UAB
Universitat Autònoma de Barcelona

Causal inference methods for generating evidence on the effects of health interventions

Borja Velasco Regúlez

Universitat
Autònoma
de Barcelona

PhD in Computer Science

Research line: Artificial Intelligence

Causal inference methods for generating evidence on the effects of health interventions

PhD Student: Borja Velasco Regúlez

PhD Advisor and tutor: Jesus Cerquídes Bueno

Contact mail: bvelasco@iiia.csic.es / borja.velasco@gencat.cat

Bellaterra (Cerdanyola del Vallès), November 10, 2024

To my family and friends

Abstract

Does the vaccine against COVID-19 cause alterations in the menstrual cycle? Does it protect against the infection-increased risk of diabetes? These are examples of causal questions about the effects of clinical interventions. They are causal because they verse about causes - in this case the COVID-19 vaccine - and effects or consequences - in these cases, alterations in the menstrual cycle and protection against diabetes-. These questions are both important and difficult. Important for the obvious reason that they concern aspects of human health. Difficult for the complexity of the systems under study: the human body, human health, and their interaction with clinical interventions. There are several approaches to answering this type of question. This thesis is concerned with the *data-driven, statistical* approach, particularly with using *observational* data, i.e., the data collected in scenarios where the clinical intervention of interest is not under the researchers' control.

Traditionally, statistical correlational methods have been used to answer these questions with this type of data. In general, these methods only provide correlations without guaranteeing their causal nature. Nevertheless, in recent years, developments in the field of *causal inference* have provided us with methods that can offer some certainty of the causality of the measured relationships under the appropriate assumptions. Until recently, researchers' adoption of these methods has been hindered by three main factors: unawareness about their existence, inertia of the traditional methods, and, to a lesser extent, lack of trust in their performance. This tendency, though, has consistently changed in recent years in the literature of clinical studies.

This thesis aims to test the hypothesis that causal inference methods should be the preferred choice for generating evidence on the effects of clinical interventions, with a particular focus on machine learning-based causal methods. For such purpose, we tackle three real-world use cases with real-world data, both using correlational and causal approaches, and we qualitatively assess and compare their performance (in a broad sense). In addition, we explore the field of machine learning (and mainly neural network)-based causal inference algorithms. The tackled questions are about the effect of the COVID-19 vaccine and vaccination timing on alterations of the menstrual cycle, the effect of the COVID-19 vaccine on the infection-heightened risk of diabetes onset, and the effect of antibiotic-loaded bone cement (a therapeutic option for patients undergoing total knee replacement surgery) on the survival of the prosthesis. Together with the aforementioned causal and correlational methods, we employ real-world observational data from large registries.

As a result, we provide answers to the posed questions. In some cases, the provided answers and/or the employed methods were novel in the literature at their time of publication. In addition, we offer qualitative evidence of the benefits of causal methods compared to correlational methods. We conclude that, in general, and when possible, causal inference methods should be the preferred choice for answering these types of questions with observational data (i.e., when randomized experiments cannot be conducted).

Contents

List of Figures

List of Tables

Acknowledgements

Doing a PhD is, like any other thing in life, an adventure. And adventures only exist in the companion of other human beings. With these lines, I want to say thanks to some of the human beings that have, in one way or another, participated in this adventure. It would not exist without them.

First and foremost, I want to say thanks to my PhD advisor, Jesus Cerquides. Someone with experience in the topic told me once that one of the most important things for successfully doing a PhD was your PhD advisor. That helped me make the decision of pursuing it, as I was convinced that I was going to be in good hands. 4 years later, I know that I was right. Jesus, thank you for your availability, your help and your guidance, most of the times about academic aspects, but a few other times about personal and life aspects too.

I want to thank my company tutor and manager, Ramon Roman, for giving me the opportunity of pursuing this industrial PhD and for effectively protecting my time for doing it. I also want to thank all my colleagues at AQuAS, especially the PADRIS team, my team, for their company throughout this journey and for having my back at work along these four years. It would have been much more difficult without your help. Also to the IIIA family, thinking about what it could have been if the pandemic had not existed.

A mi familia, mi padre, mi madre y mi hermana, pilares de mi vida y una de las principales razones de mi feliz existencia. Con un poquito de sorna diré que esta tesis va sobre causas y efectos y que vosotros, aita y ama, sois literalmente la causa de que yo esté aquí. Os estoy infinitamente agradecido por eso y por todo lo demás. A mis tios, tias, primos y primas, y en especial a mi tia Mari Carmen, por haber visto los dificiles comienzos de este trayecto (y me refiero al primer año de la carrera de ingeniería en Madrid).

Gasteizko eta Izarrako betiko lagunei, betidanik nire bizitzan egoteagatik eta betirako egongo direlako esperantzarekin. Espero dut nik zuen bizitzan zuek nirean jartzen duzuen adina zoriontasuna jartzea. Jon, Beñat, Markel, Igor, Ibon, Aitor, Ibai, Ivan, Andrea, Clara, Jessica eta beste guztiak. Maite zaituztet.

A los que llegasteis después a mi vida y compartisteis algunas de las etapas más felices de ella, y aun seguís ahí: Ruben, Carlos, Dani, Alfonso, Cesar, Guille, Alex, Mireia, María, Ares y los y las demás. A todos aquellos y aquellas con los que perdí el contacto, también.

To Florian Messerer, one of the brightest people I know, and who has most challenged me intellectually. This thesis speaks about counterfactuals, and I cannot avoid thinking of what the counterfactual would be today if you had not talked to me about The Book of Why five years back.

Aquesta tesi ha estat desenvolupada a Catalunya, i per tant sempre estarà lligada a aquesta terra, que ja és i sempre serà la meva segona casa. Gràcies a pinuts i cacahuets, a ”els de la uni”, i a ”els de la Pia”. Gracias Adriana por abrirme huella en la nieve, uno de los lugares más felices del mundo para mi, y a Andrea por integrar el equipo perfecto. Gracias a Natalia, Ash, Noelia, Tessa, Glòria, Lara, Fede, Laia, Alex, Camilla y los demás. A Pere, Roser, Niels i Glòria. To the people from Newcastle, for amazingly strange amazing times.

To Innsbruck and its mountains, if you can say thanks to a place. To Sonja, Paz and Júlia. To Uwe Siebert and Lara Hallsson, for some of the most interesting, enriching and inspiring conversations about causal inference.

Finalment, a tu, Sílvia. Aquesta tesi no existiria sense tu, ja que amb el teu exemple tu vas ser una de les principals fonts d'inspiració per fer-la. Gràcies per acompanyar-me al llarg de tot aquest camí. Traient ferro, diré que ha estat exactament el 96% del camí, però que per a mi és com si hagués estat el 100%. Gràcies, de tot cor, per tot, i per sempre.

Chapter 1

Introduction

1.1 Main goal, motivation, and context of this thesis

Determining the effect that an intervention has on the health of a person or a group of people is a task that is both difficult and important. Difficult, because the human body and human health are very complex systems in which many variables are involved. And important, for the obvious reason that health is one of the most precious things for human beings. This thesis revolves around methods for measuring the effects of interventions on human health, and in particular, it aims to determine whether a certain methodology, called causal inference, should be the chosen one for *producing evidence* about the effects of clinical interventions. The motivation for answering this question is straightforward: better-suited methods will produce better evidence, and that, in turn, will result in better health.

In the remainder of this section, we provide context for the aforementioned goal and motivation, explaining what causality, causal inference, and clinical evidence are. We divide the content into three subsections.

1.1.1 Causality

Causality, or causation, is the influence by which an event, state, or process (the cause) contributes to the occurrence of another event, state or process (the consequence). Thus, the cause is partially or totally responsible for the consequence, and the consequence is partially or totally dependent on the cause. This concept is intrinsic to almost every field of human knowledge, from philosophy to physics, and it is an abstraction about *how things work* in the universe (Mackie, 1980).

In some domains of physics, we have very precise tools to describe *how things work*: we have physical laws that can be expressed as mathematical equations. For example, we know that the gravitational force two bodies exert on each other, or the amount of heat a body releases, are phenomena that follow well-known physical laws and equations. In such cases, the concept of causality is sort of trivial, as it can be derived directly from the equations: for instance, the action of doubling the mass of the two bodies (the cause) of the first example will have as a result that the gravitational force multiplies by four (the consequence); and the action of doubling the temperature of the body (the cause) of the second example, will have as a result that the released thermal energy doubles (the consequence).

Nevertheless, when the object of interest is the human body and human health, things are not that simple. In that case, in general, we do not have laws that translate into equations that describe *how things work*, at least not in the same precise way as in the previous examples. That does not mean that we have no knowledge about the working mechanisms of the human body, on the contrary, we do have it: we know about physiology, biochemistry, genetics and much more. But if we want to know the effect of a particular intervention on the human body (for instance, exposing it to a certain amount of heat or treating it with a new drug), it is usually not possible to combine all the knowledge from the aforementioned fields to come up with a law and equation that will describe the effect of such intervention. This is mostly due to the extraordinary complexity of the system at hand, with many, many variables involved in the problem. Instead, we usually need to take a so-called *data-driven, statistical* approach. And then, the concept of causality becomes much trickier. Why? Because, as every statistician reminds us, variables can be statistically correlated without being causally related (Gershman and Ullman, 2023).

A mandatory ingredient for the *data-driven* approach to *describing how things work* is, obviously, data. When we want to know the effect of a given intervention or treatment on some outcome variable of human health, collecting data about it can be done in two ways: *experimentally* or *observationally*. The experimental way means that the researcher runs an experiment and collects data about it, which implies that they can influence the way the data is *generated* via the experimental design. The observational way means that the researcher *observes* the data that has been generated by a process over which they had no influence capacity. The distinction of these two types of data has major implications for the task of inferring and measuring causal relationships.

The most important feature of experimental data is that the intervention or treatment of interest was *randomly* assigned to patients. This is the case of the data generated through randomized controlled experiments or trials (RCTs), which have traditionally been considered the gold standard method for inferring and measuring causality in the domain of human health (Gerstman, 2023; Hariton and Locascio, 2018). The steps for conducting a randomized controlled experiment (in its basic form) are simple: a group of patients is *randomly* divided into two groups; one group gets the intervention of interest, and the other group does not; the outcome variable of interest is measured in both groups; the difference observed in such variable between the groups, if any, is the causal effect of the intervention. Remind that the causal effect is the influence that the cause (in this case, the treatment) has on the consequence (in this case, the outcome). The working principle of this method is also straightforward: randomization ensures that all factors influencing the outcome, except for the intervention, are equal in both groups (up to random variability), and thus the observed difference in the outcome can only be caused by the intervention itself.

On the contrary, in the case of observational data, the intervention of interest is not assigned randomly but based on some factors (mostly, although not exclusively, patient characteristics). If those factors, besides influencing the chances of getting the intervention, also influence the outcome, we have a problem. Because the distribution of factors will not be equal in the intervention and the nonintervention groups, the observed differences in the outcome will not only be caused by the intervention but also by the differences in the factors. That effect is usually known as confounding, and the factors producing it are known as confounders or covariates. Fortunately, statistics provides us with tools to remove the influence of confounders on the outcome, leaving only the effect of interest, i.e., that of the intervention. But these tools are not free of limitations: in order to be able to remove the effect of confounders, we need, obviously, data about them. What if we do not have such data, or worse, if we are not aware of the existence of some confounder? Then, the obtained causal effect will be biased.

If experimental data is the best for inferring and measuring causal effects, and observational data has the limitations that we just explained, why not just conduct RCTs to answer every question about the effects of interventions on human health? Well, because things are not that simple: conducting RCTs can be unfeasible or unethical. Consider, for example, the effect of smoking on the probability of developing cancer. It may sound trivial, and it is nowadays, after decades of accumulated evidence, but for some years there was a big debate about whether such an effect existed or not, and it was not possible to just randomly assign a group of people to smoke due to obvious ethical reasons. That is when observational data and observational studies come in handy.

The potential presence of bias in observational studies due to confounding does not change the fact that the goal of those studies is to infer and measure causal effects of interventions. Yet historically, some scientists have refrained from explicitly talking about causality when working with observational data and have exhorted other fellow scientists to do the same. Their justification was that they were putting a safeguard in place to avoid mistaking correlation with causation. It may have been a reasonable strategy until recently, but things have changed: advancements in the field of causal inference have *formally proven* that causality can actually be inferred and measured from observational data under certain conditions and assumptions (Judea Pearl, 2009a). But what is causal inference? we provide a definition of this concept in the following subsection.

1.1.2 Causal inference

Causal inference can be defined as a *framework* for doing exactly what we mentioned in the previous paragraph: inferring and measuring causal relationships from observational data. This framework provides us with three main things: a set of mathematically defined assumptions common to every problem, statistical estimators for measuring the quantities of interest, and formal proof or guarantee of the causal nature of the estimated quantities, given that the assumptions are met. These elements differentiate causal inference from traditional correlational analysis of observational data, which make no explicit assumptions (they still do it implicitly) and provide no guarantees about the causal nature of the measured correlations. Of course, causal inference is not free of limitations: assumptions still need to be made. However, the fact that they are explicit and thus easier to discuss or challenge represents a big step forward with respect to associational-only methods.

Note that within the framework of causal inference, there is a plethora of elements. There are two main sub-frameworks, named potential outcomes framework or Rubin causal model (D. Rubin, 1972; Imbens and D. B. Rubin, 2010), and structural causal models' framework (Judea Pearl, 1995; J. Pearl, 2000), different tools such as directed acyclic graphs, and a myriad of algorithms and estimators for as many types of scenarios. We will introduce the most important ones in Chapter 2 of this thesis.

In the next section, we define the other key element of the main goal of this thesis: evidence about the effects of clinical interventions. Recall that evidence is what determines if interventions *work*, or which which interventions *work better*.

1.1.3 Evidence generation for clinical interventions

Evidence, in this context, is *proof* in favor or against some hypothesis or claim: for instance, proof about a particular intervention being more effective than another for a certain purpose. Such evidence is obtained by analyzing data, and its accumulation is what eventually makes a hypothesis or claim a truth or a falsehood.

A particular domain of human health-related studies that requires evidence generation about interventions is that of health technology assessment (HTA). HTA is a multidisciplinary process for evaluating the properties and *effects* of a health technology (Lampe et al., 2009). Such a process must use state-of-the-art methods to consider and/or *generate* the best possible evidence. In certain situations, the process leads to the conclusion that the available evidence is insufficient or of low quality, and it is recommended that new evidence is generated. In general, ideally, this would be done through a randomized experiment study, but if that is not possible, an observational study should be conducted, and that is when causal inference comes into play. HTA is crucial for several reasons, but we highlight two of them here: from the perspective of patients, it aids in ensuring that they get the most effective interventions, and from the perspective of health providers, it aids in ensuring the sustainability of health systems.

In the next section, we define the specific research questions and the structure of this thesis.

1.2 Research questions and structure of this thesis

1.2.1 Research questions

As stated at the beginning of this chapter, the general goal of this thesis is to determine whether causal inference should be the method of choice for producing evidence about the effects of clinical interventions, in particular in the context of HTA. For that purpose, different observational data analysis methodologies will be assessed, ranging from traditional correlation analysis to causal inference methods. The investigation is carried forward by applying the different methodologies to several real-world use cases with real-world data. In particular, two health technologies are studied, which gives rise to three different questions of interest. Those health technologies are the COVID-19 vaccine on one hand and antibiotic-loaded bone cement on the other hand, which is a treatment option used during knee replacement surgery. The COVID-19 pandemic had a big influence on the fact that a big part of this thesis focuses on the health technology of COVID-19 vaccines. The questions addressed about these health technologies and the analysis methodologies are presented in the following paragraphs.

