanckriet (Oct. 12, 2009): "On integral probability metrics, \phi-divergences and binary classification". In: *arXiv:0901.2698 [cs, math]*.

Steyerberg, Ewout W; EW Steyerberg (2009): *Applications of prediction models*. Springer.

Subbaswamy, Adarsh; Suchi Saria (2020): "From development to deployment: dataset shift, causality, and shift-stable models in health AI". In: *Biostatistics* 21.2, pp. 345–352.

Tang, Eng Wei, Cheuk-Kit Wong; Peter 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.

Teasdale, Graham; Bryan Jennett (1974): "Assessment of coma and impaired consciousness: a practical scale". In: *The Lancet* 304.7872, pp. 81–84.

Topol, Eric J (2019): "High-performance medicine: the convergence of human and artificial intelligence". In: *Nature medicine* 25.1, pp. 44–56.

Verburg, Ilona Willempje Maria, Alireza Atashi, Saeid Eslami, Rebecca Holman, Ameen Abu-Hanna, Everet de Jonge, Niels Peek; Nicolette Fransisca 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.

Wen, Andrew, Huan He, Sunyang Fu, Sijia Liu, Kurt Miller, Liwei Wang, Kirk E Roberts, Steven D Bedrick, William R Hersh; Hongfang Liu (2023): "The IMPACT framework and implementation for accessible in silico clinical phenotyping in the digital era". In: *npj Digital Medicine* 6.1, p. 132.

Wong, Andrew, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen 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.

Wu, Jionglin, Jason Roy; Walter F Stewart (2010): "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches". In: *Medical care*, S106–S113.

Wyatt, Jeremy C; Douglas G Altman (1995): "Commentary: Prognostic models: clinically useful or quickly forgotten?" In: *Bmj* 311.7019, pp. 1539–1541.

Yu, Kun-Hsing, Andrew L Beam; Isaac S Kohane (2018): "Artificial intelligence in healthcare". In: *Nature biomedical engineering* 2.10, pp. 719–731.

Zhou, S Kevin, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert; Ronald M Summers (2021): "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.

Appendices

Appendix A

Chapter 5

1.1 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 $\mathbb{P}(X|A)$:

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

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 A.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 **??**,
- Verifying that setups with penalized overlap from ACIC 2016 have a higher total variation distance than unpenalized setups in Figure **??**.
- Verifying that the Inverse Propensity Weights extrema (the inverse of the ν overlap constant appearing in the overlap Assumption **??**) positevely correlates with NTV for Caussim, ACIC 2016 and Twins in Figure **??**. 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 **??** where nuisance models are fitted only on the train set.

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 A.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.

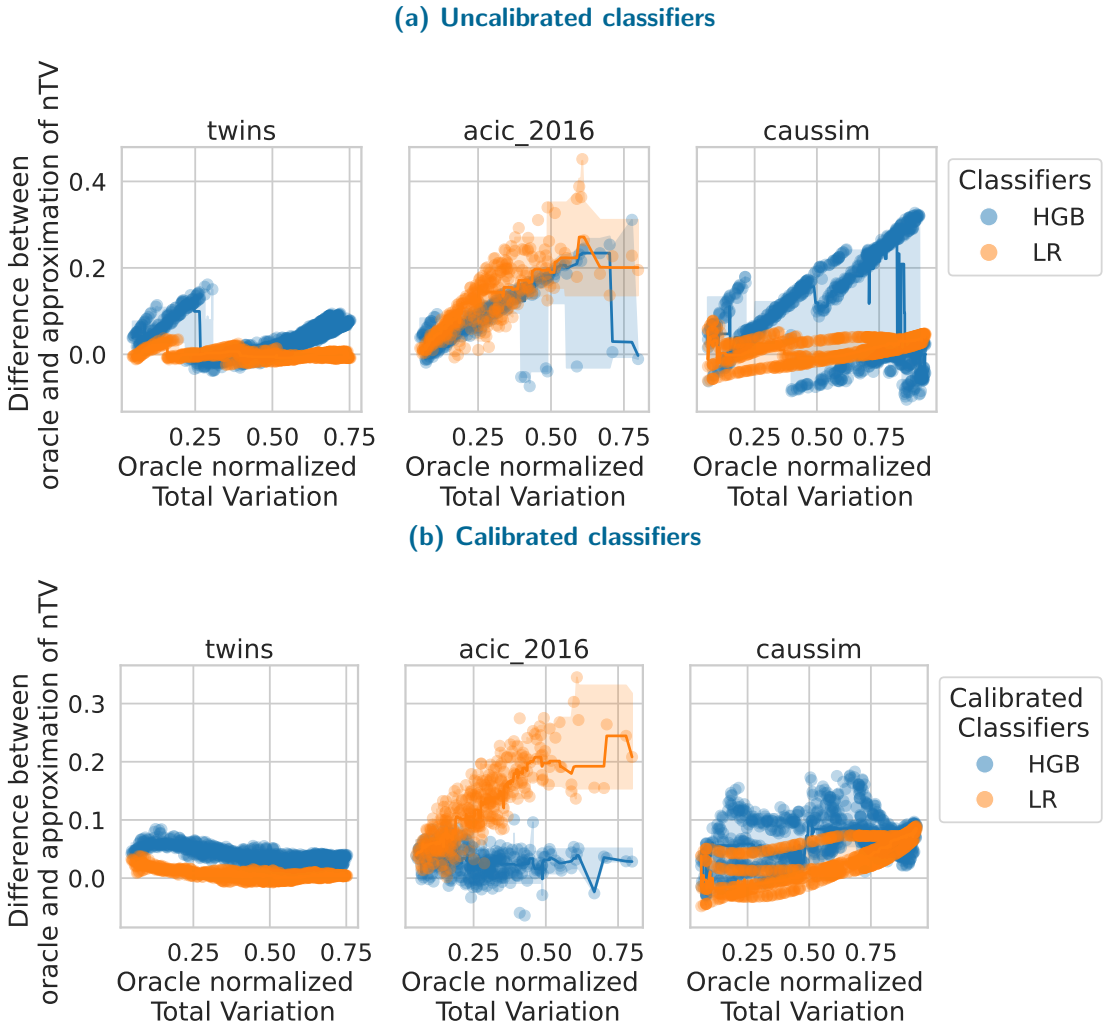

Fig. A.1. a) Without calibration, estimation of NTV is not trivial even for boosting models. b) Calibrated classifiers are able to recover the true Normalized Total Variation for all datasets where it is available.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

Titre: Titre de la thèse en français
Mots clés: Quelques mots-clé

Résumé: Blabla

Title: Thesis title in English
Keywords: Some keywords

Abstract: Blabla

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