- **Q1:** *Do the vaccine against COVID-19 and the vaccination time have any effect on the menstrual cycle?* The interest in this question sparked when several hundreds of women reported, mostly through social networks, changes in their menstrual cycles after getting the COVID-19 vaccine. Menstrual cycle stability is an important indicator of menstruating people's reproductive and overall health (Mihm, Gangooly, and Muttukrishna, 2011), and that motivates the interest in the question. This analysis was performed using a correlational approach.

- **Q2:** *Does the vaccine against COVID-19 have any effect on the risk of developing diabetes?* This question was motivated by the publication of several reports indicating that the COVID-19 infection could increase the risk of diabetes. It was natural then to wonder whether the vaccine against the infection would have any impact, either reducing that risk, leaving it unaffected, or even increasing it. This analysis was performed using a causal approach.

- **Q3:** *Does the use of antibiotic-loaded bone cement during total knee arthroplasty surgery increase the life of knee prostheses, compared with the alternative of using plain cement?* There are already several studies and meta-analyses (evidence aggregators) about this topic in the literature, but the question is still considered open: more and better evidence is needed (T. H. Leta et al., 2021). The extension of the life of the prosthesis has an important positive impact on the quality of life of patients who undergo knee arthroplasty surgery, as well as an important impact on healthcare systems in the form of savings from avoided prosthetic revisions. This analysis was performed using both a correlational and a causal approach.

- **Q4:** *What are the advantages and disadvantages, strengths and weaknesses, of correlational and causal inference methods for generating evidence about clinical interventions?* By splitting this question, we define two sub-questions, **Q4a:** *What are the advantages and disadvantages, strengths and weaknesses, of correlational methods for generating evidence about clinical interventions?* and **Q4b:** *What are the advantages and disadvantages, strengths and weaknesses, of causal inference methods for generating evidence about clinical interventions?* We conducted qualitative critical assessments of the performance (in a wide sense) of correlational and causal inference methods, as well as a qualitative comparison between their results, using the use cases of questions **Q1**, **Q2** and **Q3** as a basis. In particular, the assessment of the correlational method (**Q4a**) was performed with the use case of **Q1** (COVID-19 vaccine and menstrual cycle changes); the assessment of the causal inference method (**Q4b**), with the use case of **Q2** (COVID-19 vaccine and risk of diabetes); and the overall comparison between approaches, (**Q4**), with the use case of **Q3** (antibiotic-loaded bone cement and prosthetic survival).

- **Q5:** *Can we generalize advanced causal inference algorithms from binary treatment settings to multivalued treatment settings?* During the development of this thesis, we identified a big imbalance in the literature between causal inference algorithms for binary treatment settings and for *multivalued* treatment settings (note that a *multivalued* treatment is one for which the treatment can take more than two values). In particular, there is a trend in the field of causal inference to develop advanced algorithms, many of them based on machine learning techniques such as neural networks, with the goal of breaking state-of-the-art performance metrics. In most cases, such developments are conducted in binary treatment scenarios, and very little work is done in the realm of multivalued treatments. While studying the previously explained health technologies, we realized the usefulness of a multivalued treatment approach. For instance, consider how the question about the effects of the COVID-19 vaccine, which is originally binary (vaccine administered or not administered), becomes multivalued if we are interested in the effects of different vaccine brands. Similarly, the question about the effect of the antibiotic-loaded bone cement becomes multi-valued if we look at the effects of different antibiotics. For these reasons, we aimed to take a state-of-the-art, binary treatment causal inference algorithm and generalize it to a multivalued treatment setting.

Having defined the research questions, the next section provides an overview of how we structure this thesis to answer them.

1.2.2 Structure of this thesis

Chapter 2 presents the methods and the state of the art of this thesis. Then, Chapter 3 tackles the first research question, **Q1**, of whether the COVID-19 vaccine has any effect on the menstrual cycle. It also provides a critical assessment of the employed correlational method, answering research subquestion **Q4a** (advantages and disadvantages, strengths and weaknesses of correlational methods). Continuing with the health technology of the COVID-19 vaccine, Chapter 4 presents research question **Q2** of whether the vaccine has any effect on the risk of diabetes onset. It also provides a critical assessment of the employed causal inference method, answering research subquestion **Q4b**. Finishing the questions about health technologies, Chapter 5 presents research question **Q3** of whether antibiotic-loaded bone cement lengthens prosthetic survival. As this question is answered both with correlational and causal methods, this chapter provides a comparison between them, providing the answer to **Q4**. Then, Chapter 6 introduces multivalued treatment settings, showing how the previous clinical questions could have benefited from such an approach. It provides an answer to research question **Q5**, of whether is it possible to generalize advanced causal inference algorithms from binary to multivalued treatment settings. Finally, Chapter 7 verses about general conclusions and future work of this thesis.

To finish this introductory chapter, we provide an overview of the institutions involved in this work and the ethical aspects.

1.3 Institutions involved in this thesis

This thesis has been developed at two institutions: the Agency of Health Quality and Evaluation of Catalonia (AQuAS), which is the health technology assessment agency of that region, and the Artificial Intelligence Research Institute (IIIA) of the Spanish National Research Council (CSIC). The interest of the former, as a health technology assessment agency, was to study causal inference methods for health technology assessment. The interest of the latter focused mainly on the intersection of causal inference methods with machine learning, which is a strong trend in the field.

Within AQuAS, this work has been developed at the *Data and Artificial Intelligence* team, previously known as the Data Analysis Program for Research and Innovation in Healthcare (PADRIS). This team has access to and works with the data of the Catalan public healthcare system, combining a myriad of different sources. Among other tasks, it manages access to such databases by internal and external stakeholders, prepares data cohorts for research, and analyzes such data under the request of health system managers. This implies extensive and detailed knowledge about data-related infrastructures, data taxonomy, and data science.

1.4 Ethical aspects of this thesis

When conducting clinical studies with human data, obtaining the approval of an ethical committee is a legal requirement to safeguard participants' rights, preserve scientific integrity, and adhere to ethical research standards. Regulatory authorities in the European Union and Spain oversee the required permissions for conducting clinical observational studies within their jurisdictions. All the datasets used in this thesis are observational and have been analyzed with the approval of an ethical committee.

In particular, for the work in Chapter 3, we obtained the approval of the Spanish National Research Council's ethical committee, with internal number 129/2022. For the work in Chapter 5, we obtained the approval of Bellvitge Hospital's ethical committee, with number PR186/19, and approval of the Advisory Committee of RACat (Catalan Arthroplasty Register). The approval for the work of Chapter 4 is in progress at the ethical committee of the Spanish National Research Council (CSIC).

In the next chapter, we present the methods and the state of the art of this thesis.

Chapter 2

State of the art

Causal inference is a relatively new, multidisciplinary field of knowledge, and, as such, many of the developments and state-of-the-art methods have come and keep coming from authors working in other fields. Some of the most relevant fields are health and epidemiology, econometrics, computer science, and statistics. Because this thesis is about causal inference in healthcare settings, it is natural that this chapter will focus more on methods from that domain, but it will not be uncommon to look also at others, as methodological exchanges between domains are nowadays the rule more than the exception.

The rest of this chapter is organized as follows. First, we provide an overview of a selection of historical, critical methodological contributions to the field of causal inference. Then, we present the methods and the state of the art of this thesis. That second part is, in turn, divided into three subsections: a section about epidemiological study designs and statistical methods, another section about causal inference methods, and finally, a section about the clinical state of the art.

2.1 Brief overview of a selection of historical, critical methodological contributions to the field of causal inference

In the following paragraphs, we present a selection of works that were considered breakthroughs in the field of causal inference, especially, but not only, in its application to health. It is not an exhaustive or objective list, but it contains some of the works that have been most influential for the field and for the current thesis, according to this author.

In 1855, John Snow, a British physician, published a work titled *On the mode of communication of cholera* (Snow, 1855). A year earlier, in August 1854, a severe outbreak of cholera occurred in the Soho district of London, killing over 500 people in a few days. Cholera was a major public health threat in Europe back in those days, and there were two competing theories among physicians and scientists about its causes: the miasma theory and the germ theory. The former stated that cholera was caused by particles that would transmit through the air (airborne) and was the most accepted theory. In contrast, germ theory stated that cholera was caused by an unknown germ, that would be transmitted through contaminated water or food (waterborne). Following the outbreak, Snow set out to try to find its *cause*. He talked to local residents in Broad Street, where the outbreak had caused most casualties, and elaborated a map where he depicted the place of residence of the death. Thanks to this and other evidence, he identified the water pump at Broad Street as the potential source of the outbreak and (temporarily) convinced the local authorities to perform an *intervention*: to remove the handle of that pump. After that, the number of new cholera cases declined rapidly (nevertheless, for the sake of rigor, it must be mentioned that the decline was not only caused by the intervention but also due to the natural dynamics of the outbreak). Snow's investigations about cholera are considered a major cornerstone in the field of epidemiology, which is defined as "the study of the *determinants*," i.e., the *causes*, "occurrence, and distribution of health and disease in a defined population" (Brachman, 1996). It is one of the applied fields of knowledge that has most consistently nurtured causal inference, both with concepts and methods.

In 1986, James Robins published a work titled *A new approach to causal inference in mortality studies with a sustained exposure period — application to control of the healthy worker survivor effect* (J. Robins, 1986), in the *Mathematical Modeling* journal. In that work, he presented the *G-computation algorithm*, nowadays most popularly known as the *G-formula*. For understanding its relevance, let us bring back the definition of confounding from Chapter 1, and extend it to time-varying confounding. A confounder is a variable that influences both the treatment and the outcome of a problem, and time-varying-confounding (also called confounder-treatment loop) refers to the scenario in which confounders affect the treatment and vice-versa over time. The G-formula was the first mathematical solution to the problem of time-varying confounding, which cannot be solved with other adjusting methods (recall also that adjusting is the statistical procedure to remove the influence of factors other than the treatment from the outcome). Besides the practical implications of the development of the method, such as its application to real-world use cases, it constituted a major step forward in the field of causal inference in general, as it laid the groundwork for a more rigorous approach to causality in observational studies.

In 1995, Judea Pearl published *Causal diagrams for empirical research* (Judea Pearl, 1995) in the *Biometrika* journal. This was the first work proposing the use of directed acyclic graphs for causal inference problems as a way to graphically encode prior knowledge about the problem domain and to combine domain and statistical knowledge. The work also showed how diagrams could be used to determine if assumptions made about the problem were sufficient for identifying causal effects from observational data and for querying those diagrams to produce mathematical expressions for causal effects in terms of observed distributions. This tool enabled the description and analysis of types of variables with causal relationships that could not be described before, such as mediators and colliders. Besides, it made the task of causal query identification substantially more simple and intuitive, especially in problems with complex causal structures. This paper and its related work (Judea Pearl, 2009b) changed the paradigm of inferring causality from observational data, especially regarding causal query identification.

In 2016, Hernan and Robins introduced the concept of target trial emulation in a paper titled *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available* (Miguel A. Hernán and James M. Robins, 2016), in the *American Journal of Epidemiology*. A target trial is a hypothetical randomized controlled trial that researchers would ideally conduct. The authors explained how to systematically emulate such a trial using observational data. This paper represents another breakthrough of causal inference, especially in the field of health, because its methodology can eliminate or at least reduce some of the worst sources of bias present in correlational-only approaches for observational studies: time-related biases. Besides, empirical evidence suggests that the method *works*, in the sense that emulated target trials have obtained results that are similar to those from actual randomized trials, unlike some other observational data-based methodologies (Kuehne et al., 2022; S. V. Wang, Schneeweiss, and Initiative, 2023).

In 2021, Joshua Angrist, Guido Imbens, and David Card won the Nobel prize in economic sciences for their work on *natural experiments*, a method to identify and measure causal effects in social sciences (Card and Alan B Krueger, 1994; Imbens and Angrist, 1994; Angrist and Alan B. Krueger, 1991). Although the method itself falls slightly out of the scope of this thesis, the fact that it deserved a Nobel prize highlights the importance that causal inference has in other fields too and the contributions that those fields make.

Finally, a note on the present and the future of causal inference and its connection to artificial intelligence (AI). Many authors working in the field of AI have turned their attention to causal inference (Schölkopf et al., 2021; Berrevoets et al., 2024), some of them stating that it may play an important role in their field, potentially helping to overcome some of the current limitations of AI methods: poor performance with data drawn from a distribution different than the training set, sensitivity to spurious correlations, hallucinations and lack of factual base (particularly of large language models), among others.

2.2 Methods of this thesis and state of the art

In this section, we present the main methods employed in this thesis and the related state of the art. Any method included is relevant either because it has been directly used or because it provides the necessary context for another important method. This section also contains the state of the art of the clinical research questions addressed in this work, i.e., the latest knowledge about the topics behind those questions.

We divide this section into three subsections: Epidemiological study designs and statistical methods, causal inference methods, and clinical state-of-the-art. In the section on epidemiological study designs and statistical methods, we present the different *types* of clinical studies regarding their design, as well as the concept of evidence quality. We also introduce survival analysis, which is key for health studies. In the section on causal inference methods, we present the most relevant of them for this thesis in their most current versions. We divide this section into five parts: sub-frameworks and general methods of causal inference, causal inference methods with machine learning, causal machine learning methods (as it will be explained, they are not exactly the same), multivalued treatment settings, and algorithm evaluation. Finally, in the section on clinical state of the art, we present the latest findings about the effect of the COVID-19 vaccine on the menstrual cycle, the effects of the COVID-19 infection and vaccine on the risk of developing diabetes, and the effect of the usage of antibiotic-loaded bone cement during total knee arthroplasty on prosthetic survival.

2.2.1 Epidemiological study designs and statistical methods

As stated in the introduction of this section, the fields of epidemiology, public health, and clinical research have significantly contributed to the body of study designs and statistical methods for observational and experimental data. The most common and important study designs with observational data include case-control and cohort studies, while their experimental data counterpart includes randomized controlled trials. Furthermore, in a separate category that can combine both types of data, we find systematic reviews and meta-analyses.

On the one hand, case-control study design (Schulz and Grimes, 2002) works by selecting a group of patients based on the presence of an outcome of interest and other characteristics, forming the case group. Then, it selects another group of patients with similar characteristics but without the given outcome, forming the control group. Afterward, patients are re-grouped based on their treatment or exposure status (i.e., the treated/exposed and the untreated/unexposed), and finally, a statistical measure of the outcome is compared between groups. Additionally, usually, the researchers try to statistically remove the contribution of patients' characteristics and other variables to the outcome, which is known as *controlling* or *adjusting* for confounders. On the other hand, cohort study design (X. Wang and Kattan, 2020) selects patients based on treatment/exposure status (and other characteristics), then follows each patient in time to observe the production of an outcome of interest (or its absence) and then aims to statistically determine the contribution of the treatment/exposure and the characteristics of each patient to the production of that outcome. The case of randomized controlled trials (Stolberg, Norman, and Trop, 2004) is different, as those are always prospective studies and work by randomly assigning each member of a group of patients to a certain treatment strategy (in this case, we do not speak about exposures anymore, due to the experimental and interventional nature of RCTs), and then a measure of an outcome of interest is compared between groups. Because any factor other than the treatment is balanced across groups due to the random

Table 2.1: Basic epidemiological study designs. C-C: case-control; SR: systematic review; M-A: meta-analysis; Obs: observational; Exp: experimental.

	Obs. data	Exp. data	Controlling	Evidence quality
C-C	x		x	-
Cohort	x		x	+
RCT		x		++
SR with M-A	x	x	N/A	+++

assignment (up to random chance), the statistical differences in the outcome can only be caused by the treatment itself, i.e., controlling happens by design. Finally, a systematic review with meta-analysis (Tawfik et al., 2019) is a comprehensive summary of existing studies on a specific topic, where the systematic review part identifies and evaluates all relevant studies in a replicable way, and the meta-analysis part statistically combines their numerical results to provide a general, overall result. Table 2.1 shows a summary of the explained epidemiological designs, as well as some of their features. Note that this is a description of basic building blocks and that nuances arise in the details.

Note that the study designs that we just introduced are not the only available options. They are, in fact, the basic options, and their elements can be combined and/or modified into other alternative designs. An example of this is the self-controlled case series study (Petersen, Douglas, and Whitaker, 2016). In that design, the same patient or individual is both a case and a control, depending on the time and the exposure/treatment status. Thus, only individuals with the outcome of interest are selected; these individuals are cases when they are exposed or treated, and controls when they are unexposed or untreated (in either case, for a time window that is defined by the researchers depending on several factors such as the nature of the exposure/treatment). This design automatically accounts for time-invariant confounding, as each case-control pair is formed by the same individual with the same (time-invariant) characteristics.

Due to their nature, these different study designs have different likelihoods of providing biased estimations, and the concept of likelihood of bias is closely related to the concept of quality of evidence. Obviously, the higher the likelihood of bias, the lower the quality of the evidence. It is well established that the ascending order of the quality of evidence of the presented study designs is the following: case-control studies, cohort studies, randomized controlled trials, and systematic reviews with meta-analysis (Guyatt et al., 2011). This is why, for example, the evidence of a randomized controlled trial is considered, in general, of higher quality than the evidence of a cohort study. Note that, despite this, the risk of bias of an individual study does not only depend on its design but also on other characteristics of the study, such as the size of the analyzed sample.

To finish this section, we introduce a branch of statistics that is crucial for clinical studies: survival analysis (Clark et al., 2003; Bradburn et al., 2003). Survival analysis focuses on the expected time passed until an event of interest occurs, for example, the death of a patient, the progression of a disease, or the revision of a prosthesis. It also focuses on the factors contributing to that time. Some of the basic concepts of survival analysis include the survival and hazard functions, and their basic estimators are the Kaplan-Meier estimator and Cox's proportional hazards model. More advanced options include the weighted Kaplan-Meier estimator (Pepe and Fleming, 2018) and parametric estimators of the hazard function (Harrell, 2001). Finally, another key element of survival analysis is that of censoring, the phenomenon that occurs when only partial information is known about the outcome of a particular individual or patient.

After this overview of epidemiological study designs and related concepts, we speak about causal inference methods in the next section.

2.2.2 Causal inference methods

Subframeworks and general methods

For the purpose of reviewing the state of the art of this field, it is useful to think about causal inference as a meta-framework or a parent framework that contains other sub-frameworks and methods. The most common and relevant sub-frameworks are structural equation modeling, structural causal models, and potential outcomes. Only the latter two fall within the scope of this thesis, and we introduce them in the following paragraphs. After that, regarding specific methods, we discuss G-methods, cloning-censoring-weighting (CCW), and causal survival analysis.

Structural causal models (SCM) (Judea Pearl, 1995; Judea Pearl, 2009b) are mathematical representations of a system that describe the causal relationships among the variables of that system, giving rise to a structure that can be represented by means of a directed acyclic graph. Directed acyclic graphs are schemes formed by nodes and directed edges, with no closed loops between nodes. Those nodes represent variables, and the directed edges represent causal relationships. Figure 2.1 shows an example of a SCM. Together with SCMs, Judea Pearl and other authors also introduced the *do*-calculus (Judea Pearl, 2012). It is a set of rules forming an axiomatic system for replacing probability formulas that contain the *do* operator, which indicates intervention, with ordinary conditional probabilities. The combination of these two elements (DAGs and *do*-calculus) provides two crucial things: on one hand, it enables the identification of causal, interventional queries, i.e., the definition of the required assumptions for those queries to be computable; and on the other hand, it provides rules for their computation, i.e., their mathematical determination based on ordinary conditional probabilities.

The potential outcomes framework (or the Rubin causal model (D. Rubin, 1972; P. W. Holland, 1986; Imbens and D. B. Rubin, 2010)) sets causal inference problems in terms of potential outcomes, which are the outcomes that happen under specific interventions. For instance, in the case of a binary

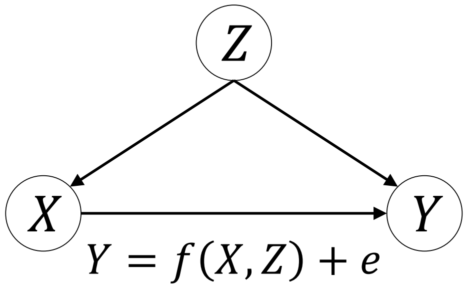


Figure 2.1: Example of a simple structural causal model, formed by a DAG and an equation, with an intervention X, a covariate or confounder Z, and an outcome Y (e is an error term).

intervention, there are two potential outcomes, one for each possible value of the intervention. The fundamental problem of causal inference is that, for each unit or patient, only one of the potential outcomes is observed, as each unit or patient can only effectively receive one of the two possible interventions. This framework gives rise to the concept of counterfactual, which is widely used in causal inference problems. Suppose we study the effect of an intervention that can take two possible values, A or B. A particular patient receives intervention A. The counterfactual outcome (or simply counterfactual) is the outcome that we would have observed for that patient had they received intervention B. Besides forming a framework, these notions help to understand the intuition behind concepts such as the propensity score, which is the probability each patient has of receiving each treatment option based on their characteristics (or covariates). Based on that, propensity score matching is a method that works by forming treated and untreated patients' pairs that have similar propensity scores (Rosenbaum and D. B. Rubin, 1983).

The aforementioned frameworks have elements in common, and, in general, concepts can be translated from one to the other, but it is worth mentioning that SCMs make causal query identification much easier than DAG-free Rubin causal models. In the following sections, we go one level down the frameworks and focus on specific causal inference methods: G-methods, cloning-censoring-weighting, and causal survival analysis.

G-methods

G-methods are a family of methods that come in handy when the problem at hand suffers from time-varying confounding. Recall that confounding occurs when a variable (usually named covariate or confounder) affects both the treatment and the outcome, and thus, time-varying confounding occurs when both the confounder and the treatment can vary over time. In such scenarios, basic adjusting strategies such as stratification fail. There are three main G-methods: the G-formula, already introduced in section 2.1, structural nested models with G-estimation (Vansteelandt and Joffe, 2014), and marginal structural models with inverse probability of treatment weighting (Stephen R. Cole and Miguel A. Hernán, 2008). The most relevant of the three for this thesis is the first one, the G-formula. Such a formula determines the conditional probability of an outcome given a particular intervention in terms of other conditional and marginal probabilities of the problem. In very simple cases, the probabilities required for computing the G-formula can be calculated nonparametrically, but as soon as the problem grows from a few binary and/or categorical covariates, modeling is required for estimating those probabilities, and clever implementation is required for computing the G-formula. Several implementation options have been proposed, each with different assumptions, pros, and cons. Wen et al. (2021) present an overview and comparison of the exiting alternatives: the iterative conditional expectation implementation (ICE), the non-iterative conditional expectation implementation (NICE), and the inverse probability weighting-based implementation. Finally, regarding the modeling part of probabilities and conditional expectations, usually linear, logistic, and/or generalized linear models are employed in the literature (McGrath, Lin, et al., 2020).

Cloning-censoring-weighting

Cloning-censoring-weighting (CCW) is a complementary method to target trial emulation (TTE), already introduced in section 2.1. CCW was presented for the first time by Cain et al. (2010), although not under that name, and further developed in other works such as the one by Gaber et al. (2024). This method aims to eliminate immortal time bias. Immortal time bias arises when, during the definition of the protocol of a target trial for its emulation, information from down the stream of time is used upstream: for instance, a patient that initiates treatment sometime after the beginning of the trial, which is a piece of information from down the stream of time, is assigned at time zero to its treatment line, which is a piece of information up in the stream of time. Between the assignation time and the treatment start time, such a patient will be, by definition, free of risk of the event of interest. This will not happen for a patient assigned to the no-treatment line, and this artificial distortion introduced by the design can bias the results. CCW aims to erase that distortion and its associated bias. The method consists of three steps: *cloning* individuals in the database, *censoring* those who deviate from the protocol of the target trial, and *weighting* the remaining ones with the inverse of their probability of not being censored, modeled as a function of patient's characteristics and time.

Causal survival analysis

Several works have given survival analysis, introduced in section 2.2.1, a causal perspective. Some examples are the work by J. Zhu and Gallego (2022), which uses the potential outcomes framework, and the work by Murray, Caniglia, and Petito (2021), which employs structural causal models. The work by Cui et al. (2023) is also relevant. We will explain this last work in depth in the next section, as it makes use of machine learning methods.

In fact, the connection between the fields of causal inference and machine learning has been established since the beginning of the current golden era of machine learning, partially because some of the breakthroughs in the field of causal inference came from authors working in computer science. Judea Pearl is the most relevant example of those authors. Since then, that connection has only continued to grow and expand. It is important to differentiate two aspects of this intersection of fields. On the one hand, we have works that focus on employing machine learning techniques for solving causal inference problems (we call that causal inference with machine learning). On the other hand, we have works that aim to develop what is known as causal machine learning. The differences between the two lie in their goals and approaches, and despite the fact that sometimes those differences can be subtle or diffuse, in general, this categorization helps to map the existing works and methods conceptually. For this thesis, the former category is more relevant than the latter. We present both categories in the following two sections.

Causal inference with machine learning

When dealing with causal inference problems, as soon as the data at hand is not extremely simple, nonparametric causal estimators suffer from the course of dimensionality and become useless: modeling is required. Among the available options for modeling, machine learning methods stand out for their power to approximate nonlinear functions and impose minimum assumptions on the data distributions. Thus, a line of work that has been especially fruitful in recent years is that of neural network-based causal inference methods. Several architectures of neural networks have been proposed for causal inference problems. Yuan, Ding, and Bar-Joseph (2020) propose a convolutional neural network for causal inference by devising a method to encode the observational data of a causal problem in an image-like matrix. Louizos et al. (2017) introduce a variational autoencoder architecture for the estimation of treatment effects at the patient level, assimilating proxies of unmeasured confounders to latent variables and exploiting the capabilities that autoencoders have with that type of variables. Yoon, Jordon, and Van Der Schaar (2018) use the generative adversarial network framework to learn counterfactuals of a causal inference problem. Shalit, Johansson, and Sontag (2017) use a rather simple architecture of few fully connected layers named TARNET (and a variation named CFR), but arranged in a clever way that optimizes the process for the task of causal inference. In particular, they propose a neural network that learns a representation of the covariates, and that has two different ”heads” or ends, one for each treatment option (given that the treatment is binary). The weights of each end are updated separately with each training data unit, ensuring that statistical power is shared in representation layers while the effect of treatment is preserved in the separate heads. In addition, another module of the network takes in the treatment value of each data unit, and the cost function adjusts for the bias introduced by treatment group imbalance during training by means of an integral probability metric defined by the authors. Taking inspiration from that work, Shi, Blei, and Veitch (2019) present another architecture named Dragonnet that further improves the result. The architecture includes another ”head” for learning the propensity score, and by defining the adequate loss function, it is ensured that the architecture exploits the sufficiency of the propensity score (Rosenbaum and D. B. Rubin, 1983) for adjustment.

The work by Shi, Blei, and Veitch (2019) not only provides a novel architecture but also a modification of the cost function that is inspired by the knowledge developed in two sub-fields of crucial importance for machine learning-based causal inference: semiparametric theory and the so-called double machine learning methodology. Semiparametric theory is, in this context, the statistical formalization of the estimation tasks present in a causal inference problem and the exploitation of theoretical and empirical knowledge about convergence, efficiency, bounds, etc., for that task. Of particular importance are the concepts of efficient influence curves, score functions, and estimating equations (Edward H. Kennedy, 2016). Double machine learning is a concept introduced by Chernozhukov, Chetverikov, Demirer, Duflo, C. Hansen, and Newey (2017) and Chernozhukov, Chetverikov, Demirer, Duflo, and Hansen (2018), and it also exploits knowledge about nonparametric and semiparametric estimation. Its goal is to develop a general framework for estimating causal effects using machine learning methods, providing confidence intervals for the estimates, and producing estimators with desirable statistical properties in terms of data efficiency and convergence. One of those desirable properties is double robustness. To explain this concept, note first that these causal inference methods work by modeling (with machine learning algorithms) two quantities of interest in the problem: the propensity score, which has been defined previously as the probability of getting the treatment given the covariates, and the conditional outcome, which is the expected value of the outcome given the covariates and the treatment. Double robustness ensures that the estimator of the causal effect of interest will converge to the correct value even if one of the two models of the aforementioned quantities is wrongly specified. This is achieved by means of orthogonalization, sample splitting, and cross-fitting (for a detailed explanation, see Chernozhukov, Chetverikov, Demirer, Duflo, C. Hansen, and Newey (2017). Finally, another method that is worth mentioning and that constitutes an alternative strategy for achieving doubly-robust estimators with nice asymptotic properties is that of targeted maximum likelihood estimation (TMLE) (Schuler and Rose, 2017). TMLE is a maximum-likelihood–based approach with two minimization steps, in which the second step optimizes the bias-variance trade-off for the target causal parameter of interest.

Finally, a subfield of work that is especially relevant for this thesis is the one combining causal survival analysis and random forests. The theoretical basis for this work was laid with Generalized Random Forests (Athey, Tibshirani, and Wager, 2019) and further developed to its most advanced version with Causal Survival Forests (Cui et al., 2023). Random forests are a flexible and data-adaptive machine learning algorithm that minimizes assumptions on the distributions of the modeled variables and delivers very good performance with tabular data with nonlinear relationships. Cui et al. (2023) employ random forests to model three key quantities present in a causal survival problem: the conditional outcome, the treatment propensity, and the censoring propensity. The former two have already been explained, and the latter is the probability of being censored, given the covariates. After modeling, these three quantities are combined into a final causal estimator that possesses desirable statistical properties.

In the next subsection, we look at the field of *causal machine learning*.

Causal machine learning

Causal machine learning is defined by Kaddour, Lynch, et al. (2022) as a collection of machine learning methods that take into account the data generating process (i.e., the real, physical underlying process that generated the observational data) and formalize it as a structural causal model. Thanks to taking this perspective, it is possible to use the models to simulate the effects of interventions by generating counterfactuals. More importantly, causal deep learning is defined by Mihaela Van der Schaar and other authors in a series of works (some of the most relevant ones being those by Balagopalan et al. (2024) and Feuerriegel et al. (2024)), as the intersection between causal inference and deep learning. Instead of limiting the efforts to simply *doing causal inference with deep learning models*, causal deep learning aims at achieving a true symbiosis between both fields, such that the whole is greater than the sum of the parts. In particular, one of the main goals of this line of work is to find practically applicable methods for healthcare problems that can relax some of the assumptions associated to causal inference strategies, which are often strong. This type of effort is crucial for the successful adoption and usage of causal inference methods in real-world problems.

So far, most of the discussed algorithms and methods that have been presented belong to scenarios with binary treatment. Nevertheless, often, real-world applications have multivalued treatments. In the next section, we provide an introduction to multivalued treatment settings.

Multivalued treatment settings

Multivalued treatment settings are those in which the treatment or intervention of interest is not binary but categorical, with more than two possible values. Because of the intrinsically more complex nature of such settings in comparison with binary treatments, methods for multivalued treatments have traditionally been somewhat neglected in the literature. Nevertheless, many real-world problems have multivalued treatments or could benefit from a multivalued treatment-based approach. One of the authors who has more thoroughly studied these settings is Mattias D. Cattaneo (Cattaneo, 2010; Cattaneo, Drukker, and A. D. Holland, 2013). In his works, the author not only lays the formal mathematical definitions of causal quantities in multivalued settings and provides efficient estimators for them but also demonstrates some key findings. For instance, the joint estimation of multivalued causal effects is necessary for correct statistical inference, as opposed to estimating each effect separately. Other relevant works proposing algorithms for multivalued treatment settings are those by Kaddour, Y. Zhu, et al. (2021), who use neural networks and focus on *structured treatments* such as graphs, images, texts, etc., or the work by Schwab, Linhardt, and Karlen (2019), who use support vector machines, or the work by Künzel et al. (2019), who use *meta-learners*, i.e., aggregators that combine the outputs of individual algorithms.

Finally, in the next part, we bring attention to a topic that is cross-sectional to all previously presented methods and algorithms: the evaluation of algorithmic performance.

Algorithm evaluation

Any of the presented algorithms so far can and should be subject to the evaluation of its performance. For that, a metric or a set of metrics is required. Such metric is usually a measure of the deviation between the estimation of a given effect provided by the algorithm under evaluation and the *ground truth* effect. The philosophy, thus, is very similar to the one used with machine learning algorithms. As an example, in an image recognition task, the training and evaluation of a neural network is done based on the ground truth information, which is the label for each image provided, usually by one or (some consensus of) several humans. Causal inference, nevertheless, presents a specific challenge: in general, when using real-world data, the ground truth of the effect of interest is not available for the evaluator by any means. This is a direct consequence of the aforementioned fundamental problem of causal inference, which refers to the fact that for each individual or patient, we do not get to observe one or some of the potential outcomes, the counterfactual(s). Thus, usually, synthetic or semi-synthetic data is required for the evaluation of causal inference algorithms, where the data-generating process is fully or at least partially under researchers' control, and they can simulate the required information to calculate ground truth effects. The development of data-generating processes for algorithm evaluation is carried out mostly *ad hoc* in the literature, although some datasets have been established as *de facto* benchmarks for comparisons. Different works employ data of different nature, and thus we have examples using **real data** (Brost, Mehrotra, and Jehan, 2020; Schnabel et al., 2016; Linden and Yarnold, 2016; Uysal, 2015; Hong, 2012; Esposti, 2017), **semi-synthetic data** (Kuang et al., 2021; Bica, Jordon, and Schaar, 2020; Schwab, Linhardt, Bauer, et al., 2020; Franklin et al., 2014; Kaddour, Y. Zhu, et al., 2021), or **synthetic data** (Lopez and Gutman, 2017; Austin, 2018; Y.-Y. Lee, 2018; Garrido, Lum, and Pizer, 2021; Graham and Pinto, 2022; A. Li and Judea Pearl, 2022). Regarding these types of data, there are two properties at trade-off: realism and control over the data-generating process (DGP). With real data, realism is maximized, but we have no control over the DGP (and most times, we do not even have access to the ground truth effects); with fully synthetic data, we have full control over the DGP, but at the cost of realism. Due to this, most of the works in the literature choose the semi-synthetic data option. But this option is not free of limitations, as some authors have indicated (Curth, Svensson, and Weatherall, 2021). A clever way to overcome or at least minimize these limitations has been presented by Neal, Huang, and Raghupathi (2021). This work proposes a method to fit models to existing, real data and then use those models to generate synthetic but *a priori* realistic data. The authors also show that the distributions of the generated data are statistically indistinguishable from those of the real data.

After having presented the methods and state of the art of this thesis, we introduce now the clinical state of the art.

2.2.3 Clinical state of the art

In this section, we present the state of the art of the clinical aspects of this thesis. In particular, we divide the section into three parts: the first about the side effects of the COVID-19 vaccine on the menstrual cycle, the second about the effect of the COVID-19 infection and vaccine on the risk of developing diabetes, and the last about the effect of the usage of antibiotic-loaded bone cement in knee prosthesis survival.

Effect of the vaccine against COVID-19 on the menstrual cycle

The COVID-19 epidemic started with an outbreak in December 2019 in China and was declared a pandemic by the World Health Organization (WHO) the 11^{th} of March 2020. It constituted one of the biggest health threats that humanity has faced in modern times. The first vaccines against COVID-19 were developed in 2020 and were authorized for their administration to the general population by the end of that year at an unprecedented speed. Nowadays, three and a half years after vaccination campaigns started worldwide, and with more than 13.000 million doses administered, there is solid evidence showing that the rate and type of side effects of the vaccine are lower risk than the COVID-19 infection itself (Wise, 2024; Amer et al., 2024): COVID-19 vaccines are, in general, safe and worth getting.

Most typical side effects are mild and do not require specialized medical care. Those include soreness, redness or inflammation of the vaccination site, fatigue, headache, or myalgia. In addition, potential effects on other health outcomes have also been studied, one of them being the menstrual cycle. This is relevant, among other reasons, because the characteristics of the menstrual cycle are important indicators of the reproductive and overall health of menstruating people. Edelman, Boniface, Benhar, et al. (2022) found an association between receiving the COVID-19 vaccine and experiencing a small and temporary increase in the length of the menstrual cycle, with the effect being most noticeable in the cycle immediately following vaccination. In a later study, Edelman, Boniface, Male, et al. (2024) also found that the phase of the menstrual cycle at which the vaccine was administered, i.e., the menstrual cycle timing, was associated with the presence or absence of the previously described cycle changes. Ramaiyer et al. (2024) found similar results using data collected from a period-tracking app. Finally, two systematic reviews about the topic (Nazir et al., 2022; Smaardijk et al., 2024) found aggregated evidence of alterations of the menstrual cycle associated with the COVID-19 vaccine, although those were mild and did not last over time.

Effects of the COVID-19 infection and vaccine on diabetes

In this subsection, we explore the evidence about a potential beneficial side effect of the COVID-19 vaccine: the protection it provides against the increased risk of diabetes due to COVID-19 infection.

Several studies in the literature have shown that the incidence of diabetes onset increased during the COVID-19 pandemic. The work by Xie and Al-Aly (2022) found an increased risk of incident diabetes in a group of patients who had COVID-19 in comparison with control groups. Another study by Wander et al. (2022) also found an association between COVID-19 infection and a higher risk of incident diabetes, although only in men. Another work by Salmi et al. (2022) reported that more children with type 1 diabetes had severe diabetic ketoacidosis (DKA) at diagnosis during the pandemic, but authors hypothesized that it was not a consequence of COVID-19 infection itself but of delays in diagnosis.

Given these discoveries, it was then natural to wonder about the effect of the vaccine on the relationship between COVID-19 and diabetes. Taylor et al. (2024) looked at the association between COVID-19 and the incidence of any type of diabetes and the effect of COVID-19 vaccination on that association. They found that elevated incidence of type 2 diabetes after COVID-19 was less apparent in people who had been vaccinated. Similarly, another work by Kwan et al. (2023) reported that diabetes risk after COVID-19 infection was higher in unvaccinated patients. Xiong et al. (2023) evaluated the risk of diabetes following different COVID-19 vaccines and SARS-CoV-2 infections and found no increased risk of diabetes post-vaccination but a higher risk of type 2 diabetes following infection (especially with the Omicron variant). Finally, two systematic reviews about the topic have been published: Alsudais et al. (2023) only considered type 1 diabetes and had a very low number of patients, and He et al. (2023) concluded that the complex relationship between vaccination and diabetes had a bidirectional effect: vaccination could contribute to the risk of worsening blood glucose in diabetic patients, and diabetic patients could have a lower antibody response after vaccination than the general population.

In the next subsection, the last of this chapter, we jump to the next health technology and speak about the state-of-the-art knowledge about the relationship between antibiotic-loaded bone cement and knee prosthesis survival.

Effect of the usage of antibiotic-loaded bone cement during total knee arthroplasty on the survival of the prosthesis

In these lines, we discuss the state-of-the-art knowledge of the effect of antibiotic-loaded bone cement on knee prosthesis survival after total knee arthroplasty. To do so, we first provide some context about the interest in this topic. Peri-prosthetic joint infection is a major complication of total knee arthroplasty surgery (TKA) and happens in between 1% and 2% of the cases. It usually requires revision surgery, and this has a big impact on patients' life quality and satisfaction (Rachel Frank, Michael Cross, and Craig Della Valle, 2014). For this reason, surgeons and researchers have been looking for interventions that decrease the incidence of peri-prosthetic joint infection. In the case of total *hip* arthroplasty surgery, an example of such intervention is the addition of antibiotics to the bone cement employed to fix the prosthesis to the patient's bone. But in the case of total *knee* arthroplasty, the evidence about the benefits of this intervention is inconclusive. Some studies show results in favor of it, others against it, and yet others show no effect at all. Thus, this topic remains an open question in the specialized literature, and studies at all levels of evidence quality (observational, randomized, and systematic reviews with meta-analysis) are still being conducted and published.

Jameson et al. (2019) analyzed 731.214 prostheses and found a lower risk of prosthetic revision surgery among prostheses with antibiotic-loaded bone cement (ALBC), in comparison with prostheses with plain cement, after adjusting for other variables. That work also found no evidence of prosthetic mechanical problems induced by the antibiotic, a hypothesis that some authors have proposed in other works. Similarly, Bendich et al. (2020) found lower rates of revision in the ALBC group and a protective effect of ALBC against revision in a multivariate analysis. Another study finding evidence in favor of ALBC was that by Randelli et al. (2010). On the contrary, among individual studies that report no effect or a harmful effect of ALBC, we find those by Bohm et al. (2014), Namba, Chen, et al. (2009), Namba, Inacio, and Paxton (2013) and Hinarejos et al. (2013).

Furthermore, we find several systematic reviews with meta-analyses about this topic in the literature. In general, the works by King et al. (2018), T. Leta et al. (2024), and H.-Q. Li et al. (2022) showed no statistically significant differences in infection rates between ALBC and plain cement groups. Nevertheless, it is worth mentioning that T. Leta et al. (2024) reported that four out of the nine included databases showed results in favor of ALBC, and H.-Q. Li et al. (2022) reported that the two largest included studies reached the same conclusion.

Finally, note that patients' preoperative characteristics can also be associated with the risk of developing peri-prosthetic joint infection. In general, there is consensus in the literature about the fact that gender (being male), age (being older), having previous comorbidities such as diabetes or rheumatoid arthritis, and habits like smoking or alcohol abuse are all risk factors for prosthetic infection (Kurtz et al., 2010; Jämsen et al., 2009; Namba, Inacio, and Paxton, 2013; Resende et al., 2021).

2.3 Conclusions

In this chapter, we have reviewed the methods that are more relevant to this thesis and their state of the art. First, we have provided an overview of some historical, crucial methodological contributions to the field of causal inference. Then, we have gone through epidemiological study designs and correlational approaches to observational data analysis and statistical methods. Afterward, we focused on causal inference methods. And finally, we have provided context and state-of-the-art knowledge about the clinical aspects of this thesis.

In the next chapter, we delve into the application of these methods to the first analyzed real-world use case of this thesis: the effect of the COVID-19 vaccine on the menstrual cycle.

Chapter 3

Effects of the vaccine against COVID-19 and its administration time on the menstrual cycle

In this chapter, we present an answer to research question **Q1**, *Do the vaccine against COVID-19 and the vaccination time have any effect on the menstrual cycle?*

The chapter is a longer version of the journal article ”Borja Velasco-Regulez, Jose L. Fernandez-Marquez, Nerea Luqui, Jesus Cerquides, Josep Analia Fukelman, & Josep Perelló (2022). Is the phase of the menstrual cycle relevant when getting the covid-19 vaccine? *American Journal of Obstetrics and Gynecology*, 227, 913-915. DOI: 10.1016/j.ajog.2022.07.052.”

In addition, we also provide an answer to research subquestion **Q4a** *What are the advantages and disadvantages, strengths and weaknesses, of correlational methods for generating evidence about clinical interventions?* basing our analysis in the use case of **Q1**.

3.1 Background

During the development of a new drug or vaccine, usually the first pieces of evidence about its security are generated through experiments. Then, randomized controlled trials for assessing effectiveness are performed (Spieth et al., 2016). The assessment of the effectiveness requires monitoring one or several outcomes of interest. In the particular case of COVID-19 vaccines, those outcomes could be the number of infected patients or the severity of the infections. Besides these outcomes, which are directly related to the main goal of the vaccine (protecting from COVID-19), it is also usual to look at other indicators, normally with the intention of discovering *side effects* (i.e., effects other than the main or desired ones). Thus, usually any abnormal health symptom will be detected and reported during the trials. Nevertheless, it is not possible to analyze every single health variable (which is not even a defined or closed category) or to extend the trial indefinitely. So, after trials end, vaccines are approved, administration starts among the general population, and monitorization of side effects continues through pharmacovigilance protocols.

In the case of the COVID-19 vaccines, the most typically reported side effects were mild and did not require specialized medical care. Those included soreness, redness or inflammation of the vaccination site, fatigue, headache, or myalgia. Some specific formulations of the vaccine were also associated with more serious side effects, in particular with blood clots (Zhao et al., 2024). Nevertheless, three and a half years after vaccination campaigns started worldwide, and with more than 13.000 million doses administered, the accumulated evidence shows that the rate and type of side effects of the vaccine are lower risk than the infection itself (Wise, 2024; Amer et al., 2024): COVID-19 vaccines are, in general, safe and worth to get.

During the first half of 2022, public attention was drawn to possible side effects of the COVID-19 vaccine on the menstrual cycle, mostly through anecdotal evidence in the form of patient reports shared on social media. Characteristics such as the stability of the menstrual cycle are important indicators of the reproductive and overall health of menstruating people, as alterations can affect physical, emotional, sexual, and social aspects of their lives (Critchley et al., 2020). Thus, this topic became relevant among clinicians and researchers, and formal studies were planned and conducted. The way this topic emerged and gained attention is connected to the concept of *citizen science*, which is relevant to the current chapter of this thesis and is explained in the next paragraph.

In the field of healthcare, the process of posing research questions and collecting data for analysis is carried out most times by health researchers and/or clinicians. Nevertheless, occasionally, this process can also be led by non-professional, volunteer citizens, giving rise to what is known as citizen science. The concept of citizen science was first defined in the mid-1990s by Rick Bonney and Alan Irwin (Riesch and Potter, 2014; Irwin, 2024) and can be summarized as the collaborative process by which members of the general public engage in scientific research, usually in partnership with professional scientists. In the field of healthcare in particular, this process can be oriented towards disease prevention, community engagement in public health, and health promotion in general (Vohland et al., 2021; Laird et al., 2023). Citizen science can help put the scientific focus on topics that have been otherwise neglected, collecting big amounts of data that would otherwise be difficult or expensive to collect and increasing the variety in the data, reducing potential biases. This process is further enhanced by means of technologies such as social networks and smartphone applications. The work of this chapter is a successful example of citizen science. Firstly, as we already mentioned, the attention towards the topic (COVID-19 vaccine and menstrual cycle) was originally drawn by citizens instead of professional researchers. Secondly, for developing the work of this chapter, we employed data from a menstrual cycle tracking smartphone application, Lunar App (APP Lunar 2024), used by individuals for information purposes and without a primary focus on research. The work was a successful collaboration between different institutions with different domains of expertise: Members of the Lunar App team as the data providers, gynecologists from the *Hospital de la Santa Creu i Sant Pau* as the health experts, a member of the University of Geneva as a citizen science expert, and finally members from the IIIA (Artificial Intelligence Research Institute) and AQuAS (Agency for Health Quality and Assessment of Catalonia) as the data analysis experts. The author of this thesis acted as the coordinator of the project.

Some of the first published scientific studies about the relationship between the COVID-19 vaccine and the menstrual cycle provided evidence of alterations in the first cycle after vaccination (Edelman, Boniface, Benhar, et al., 2022; Nazir et al., 2022). It must be mentioned, though, that those alterations were minor, mostly in the form of increased cycle length, and tended to disappear at later cycles. Gynecologists and other experts wondered then about the effect of vaccination *timing* on those alterations, i.e., whether the moment of getting the vaccine with respect to the menstrual cycle (which has different phases, as we will explain soon) would have any effect on the observed alterations. This is exactly the research question that we aim to answer in this chapter, **Q1**, *Do the vaccine against COVID-19 and the vaccination time have any effect on the menstrual cycle?* Note that the average menstrual cycle lasts 28 days and that any cycle between 21 and 35 days is considered clinically normal (Critchley et al., 2020; Mihm, Gangooly, and Muttukrishna, 2011). Also, note that the menstrual cycle consists of two phases: the follicular phase (between days 1 and 14 in a 28-day cycle) and the luteal phase (between days 14 and 28 in a 28-day cycle). Ovulation occurs between both phases. For answering research question **Q1** we will employ the epidemiological study design of self-controlled case series (SCCS) (Petersen, Douglas, and Whitaker, 2016), and correlational statistical analysis methods that we will detail in the next section. This use case will allow us to answer research subquestion **Q4a**, *What are the advantages and disadvantages, strengths and weaknesses, of correlational methods for generating evidence about clinical interventions?*

After this introduction, the remainder of this chapter is divided into the sections of data and methods, results, and discussion and conclusions.

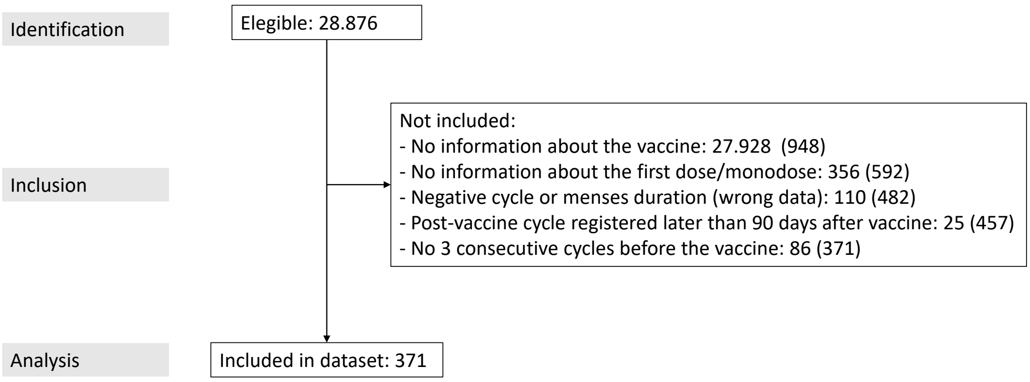

Identification
Elegible: 28.876
Not included:
- No information about the vaccine: 27.928 (948)
- No information about the first dose/monodose: 356 (592)
- Negative cycle or menses duration (wrong data): 110 (482)
- Post-vaccine cycle registered later than 90 days after vaccine: 25 (457)
- No 3 consecutive cycles before the vaccine: 86 (371)
Inclusion
Analysis
Included in dataset: 371

Figure 3.1: STROBE flow diagram of the database filtering process. (STROBE: Strengthening the Reporting of Observational studies in Epidemiology).

3.2 Data and methods

We analyzed data collected by the menstrual cycle tracking smartphone application Lunar App. This application allows users to track their menstrual cycle and menstruation (also called menses). In particular, it allows the recording of beginning and end dates and storing the pain intensity and the blood loss quantity during menstruation (both aspects recorded as *less*, *equal*, or *more than usual*). Finally, it also allows for storing the COVID-19 vaccination status of the users.

The database we analyzed contained 28,876 users and 162,529 cycles. We filtered the data, keeping only those users who had reported their vaccination status and at least five consecutive cycles. We considered the first doses (or monodoses) of the vaccine for the analysis, and we removed incomplete and/or wrong data. After this filtering process, we ended up with 371 users and 1855 cycles, registered between September 2020 and February 2022. Figure 3.1 shows the STROBE Strengthening the reporting of observational studies in epidemiology) (Elm et al., 2008) diagram of the filtering process, with the details. The relatively small size of the final sample was caused by the restrictive inclusion and exclusion criteria, imposed to ensure the maximum attainable data quality.

The intervention of interest was binary: getting the COVID-19 vaccine during the follicular phase, or getting it during the luteal phase. The luteal phase was defined as the period between the beginning of menstruation and the 14 days prior to it, due to the relative robustness of that phase. The rest of the cycle was considered to be the follicular phase. The primary outcome was menstrual cycle length *change* in days. Secondary outcomes were menses length change in days, and variations in the usual blood quantity and pain intensity during the menses. Users reported abnormalities when they had more or less blood loss quantity or pain intensity than usual during menses.

The self-controlled case series design was employed for analysis. Recall that in this design, each participant is a control before the intervention of interest, and a case after. This design automatically controls for time-invariant confounders, as each case-control pair is formed by the same patient with the same baseline covariates. No other covariates were included in the analysis, as they were not available in the database or they did not contain information before and after the intervention.

For calculating the menstrual cycle length change, we computed, for each user, the difference between the median length of the three cycles before the vaccine, and the length in which the vaccine was given (4th cycle). Then, we computed the median over all the users, as well as the 95% confidence intervals of the point estimate. We used medians because the data was not normally distributed. We proceeded identically for the menses length, but employing data from the 5th cycle. For the blood loss quantity and pain intensity, we computed the differences in the percentages of cycles with abnormalities in each endpoint before and after the vaccine, and the 95% confidence intervals of the point estimates. Finally, for effectively analyzing the intervention of interest, we stratified the analysis of all outcomes by the phase of the menstrual cycle of the user at vaccination time (note that we also provide results of the overall, unstratified data for reference). We employed Wilcoxon signed-rank and Chi-squared tests for statistical hypothesis testing of medians and proportions, respectively. Statistical significance was set for a p-value smaller than 0.005.

3.3 Results

First, we present some informative results that provide context of the morphology of the database. The distribution of percentages of users' age range (of the final dataset for analysis) was the following: 11.85% between 18 and 24 years; 49.15% between 25 and 34; 28.56% between 35 and 44; 8.31% between 45 and 54; 2.13% others. The frequency of each vaccine brand identifier was the following: Sinopharm BIBP, 85; Oxford–AstraZeneca (Covishield), 102; Sputnik V, 62; Pfizer–BioNTech (Comirnaty), 84; Moderna (Spikevax), 17; Janssen (Johnson & Johnson), 7; others, 14. The distribution of the medians of cycle lengths of each user before the vaccine had a median value of 28 days, with a (5, 95) inter-percentile range of (22, 34) days, indicating that the cycles of the sample were relatively stable.

In the overall, unstratified analysis, we observed a statistically significant increase in the median cycle length of 0.5 days (confidence intervals: (0.0–1.0)) for all individuals, as it can be seen in Figure 3.2 (left). We also observed that 8.08% of the individuals had an increase of 8 or more days of the cycle length, which is considered clinically significant (Mihm, Gangooly, and Muttukrishna, 2011) (see Table 3.2). We observed no variation in menses length, which is in line with results previously reported in the literature (Edelman, Boniface, Benhar, et al., 2022). In addition, we

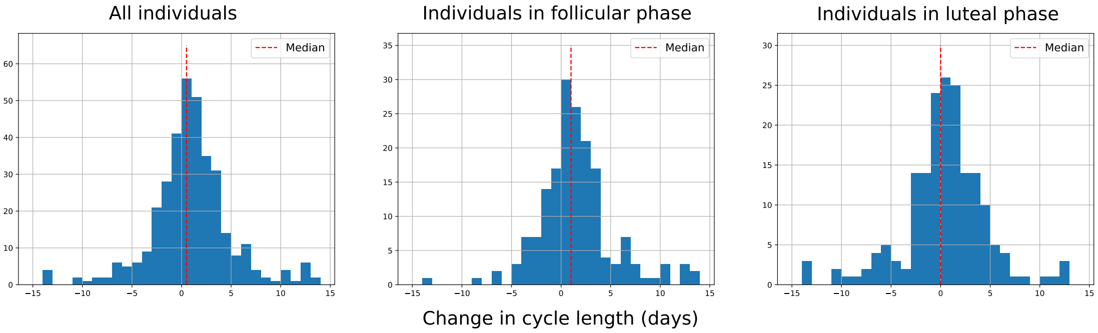


Figure 3.2: Histograms of the differences between the median length of the three cycles before the vaccine and the length of the cycle of the vaccine, for each user. The median value is depicted in each plot. Left: all individuals; Center: individuals vaccinated during the follicular phase; Right: individuals vaccinated during the luteal phase.

observed no significant variations in the percentages of cycles with abnormal blood loss or pain intensity. All the results can be seen in Table 3.1.

The stratified analysis showed an association between the phase of the menstrual cycle at vaccination time and the cycle length change. Thus, individuals vaccinated during follicular phase showed a statistically significant median increase cycle length of 1 (0.0, 1.0) day (see Figure 3.2, center), with 11.82% of the users having an increase of 8 or more days (Table 3.2), while individuals vaccinated during luteal phase showed no change. These results can be seen in Table 3.1.

Table 3.1: Information about the COVID-19 vaccine association with menstrual cycle disorders. * for a p-value$<$0.005.

	All vaccinated individuals		Individuals vaccinated during follicular phase (186; 50.13%)		Individuals vaccinated during luteal phase (185; 49.87%)	
	Change	P-val	Change	P-val	Change	P-val
Cycle length	0.5 (0, 1)	*	1 (0, 1)	*	0 (0, 1)	0.96
Menses length	0 (0, 0)	0.01	0 (0, 0)	0.1	0 (0, 0)	0.05
Percentage of cycles with abnormal blood loss during menses	-2.9 (-7.8, 2)	0.15	-3.8 (-10.9, 3.4)	0.2	-1.9 (-8.6, 4.6)	0.46
Percentage of cycles with abnormal pain intensity during menses	-0.5 (-5.7, 4.8)	0.83	-1.1 (-8.7, 6.6)	0.72	0.2 (-7, 7.4)	0.95

3.4 Discussion and conclusions

Our analysis showed an association between the vaccine and an increase in cycle length, similar to what some other previous and also posterior studies have shown, including meta-analyses (Edelman,

Table 3.2: Distribution of percentages of users with different cycle length increases after vaccination.

	All users	Users vaccinated during follicular phase	Users vaccinated during luteal phase
Increase <=0 days	49.86%	44.08%	55.67%
Increase in (0,2] days	23.71%	25.80%	21.62%
Increase in (2,8) days	18.32%	18.27%	18.37%
Increase >=8 days	8.08%	11.82%	4.32%

Boniface, Benhar, et al., 2022; Nazir et al., 2022; Smaardijk et al., 2024).

Besides that, our results also showed an association between the phase of the menstrual cycle at vaccination time and the change in cycle length. Thus, vaccination during the luteal phase had a *protective effect* over COVID-19 vaccine-related menstrual cycle disorders, compared to vaccination during the follicular phase. This suggests considering the phase of the menstrual cycle for the design of COVID-19 vaccination policies, recommending vaccination during the luteal phase to menstruating individuals. These conclusions were novel in the literature when we published them (Velasco-Regulez et al., 2022), and were replicated by a larger study later (Edelman, Boniface, Male, et al., 2024). It is also worth mentioning that our study was included in a posterior meta-analysis (Smaardijk et al., 2024), and that the work by Ramaiyer et al., 2024, which was published later and reached conclusions similar to ours, also employed data from a menstrual cycle tracking app.

3.4.1 Critical assessment of the employed methodology

The first research question of this chapter, **Q1**, (*Do the vaccine against COVID-19 and the vaccination time have any effect on the menstrual cycle?*) has been answered with observational data, with the SCCS epidemiological study design, and with correlational-only statistical methods (hypothesis testing). Causality has not been explicitly addressed and no causal inference method has been employed. In the following paragraphs, we indicate the main advantages and strengths, as well as disadvantages and weaknesses of this approach, with the goal of answering research subquestion **Q4a**, (*What are the advantages and disadvantages, strengths and weaknesses, of correlational methods for generating evidence about clinical interventions?*)

Among advantages and strengths we find the simplicity of this method. It allowed us to answer an important question about a complex topic with a database that contained limited information. In such a scenario, other more powerful but more complex approaches could have failed due to insufficient data. The employed method is able to remove potential confounding from any variable that is time-invariant in a straightforward manner.

Among disadvantages and weaknesses, on the one hand, we find that some of the assumptions of the employed epidemiological design may be unrealistic, in particular, the assumption that no characteristic of the patient changes with time or between before and after vaccination. On the other hand, the most crucial weakness is that the employed correlational-only approach provides no guarantee of the causal nature of the measured association, i.e., it provides no guarantee that our intervention of interest (vaccine administration timing) is actually the cause of the observed changes in the outcome (menstrual cycle length variation). This means that, technically, it would be correct for us to say that the observed correlation may be due to pure chance, or that it may be caused by a common variable that we are not aware of. The problem is that no correlational-only, observational study ever concludes that, because that would render the study partially useless: causality is *implicit*, despite the disclaimer authors usually make about the correlational nature of the conclusions. Gershman and Ullman (2023) provide evidence supporting the hypothesis that people do infer causality from statements of association, under minimal conditions. Thus, this methodology presents serious limitations that hinder the reached conclusions.

In the next chapter, we continue analyzing the same health technology, the COVID-19 vaccine, but another outcome of interest, diabetes onset. We employ a causal approach for the analysis.

Chapter 4

The effect of covid-19 vaccine on the risk of diabetes onset

In this chapter, we present an answer to research question **Q2**, *Does the vaccine against COVID-19 have any effect on the risk of developing diabetes?*

The chapter presents a causal analysis of the question of interest, employing an integral causal approach with some of the most complete and advanced methods at our disposal. This work is still ongoing, and we intend to publish it when it is finished in a high impact journal.

In addition, we also provide an answer to research subquestion **Q4b** (*What are the advantages and disadvantages, strengths and weaknesses, of causal inference methods for generating evidence about clinical interventions?*) basing our analysis in the use case of **Q2**.

4.1 Background

Approximately two years after the COVID-19 pandemic started, some studies in the literature began to provide evidence that diabetes onset had increased during that period. Xie and Al-Aly (2022) found increased risk and 12-month prevalence of incident diabetes in a group of patients who had COVID-19, in comparison with two control groups without the infection. Wander et al. (2022) also found an association between COVID-19 infection and a higher risk of incident diabetes, although only in men. And Salmi et al. (2022) reported that more children with new-onset type 1 diabetes were diagnosed with severe diabetic ketoacidosis at admission to pediatric intensive care units during the pandemic. Knowing that viral infections can trigger type 1 diabetes (Rajsfus, Mohana-Borges, and Allonso, 2023), the authors of that work aimed to determine whether COVID-19 infection had a direct role in such increase (i.e., whether the infection was the *cause*). They concluded that the observed increase was probably due to ”delays in diagnosis following changes in parental behavior and healthcare accessibility.” Another (non-peer-reviewed) report from the Health Quality and Assessment Agency of Catalonia, co-authored by the author of this thesis (Troncoso et al., 2022), described that in 2020, no increase in type 1 diabetes incidence was observed, but 2021 witnessed a 28% increment. Nevertheless, it concluded that it was not possible to establish a causal relationship between the pandemic or the infection and the aforementioned increase, neither through biological nor through *social* mechanisms (such as the disruption of normal healthcare assistance), and that further research in the topic was warranted.

Given these discoveries, it was then natural to wonder about the effect of the COVID-19 vaccine on the risk of diabetes. Studies researching this topic have been published only in the last two years, 2023 and 2024, given the novelty of the issue. Taylor et al. (2024) investigated the association between COVID-19 and the incidence of any type of diabetes and the effect of COVID-19 vaccination in such association. The authors of that study found that elevated incidence of type 2 diabetes after COVID-19 was less apparent in people who had been vaccinated. Kwan et al. (2023) reported that diabetes risk after COVID-19 infection was higher in patients who were not vaccinated than in those who were, also suggesting a beneficial protective effect of the vaccine. They proposed a possible pathway for the explanation of the higher diabetes risk after infection (”inflammation contributing to insulin resistance”) but mentioned that ”additional studies are needed to understand cardio-metabolic sequels of COVID-19 and whether COVID-19 vaccination attenuates the risk of cardio-metabolic diseases”. Finally, a systematic review by He et al. (2023) analyzed studies discussing the effect of diabetes on vaccination and the effect of vaccination on diabetes. The authors concluded the existence of a complex relationship with a bidirectional association: vaccination could contribute to the risk of worsening blood glucose in diabetic patients, and diabetes could induce a lower antibody response after the vaccine. All in all, further evidence was required.

None of the aforementioned studies nor any other found in the literature, to the best of our knowledge, employed causal inference for the question of interest. Kwan et al. (2023) used a self-controlled exposure-crossover design, Xiong et al. (2023) a case-control design, and Taylor et al. (2024) a cohort design, and all of them employed causality-free, correlational-only methods. On the contrary, we posed a precise causal question, *Does the vaccine against COVID-19 have any effect on the risk of developing diabetes?* (research question **Q2**), and we aimed to answer it using a causal approach and causal inference methods. In particular, we developed a Directed Acyclic Graph (DAG) of the involved variables, we employed the target trial emulation framework, together with Cloning-Censoring-Weighting (CCW), and we adapted a machine learning-based, non-iterative conditional expectation (NICE) implementation of the G-formula, by modifying a previously existing algorithm. The reason for employing the G-formula was that time-varying confounding was present in our problem. Then, using this study as a base, we aimed at answering research subquestion **Q4b**, *What are the advantages and disadvantages, strengths and weaknesses, of causal inference methods for generating evidence about clinical interventions?*

After this introduction, the rest of this chapter is divided into the sections of methods and data, results, and discussion and conclusions. The work presented in this chapter was conducted in collaboration with UMIT TIROL University, in particular with the Public Health, Health Services Research and Health Technology Assessment department.

4.2 Methods and data

4.2.1 Methods

In this section, we describe the methodology employed for the current study. First, we start by introducing the PICO framework and question (Hosseini et al., 2024), which is a framework for formally and systematically posing research questions employed by the evidence-based medicine (EBM) approach. EBM has been defined as the systematic approach to clinical problem solving that takes into account the best available research evidence (Akobeng, 2005). Afterward, we define the protocol of the target trial, i.e., the trial that we would have ideally conducted. Recall that target trial emulation (TTE) (Miguel A. Hernán and James M. Robins, 2016) was introduced in the state of the art of Chapter 2. Then, we define the directed acyclic graph (DAG) of the problem, with the involved variables and their causal relationships. And finally, we introduce the causal quantities and effects of interest, providing the estimators and algorithms employed for computing them.

PICO question

PICO is a framework for formally posing research questions that is employed by the evidence-based medicine (EBM) approach. PICO stands for population, interventions, comparisons, and outcomes, and requires researchers to clearly determine those four components in their study. In our case, the *population* of the study were all the citizens of Catalonia covered by the public healthcare system's service, around 8 million individuals. The *interventions* of interest were defined by the number of received doses of the COVID-19 vaccine, and in particular, four interventions were possible: not receiving any dose (0 doses), receiving 1 dose, 2 doses, or 3 doses. The outcome of interest was new-onset diabetes throughout the study period, identified by the presence of a first diabetes diagnostic code during such time. We employed the International Classification of Diseases (ICD) (Organization, 2004), revisions 9 and 10, for diagnostic codes. In particular, the considered ICD-9 codes were those starting with '250' ('diabetes mellitus'), and the ICD-10 codes were those starting with 'E10' ('Type 1 diabetes mellitus'), 'E11' ('Type 2 diabetes mellitus') or 'E13' ('Other specified diabetes mellitus'). We considered type 1, type 2, and unspecified diabetes for remaining agnostic about the potential mechanisms of influence of the COVID-19 infection and vaccine on the disease.

Target trial protocol, with CCW

We employed the target trial emulation method, introduced in Chapter 2, following the guidelines of the work by Miguel A. Hernán, W. Wang, and Leaf (2022) and Kuehne et al. (2022). Those guidelines require the definition of a target trial protocol, which we present in Table 4.1. In particular, the table contains information about patients' eligibility criteria, treatment strategies, assignment procedures (of patients to treatment lines), outcome(s) of interest, follow-up period, causal quantities and contrast of interest, statistical methods employed, and confounder variables considered.

Besides emulating this target trial, we employed the CCW strategy (Gaber et al., 2024) to diminish the risk of immortal-time bias. Immortal time bias appears when, in an observational study, information from the *future* of a variable (typically the treatment) is used to assign individuals to a treatment line at *baseline time*. As an example, in the present study, this occurs when an individual is assigned to the treatment line of taking three vaccine doses: until the time that individual gets the third vaccine, they are free of suffering the outcome by definition (because otherwise they would not have been assigned to that treatment line in the first place). This would not happen in a real trial. Besides, a patient assigned to the treatment line of getting zero vaccine doses does not have such a "risk-free" period, and that difference, artificially introduced by the fact that we are emulating a trial with observational data, can induce bias in the results. CCW is a method for correcting this. The cloning and censoring stages of this method were implemented directly as in Gaber et al., 2024, i.e., each patient was *cloned* in the database as many times as different treatment strategies were available (in this case four), then, each clone was assigned to each one of the available treatment strategies, and then clones were *censored* whenever they would violate the treatment strategy assigned to them in the protocol. In this context, censoring means that we stop tracking

Table 4.1: Table of the protocol of the emulated target trial.

Eligibility criteria	Individuals living in Catalonia, under the insurance of the Catalan public healthcare system (CatSalut), who did not have any diabetes diagnostic before the beginning of the study period (1/1/2021).
Treatment strategies	Four possible treatments with the COVID-19 vaccine: no dose (0 doses), 1 dose, 2 doses, or 3 doses. Protocol vaccination dates are assigned based on the vaccination strategy of the Spanish health authorities (*Estrategia de vacunación COVID-19* 2024). Thus, people above 50 years old should get the first vaccine dose after 1/2/2021, the second dose after 1/3/2021, and the third dose after 1/4/2021. Similarly, people below 50 years old should get the first vaccine dose after 1/6/2021, the second dose after 1/7/2021, and the third dose after 1/8/2021. For each individual, a grace period of 180 days is granted for the fulfillment of the conditions of the assigned treatment *strategy*. If, after this grace period, the individual does not follow the assigned treatment, they are censored.
Assignment procedures	Each patient is assigned to the treatment line they followed, looking retrospectively. Clarifying note: if a patient got three vaccine doses but had a diabetes diagnostic after the second dose, it is effectively assigned to the two-dose intervention group.
Outcome	Diabetes diagnostic and its discrete time of occurrence. Type of outcome: discrete time-to-event, survival-type. The diabetes diagnostic is identified by ICD codes. ICD-9: starting with ’250’. ICD-10: starting with ’E10’, ’E11’, or ’E13’
Follow-up	Start of follow-up: 1/1/2021. End of follow-up: 31/12/2023.
Causal quantities and contrasts of interest	Per protocol effect. Cumulative risk of new-onset diabetes under the different intervention strategies at the end of follow-up, and ratios of those risks with respect to a reference. The reference is the cumulative risk of the ”natural course,” i.e., the risk under no specified intervention (just the observed ones).
Statistical methods	Parametric, non-iterative conditional expectation (NICE) implementation of the G-formula; controlling for time-invariant and time-varying confounding.
Confounder variables	Time invariant confounder variables: date of birth, sex assigned at birth, country of birth, area of residence within Catalonia, and an indicator of the socioeconomic level of that area of residence. Time-varying confounder variables: body mass index, systolic and diastolic blood pressure levels, cholesterol level, blood glucose level, abdominal perimeter, smoking status, adjusted comorbidity index (Monterde, Vela, and Clèries, 2016) (similar to Charlson index) and COVID-19 infection status.

those patients/clones without knowing whether they developed the outcome (new-onset diabetes) afterward or not. All we know is that they did not develop the outcome during the time they were followed, i.e., we have *partial information* about them. This concept of censoring and partial information was introduced in Chapter 2, section 2.2.1, in survival analysis. For the weighting stage, we employed stabilized inverse probability of censoring weights (sIPCW), computed as

$$sIPCW = \prod_{k=0}^{K} \frac{P(C_k = 0|U_0, C_{k-1} = 0)}{P(C_k = 0|U_0, V_k, C_{k-1} = 0)} \tag{4.1}$$

where k is a discrete-time index of the follow-up period. In this case, three possible intervention times existed, as it was possible to get a maximum of three vaccine doses, thus $k \in [0, 2]$); C_k is the censoring indicator, i.e., a binary variable indicating whether the patient has been censored at time k or not; U_0 is the vector of baseline, time-invariant confounders; and V_k is the vector of time-varying confounders at time k. A weight was computed for each uncensored patient/clone in the cohort at the end of follow-up time. The probabilities present in the weights' formula were modeled with pooled logistic regressions.

Note that, due to censoring, not all patients were followed up for the same amount of time, hence the need to employ a time-to-event, survival-type outcome variable.

DAG of the problem

We developed a directed acyclic graph of the problem, consulting experts in the field for coding assumptions about causal relations among the variables and for identifying the adjustment set (i.e., the minimum set of variables to adjust for, in order to obtain an unbiased estimate of the effect of interest). Because the graphical representation of the full DAG had many nodes and edges, Figure 4.1 shows a schematic version of it, where four types of nodes are depicted: treatment variable nodes, the outcome variable node, time-varying covariate nodes and time-invariant covariate nodes. Note that there are three possible treatment nodes and, thus, three time-varying confounder nodes. Besides the schematic version, we include in Appendix A a version of the DAG in which time has been *collapsed* into a single step, but which contains all the variables of the problem (except for latent ones, i.e., variables that are known but uninformed and that do not introduce confounding). The adjustment set was formed by the confounder variables mentioned in Table 4.1.

Note that the confounder selection was based on the main risk factors of type 2 diabetes. For the case of type 1, these factors are mostly unknown, except for the genetic ones (ElSayed et al., 2023). The genes that contribute to the development of type 1 diabetes provide instructions for making proteins that play a role in the immune system (H. S. Lee and Hwang, 2019). We assumed that such genes would have no influence on the probability that a patient would get the COVID-19 vaccine and thus would not induce confounding.

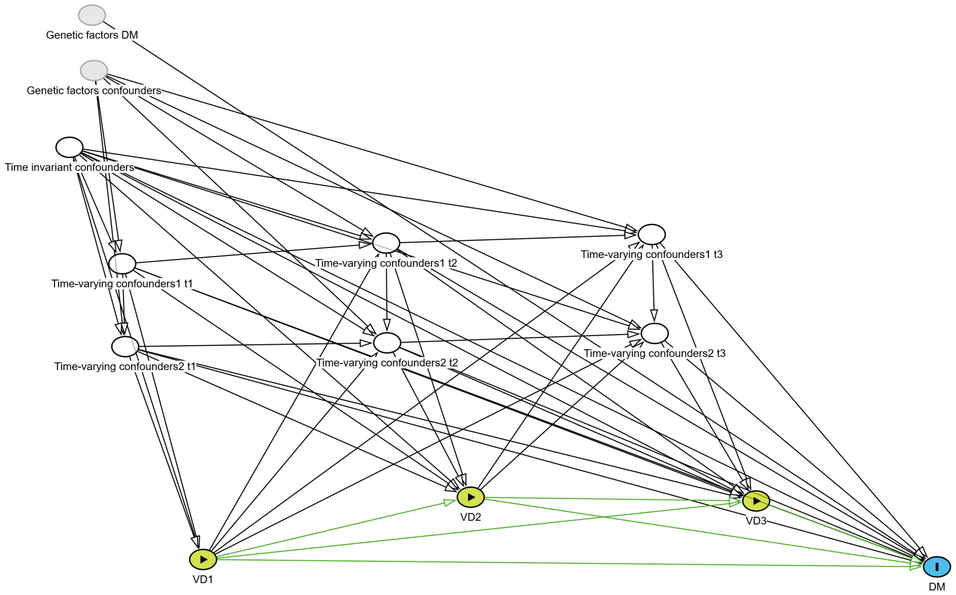

Genetic factors DM
Genetic factors confounders
Time invariant confounders
Time-varying confounders1 t1
Time-varying confounders2 t1
Time-varying confounders1 t2
Time-varying confounders2 t2
Time-varying confounders1 t3
Time-varying confounders2 t3
VD1
VD2
VD3
DM

Figure 4.1: Simplified version of the DAG of the problem. Observed variables: time-invariant confounders; two sets of time-varying confounders at three different time points. Intervention variables: vaccine doses (VD) at three different time points. Outcome variables: diabetes (mellitus, DM).

Causal quantities and effects of interest

In this section, we define the causal quantities and contrasts of interest. To do so, we first need to provide some definitions. Thus, let our discrete-time outcome, new-onset diabetes at time k, be Y_k, a binary variable taking value 1 if the patient had a diabetes diagnostic at that time and 0 otherwise. Similarly, Z_k is the vector of covariates (time-varying and time-invariant) at time k, and C_k is the censoring indicator at that time. In general, for a random variable A, $\bar{A}_k$ denotes the *history* of A through k, $(A_0, ...A_k)$. Let us denote a treatment *strategy* by the random variable $\bar{X}$ and a realization of it by the lowercase $\bar{x}$. Note that a treatment strategy $\bar{X}$ is formed by the sequence of (binary) treatment realizations $X_k = x_k$ at each time step, thus $\bar{X}_K = (X_0 = x_0, ...X_k = x_k, ...X_K = x_K)$, being K the total number of time steps. In our particular case, the strategy of receiving 0 doses of the vaccine is $\bar{X} = (X_0 = 0, X_1 = 0, X_2 = 0)$, and we refer to it as $0\bar{D}$, 1-dose strategy is $1\bar{D} = (1, 0, 0)$, 2-doses strategy is $2\bar{D} = (1, 1, 0)$, and finally, 3-doses strategy is $3\bar{D} = (1, 1, 1)$. Recall that our discrete-time outcome for time k is Y_k. Then, the discrete, *hazard* or risk of the outcome at time k is $P(Y_k = 1 | Y_{k-1} = C_{k-1} = 0)$, and the cumulative risk is $\sum_k P(Y_k = 1 | Y_{k-1} = C_{k-1} = 0)$. This (the cumulative risk) is the causal quantity of interest that we want to compute. In particular, we want to compute it using the previously defined strategies. We can express that, with counterfactual notation, as $P(\bar{Y}_K^{0\bar{D}} = 1)$,

$P(\bar{Y}_K^{\bar{1D}} = 1)$, $P(\bar{Y}_K^{\bar{2D}} = 1)$ and $P(\bar{Y}_K^{\bar{3D}} = 1)$. Finally, the causal effects or contrasts that we want to compute are the ratios of those cumulative risks with respect to the risk of the so-called *natural course* (expressed as $P(\bar{Y}_K^{\bar{NC}} = 1)$), i.e., $P(\bar{Y}_K^{\bar{0D}} = 1)/(\bar{Y}_K^{\bar{NC}} = 1)$, etc. The *natural course* risk is the risk that would be observed if no treatment policy was applied, just the observed treatments (Young et al., 2011).

Estimator and algorithm: Parametric NICE G-formula with random forests

For estimating the aforementioned causal quantities, we employed the parametric non-iterative conditional expectation (NICE) implementation of the G-formula (Chiu et al., 2023). Other options such as the iterative conditional expectation (ICE) or the inverse probability of treatment weighting (IPTW) were also available. Evidence shows that their performance is, in general, similar (Wen et al., 2021). The NICE implementation requires some extra assumptions compared to ICE or IPTW-based implementations, as it requires researcher-defined regression models. We divide the explanation of this estimator in two parts: first, we explain the G-formula itself, and then, its parametric NICE implementation.

The G-formula was introduced by J. Robins, 1986 for measuring causal effects in settings with time-varying confounding. In such settings, traditional adjustment methods fail, providing biased effects estimates even with all confounders correctly identified. This formula takes different shapes depending on the particular characteristics of the problem at hand. Our case is one of discrete-time survival with static and deterministic treatment regimes. A static treatment is one that does not depend on past covariate *history*, and deterministic means simply that it is not random. In such a scenario, the cumulative risk under a given treatment strategy is given by the G-formula as

$$\int_{\bar{z}_{k-1}} \sum_{k=0}^{K} P(Y_k = 1 | Y_{k-1} = C_k = 0, \bar{z}_{k-1}, \bar{x}_{k-1})$$
$$\times \prod_{s=0}^{j-1} P(Y_s = 0 | Y_{s-1} = C_s = 0, \bar{z}_{s-1}, \bar{x}_{s-1}) f(z_s | Y_s = C_s = 0, \bar{z}_{s-1}, \bar{x}_{s-1}) \tag{4.2}$$

given that the identifiability conditions (*positivity*, *consistency* and *sequential exchangeability*) hold, and with $f(z_k | Y_k = C_k = 0, \bar{z}_{k-1}, \bar{x}_{k-1})$ being the joint density of confounders at time k under the given treatment. For a detailed explanation, see the works Miguel A Hernán and James M Robins, 2020; Wen et al., 2021.

The parametric NICE implementation of the G-formula works in two stages: first, it is necessary to fit models of the components present in the G-formula. Those components are, on the one hand, the conditional distribution of each covariate given censoring, past covariate, and treatment histories, and, on the other hand, the probability of an outcome given censoring, past covariate, and treatment histories. In both cases, this is done using regression models provided by the user.

The specific regression models employed in this case can be found in Appendix B. Then, the second stage requires approximating the integral and sum of the G-formula by performing Monte-Carlo simulations (n times) in four steps: 1) at time step $k = -1$, sampling from the observed baseline confounders, and assigning the treatment of interest; 2) at time steps $k \geq 0$, simulating the confounders using the fitted models from the previous stage and the confounder values from $k - 1$ as inputs for the models, as well as the treatment assigned per the strategy; 3) for each $k + 1$ time, simulate the outcome, using the fitted outcome model from the previous stage, and using the confounders simulated in step 2 and the treatment as per strategy as inputs; 4) finally, compute the cumulative risks approximating expression 4.2.1 with the Monte Carlo integration method.

We employed the *pygformula* python package (*pygformula* 2024) as the base for our algorithm. Nevertheless, we introduced two modifications. Firstly, the base algorithm used generalized linear models (GLM) for the modeling steps, and we changed those by random forests. Secondly, we introduced the computation of the weights of the CCW method, explained in section 4.2.1. After their computation, we used them as sample weights provided to the random forest during the modeling of the conditional outcome, which is part of the first stage of the NICE G-formula. The pseudo-code of this algorithm can be seen in 4.1.

Note that the substitution of GLMs for random forests in the NICE G-formula improved the efficiency of the resulting algorithm. Our modified algorithm was faster and less memory-consuming than the original. In fact, the original version would deplete our memory resources ($\sim$300GB of RAM) even for small subsets (as small as 1%) of the whole cohort of patients. Optimization attempts such as paralleling or dividing in batches the minimization process of the fitting stage of GLMs or storing the data in efficient formats were still slower and more memory-consuming than our algorithm. In that sense, it is worth highlighting that our adapted algorithm was more suitable for working with large amounts of data than the original.

For computational reasons we randomly divided our cohort into ten equal chunks, we estimated the causal effects of interest in each of them and then we computed averages and standard errors. The underlying assumptions of this strategy are that, across chunks, samples are independent and identically distributed and that the causal effects of interest follow a normal distribution.

4.2.2 Data

In this subsection, we describe the employed data and the preprocessing steps. Catalonia has a population of around 8 million inhabitants and a public healthcare system that provides coverage to all of them. Despite the existence of private healthcare services, the public system is predominant in domains such as the surveillance of infectious diseases, the implementation of vaccination programs, and the attention to chronic diseases. In the particular case of the vaccination program against the COVID-19 pandemic, given the extraordinary circumstances in which it unfolded, the

Algorithm 4.1: RF-based, parametric NICE G-formula with CCW

```
   input  : Cohort of patients
   output : $P(\bar{Y}_K^{\bar{x}} = 1) \quad \forall \bar{x} \in (0\bar{D}, 1\bar{D}, 2\bar{D}, 3\bar{D})$
1  for $patient \in$ Patients do
2      Clone: clone $patient$ $T - 1$ times, with $T$ the number of possible interventions; we
         refer to clones and $patient$ simply as clones
3      Assign: Assign clones to treatment strategies
4      for $i \in [0, T - 1]$ do
5          clone$_i \to \bar{X} = i\bar{D}$
6          Censor: Censor clones that violate the protocol
7          if violatesProtocol(clone$_i$) then
8              $C_{\mathbf{clone}_i} = 1$
9          end
10     end
11 end

12 Fit $P(C_k = 0 | U_0, C_{k-1})$ $\forall k \in [0, K]$ with logistic regression, data of all clones
13 Fit $P(C_k = 0 | U_0, V_k, C_{k-1})$ $\forall k \in [0, K]$ w. logistic regression, data of all clones
14 Compute sIPCW$_i$ using the models of steps 12, 13, as

       $$\prod_{k=0}^{K} \frac{P(C_{k,i} = 0 | U_{0,i}, C_{k-1,i} = 0)}{P(C_{k,i} = 0 | U_{0,i}, V_{k,i}, C_{k-1,i} = 0)} \quad \forall i \in [1, I], \text{with } I \text{ the number of clones}$$

15 Fit model of $f(z_k | Y_k, C_k, \bar{z}_{k-1}, \bar{x}_{k-1})$ with user-provided covariate regression models,
     random forests, data of all clones
16 Fit model of $P(Y_k = 1 | Y_{k-1}, C_k, \bar{z}_{k-1}, \bar{x}_{k-1})$ with user-provided outcome regression
     model, random forests, sIPCW, data of all clones

17 for $\bar{x} \in (0\bar{D}, 1\bar{D}, 2\bar{D}, 3\bar{D})$ do
18     for $m \in [1, n]$ do            // with $n$ the number of MC simulations
19         for $k \in [-1, K]$ do
20             if $k$=-1 then
21                 Sample covariates
22             else
23                 Compute components of expression 4.2.1 using models from steps 15, 16
24                 Compute result of expression 4.2.1 using results of step 23
25             end
26         end
27     end
28     Compute average of results over the $n$ simulations (Monte Carlo integration method)
29 end
```

public health system was the only healthcare provider involved in its development (i.e., the only provider administering vaccines).

Several electronic health records from the Catalan health service were employed. In particular, those were the databases of hospital discharges, emergency rooms, primary care, results of laboratory tests of primary care, vaccination status, the specific registry of COVID-19, and the central registry of insured patients. In all databases, patients were identified by a unique and common pseudo-anonymous identifier, which allows for the linkage of the data.

The starting point of the information was the central registry of insured patients, which contained the all-time list of insured patients of the Catalan public health system. It contained more than 12 million registries. We filtered patients who died before the beginning of the study date, 1/1/2021, and patients who were not active at the extraction time, January 2024. This resulted in 8,049,335 unique patients. The table contained information about the sex assigned at birth, birth date, country of origin, and region of residence within Catalonia, besides the date of death. This list of patients was the base employed for the lookup of the rest of the variables in the rest of the databases.

Three index dates were assigned to each patient. The values of those index dates were taken from the national vaccination strategy of the Spanish health authorities *Estrategia de vacunación COVID-19* 2024. Each of these three dates represents the ideal time in which a patient should have received each dose of the COVID-19 vaccine.

The specific registries of COVID-19 contained information about infections and vaccination status. Thus, from these databases, we obtained infections, infection dates, number of administered vaccine doses (0, 1, 2, or 3 doses), and dose administration dates.

The databases of hospital discharges, emergency rooms, and primary care register the activity in those domains as visits. Each visit was associated with one or more medical diagnoses coded with the ICD9 and ICD10 standards. Thus, we obtained the diabetes diagnoses and their dates from these databases. After the search, episodes were ordered by date, and the first one was kept. Patients with a diabetes diagnosis before the study start date were excluded from the cohort.

The database of results of laboratory tests of primary care collects information on clinical tests ordered by primary care specialists. From this database, we obtained results and dates of tests of blood glucose, systolic and diastolic blood pressure, abdominal perimeter, body mass index, and cholesterol level.

Finally, from other tables, we obtained an indicator of the socioeconomic status of the region where each patient lived, their smoking status, and an adjusted comorbidity indicator (Monterde, Vela, and Clèries, 2016) similar to the Charlson index.

All this information was combined through a merging algorithm into a single table. The pivotal information was the vaccination status. Recall that, for each patient, we had the number of vaccine doses administered (0, 1, 2, or 3) and each dose's administration date or its index date. Then, the information on the time-varying confounders was added by selecting the closest available value of each variable prior to the vaccine administration date or index date. This process was repeated for each person and each vaccine dose. Finally, the information on the time-invariant confounders was merged by the patient identifier. Patients who had information about the second and/or third dose without information about the first dose were dropped, assuming that these errors occurred randomly. This resulted in 7,499,081 patients included in the final cohort.

Data preprocessing

After the database was constructed, we took some preprocessing steps. Those included

- removing outlier values of BMI above 70 kg/m^2 (set to missing) and dropping patients aged above 110 years (assumed dead),
- imputing missing values of country of birth and area of residence (sampling from a distribution with probabilities equal to the frequencies of each category of those variables),
- causally imputing missing values of time-varying confounders, using the same confounder models employed in the G-formula,
- imputing missing values of time-varying confounders' value dates, using the beginning of study date,
- and finally, carrying forward the information of time-varying confounders with missing values at time steps 2 and/or 3 (recall that our database contains information on three different moments in time).

4.3 Results

In this section, we present the results of the described study. Table 4.2 shows the cumulative risk of new-onset diabetes in each intervention group after the whole study period and the ratio of those risks with respect to the natural course of the disease. Point estimates, together with their 95% confidence intervals, are provided. Recall that the different interventions were receiving 0, 1, 2, or 3 doses of the vaccine against COVID-19, and natural course refers to the result under no intervention strategy, just the observed treatments. The results show that getting the vaccine against COVID-19 has a protective effect against diabetes onset, as the accumulated risk of developing the disease is 0.096 (0.072, 0.120) (risk ratio of 1.465 (1.354, 1.575)) in the group who got 0 doses of the vaccine, versus probabilities of 0.058 for 1 and 2 doses, and 0.054 for 3 doses (risk ratios of 0.899, 0.887

Table 4.2: Cumulative risk of diabetes onset for the different interventions (left column), and the ratio of those risks with respect to ”Natural course” risk (right column), after a total study time of 3 years and setting the grace period for protocol violations to 180 days. Point estimates together with 95% confidence intervals (between brackets) are provided.

Intervention	Cumulative risk of diabetes onset at the end of study time [95% CI]	Risk ratio [95% CI]
Natural course	0.065 [0.052, 0.078]	1.000 [1.000, 1.000]
0 doses	0.096 [0.072, 0.120]	1.465 [1.354, 1.575]
1 doses	0.058 [0.048, 0.069]	0.899 [0.870, 0.927]
2 doses	0.058 [0.048, 0.068]	0.887 [0.855, 0.919]
3 doses	0.054 [0.045, 0.062]	0.827 [0.770, 0.884]

and 0.827 respectively). This means around 45% more risk of diabetes for patients who got 0 doses of the vaccine with respect to the natural course, and around 10, 11, and 18% less risk for patients who got 1, 2, and 3 doses, respectively. Some dosage effect is observed, as the risk decreases when the number of doses increases.

4.4 Discussion and conclusions

The epistemic starting point for this study was the evidence indicating that the infection of COVID-19 increased the risk of diabetes onset and the evidence suggesting that the vaccines protected from that increased risk. Based on that information, we aimed to estimate the effect of the vaccine on the risk of diabetes. Because we have included the COVID-19 infection status as a variable in our problem (i.e., its effect in the outcome is partialled out by the G-formula), and because our only intervention variable was the vaccine, we have effectively estimated the *direct* effect of the vaccine on the outcome, i.e., $Vaccine \rightarrow Diabetes$, as opposed to the effect mediated by the COVID-19 infection, $Vaccine \rightarrow Infection \rightarrow Diabetes$ (note that, in the DAG, both of these causal paths are present). Our initial hypothesis was that this direct effect should be very small or zero, but we found otherwise. One possible explanation is that our results could be biased due to under-reporting of the COVID-19 infection status variable or due to the lack of a finer indicator, such as infection severity. In any case, the overall picture clearly shows that getting the COVID-19 vaccine is an effective strategy for avoiding the diabetes onset risk increase caused by COVID-19 infection.

The main limitation of this study is that the employed data was not specifically recorded for research but generated through routine provision of healthcare services and re-utilized for research instead.

Despite being a limitation, this is also the main feature of the so-called real world data (RWD), and the evidence generated with it, real-world evidence (RWE) (M. Li et al., 2021). RWD can be less complete, but in turn, it is usually much larger, with sample sizes that would be difficult or expensive to obtain solely for research. In addition, it provides evidence about real-world scenarios and healthcare provision, in contrast with experimental designs, which can sometimes face problems with the generalizability of their findings (Nieto-Gómez et al., 2024). Another potential limitation is the existence of hidden confounder variables. Finally, on the methodological side, a potential limitation derives from the so-called G-null paradox, related to model misspecification affecting parametric implementations of the G-formula (for a detailed explanation, see for instance (McGrath, Young, and Miguel A. Hernán, 2022)). Nevertheless, it has been shown that the risk of the G-null paradox can be minimized by not using overly parsimonious models, and we consider that we fulfill this condition. Furthermore, the employed software package provides nonparametric estimates of the G-formula for the natural course line as an indicator of model misspecification: divergence between the parametric and nonparametric estimates would constitute evidence of model misspecification. We did not observe significant differences between those estimates in our experiments.

An important line of future work will include obtaining further evidence in favor of the previously mentioned hypothesis, i.e., that the COVID-19 vaccine protects from the increased risk of diabetes through COVID-19 infection (mediated by it). For that purpose, we could perform another intervention with the G-formula, thanks to the fact that this method allows us to intervene in any confounder of the problem. Thus, we can impose different rates of COVID-19 infection in the group with 0 doses and the groups with at least 1 dose. A bigger protective effect of the vaccine against diabetes in any of the vaccinated groups would constitute evidence in favor of the aforementioned hypothesis. Furthermore, trying other implementations of the G-formula or considering other G-methods, such as G-estimation, would increase the robustness of the results. Another line of future work is to assess the impact of the grace period. For this purpose, we will conduct a sensitivity analysis of the results with respect to that parameter.

4.4.1 Critical assessment of the employed methodology

In this chapter, we have answered research question **Q2** (*Does the vaccine against COVID-19 have any effect on the risk of developing diabetes?*) with observational data, a causal approach, and causal inference methods. In particular, we developed a DAG of the problem, employed the TTE with CCW framework, and adapted and used an advanced implementation of the G-formula. In the following paragraphs, we discuss the main advantages and strengths, as well as disadvantages and weaknesses, of this approach to answer research subquestion **Q4b**, *What are the advantages and disadvantages, strengths and weaknesses, of causal inference methods for generating evidence about clinical interventions?*

Among advantages and strengths, there is a critical remark to make: without the G-Methods (family to which the G-formula belongs), effectively controlling for time-varying confounding is not possible. Thus, in this scenario, causal inference methods are not just an option but a must. Besides that, employing the target trial emulation framework with CCW mimics the dynamics of a pragmatic randomized controlled trial (pragmatic meaning that the treatment is not blinded to patients nor health workers), and that has important implications in two aspects. On the one hand, speaking the clinical trial *language* facilitates the communication between researchers, accustomed to working with observational data, and clinicians, accustomed to working with trial data. On the other hand, as we already mentioned, some evidence suggests that this design can actually close the gap between results from observational studies and clinical trials (Kuehne et al., 2022; S. V. Wang, Schneeweiss, and Initiative, 2023). More comparisons between actual randomized trials and trial emulations are currently being performed, so more evidence in one direction or the other is to be expected. Finally, our explicitly causal approach allows us to speak about actual causal effects and not just correlations, even when the observed results may deviate from our original hypothesis, and we recommend further research. Out of the works analyzed in the literature about this topic, none possesses this feature, as they are all correlational works.

Among disadvantages and weaknesses, it must be mentioned that G-methods have relatively low penetration in observational studies due to their allegedly higher complexity compared to other more straightforward approaches (in scenarios without time-varying confounding). Thus, it is often the case that G-methods are only employed when disregarding the time-varying nature of confounding in the problem at hand is unaffordable. Nevertheless, this tendency is changing and will continue to do so, partially thanks to causal methods gaining popularity and being available in open-source software packages (McGrath, Lin, et al., 2020; *pygformula* 2024).

In the next chapter, we analyze another health technology, antibiotic-loaded bone cement, and we assess its impact on knee prosthesis survival. We will perform both a correlational and a causal analysis, which will allow us to make a direct comparison.

Chapter 5

The effect of antibiotic-loaded bone cement on the survival of the prosthesis after total knee arthroplasty

This chapter presents the answer to research question **Q3**, *Does the use of antibiotic-loaded bone cement during total knee arthroplasty surgery increase the life of knee prostheses, compared with the alternative of using plain cement?*

The chapter is divided into three main parts:

- An introduction about the motivation and relevance of question **Q3**, and a description of the observational data we have used to answer it.
- A correlational analysis of the problem. This part contains the work of the journal article ”Gil-Gonzalez Sergi, Velasco-Regúlez Borja, Cerquides Jesus, et al. (2024). Antibiotic-loaded bone cement (ALBC) is associated with a reduction of the risk of revision of total knee arthroplasty: Analysis of the Catalan Arthroplasty Register. *Knee Surgery, Sports Traumatology, Arthroscopy*. DOI: 10.1002/ksa.12361.” An earlier version of that work (”Sergi Gil-Gonzalez, Borja Velasco-Regúlez, Jesus Cerquides, et al. (2023). ¿El cemento con antibiótico reduce el riesgo de infección protésica en artroplastia primaria total de rodilla? Análisis del registro catalán de artroplastias. *10º Congreso de la AEA-SEROD*”) was presented as an oral poster communication at the 10^{th} congress of the Spanish Arthroplasty Association (AEA) and the Spanish Knee Association (SEROD), receiving the award to the best oral communication.
- A causal analysis of the problem. This work has been sent to the Journal of Healthcare Informatics Research, under the title ”Causal analysis of the effect of antibiotic-loaded bone

cement on knee prosthesis survival" and is currently under peer review.

In addition, we also present an answer to research question **Q4**, *What are the advantages and disadvantages, strengths and weaknesses, of correlational methods and causal inference methods for generating evidence about clinical interventions?*. We do so by carrying out a comparison between both approaches using the use case of **Q3** as a basis.

5.1 Background

Peri-prosthetic joint infection (PJI) is a major complication of total knee arthroplasty (TKA), which happens in between 1% and 2% of the cases (R. Frank, M. Cross, and C. Della Valle, 2015). This complication often necessitates revision surgery, which significantly reduces the patient's quality of life and satisfaction (Garvin and Konigsberg, 2011). The last decades have witnessed many contributions and improvements aimed at reducing the rate of PJI. Some examples are the use of prophylactic antibiotics, improvements in orthopedic theatres, and modifications in preoperative patient preparation (Parvizi, Cavanaugh, and Diaz-Ledezma, 2013). In the case of total hip arthroplasty, the use of antibiotic-loaded bone cement (ALBC) during primary arthroplasty has also been shown to decrease the rate of PJI (Engesæter et al., 2006; Hinarejos et al., 2013). However, in the case of TKA, the evidence of the benefit of that strategy is inconclusive (T. H. Leta et al., 2021; H.-Q. Li et al., 2022). Downsides such as the possibility of altering the mechanical properties of the cement, the generation of antibiotic microbial resistance or the increasing cost (Dunne et al., 2007; Hoskins et al., 2020; King et al., 2018) cause a lack of consensus about this intervention across countries, and there is substantial variability in the findings reported by studies carried out in different countries or regions (T. H. Leta et al., 2021). Places like the United Kingdom or the Scandinavian countries use ALBC in primary TKA in more than 90% of cases, but this percentage is much lower in places like the United States, Spain, or Russia (Randelli et al., 2010). Furthermore, patients' preoperative characteristics can also be associated with the risk of developing PJI. Gender, age, previous surgeries, and comorbidities, such as diabetes, obesity, or inflammatory diseases, increase the probability of septic complications (Namba, Inacio, and Paxton, 2013).

The question of whether ALBC usage during total knee arthroplasty has an impact on prosthetic survival, be it positive or negative, is a causal question, as it speaks about a cause-and-effect relationship and can be subjected to confounding bias (Miguel A Hernán and James M Robins, 2020; Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell, 2016). As we have already discussed in previous chapters, these types of questions have traditionally been tackled in the literature of clinical observational studies with causality-free, associational-only approaches, despite sometimes using causal vocabulary. Studies about prosthetic survival are no exception to that trend. Some examples: in Jameson et al., 2019, authors speak about ”associations” between ALBC and prosthetic survival, although they also use causal concepts such as ”adjusting”; in Bohm et al., 2014, authors speak about ”effects” of ALBC and ”confounding” of other factors, both of which are causal concepts, but then refrain from drawing causal conclusions from their results. Some authors have referred to this situation as the ”causal-word ban” (Miguel A. Hernán, 2018), a ”ban” on using the word ”causal” when employing observational data. As we have already discussed before, this constitutes an epistemic limitation, as avoiding explicitly using the word ”causal” does not change the causal nature of the question, and that can create confusion (Gershman and Ullman, 2023).

The goals of this chapter are two-fold. On the one hand, we want to provide an answer for research question **Q3**, *Does the use of ALBC during total knee arthroplasty surgery increase the life of knee prostheses, compared with the alternative of using plain cement?* We do that using observational data and employing both a correlational and a causal approach (separately). To the best of our knowledge, this is the first time that causal inference methods have been employed to tackle this question. Using both approaches will allow us to compare them and provide an answer to research question **Q4**, *What are the advantages and disadvantages, strengths and weaknesses, of correlational methods and causal inference methods for generating evidence about clinical interventions?*. Recall that we have already provided partial answers for this question in the two previous chapters through subquestions **Q4a** and **Q4b**, but in this case, we will be able to directly compare the performance of both approaches with the same use case.

The rest of the chapter is distributed as follows. First, we describe the data sources and the data that have been employed in this work, which were common for both the correlational and the causal analyses. Then, we delve into the correlational analysis, explaining the methods, results, and conclusions of that approach. After, we do the same with the causal analysis. Finally, we present the overall conclusions of the chapter, with a critical assessment of both approaches and their comparison. Note that due to the different idiosyncrasies of the correlational and the causal approaches, their respective sections do not run exactly in parallel, and some particularities in the framing of the problem, vocabulary, etc., should be expected. Also note that the conclusions section of each approach is independent, and the comparison between approaches is only established in the overall conclusions section.

5.2 Data

Our study primarily relied on data from the Catalan Arthroplasty Register (RACat), a population-based registry collecting information on knee and hip replacements performed in Catalonia since 2005. Initially voluntary, RACat became mandatory in 2015 and currently covers the activity of 51 out of 56 public hospitals. From this database, we obtained information such as the primary surgery date, the employed type of bone cement (plain or loaded with antibiotic), the *revision surgery* date,

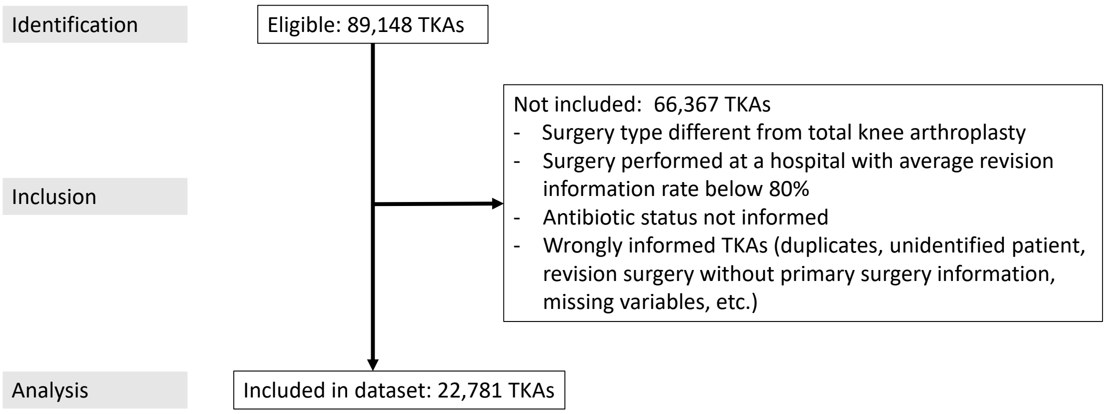

Identification
Eligible: 89,148 TKAs
Not included: 66,367 TKAs
- Surgery type different from total knee arthroplasty
- Surgery performed at a hospital with average revision information rate below 80%
- Antibiotic status not informed
- Wrongly informed TKAs (duplicates, unidentified patient, revision surgery without primary surgery information, missing variables, etc.)
Inclusion
Analysis
Included in dataset: 22,781 TKAs

Figure 5.1: STROBE (Strengthening the reporting of observational studies in epidemiology) diagram of the study. Starting point of 89,148 eligible TKAs, for finally analyzing 22,781 TKAs.

if any, and the revision surgery reason, if any. Additionally, we employed data from the Catalan Institute of Health (ICS) and the Catalan Health System (CatSalut) for information on hospital activity and primary care assistance. In particular, we used the registries of the Basic Minimum Set of Data (BMSD) of hospital discharges and primary care databases to obtain patients' diagnostics before knee surgery. BMSD and primary care databases are mandatory registries, and thus they constitute the gold standard of our datasets.

Case inclusion criteria and data preprocessing

We began by selecting all knee arthroplasty procedures recorded in RACat between 2011 and 2020 (both included). Arthroplasties were followed up until 31/12/2023. We excluded non-standard procedures like uni-compartment replacements and those using less common cementing techniques. To minimize potential biases caused by missing revision data, we excluded hospitals with a revision information rate below 80%. This rate was calculated by linking RACat with BMSD data by patient identifier, surgery type, and date of surgery and comparing the number of revisions reported in each registry (in particular, the rate was computed as the fraction of procedures registered by RACat with respect to those registered by BMSD; recall that BMSD is a mandatory registry and thus contains the information of all performed surgeries). Finally, entries with incomplete or inaccurate information were removed. This selection process resulted in a final dataset of 22,781 total knee arthroplasties for analysis. Figure 5.1 shows a flow diagram of this process, using the STROBE framework (Elm et al., 2008).

Note that we conducted a sensitivity analysis of the revision information rate to assess the impact of this parameter. We tried thresholds in the range between 75% and 95%, varying in 5% increments, and analyzed how that affected the overall revision and infection rates. As expected, increasing the threshold drove down the number of included hospitals and the overall size of the dataset but also led to a higher observed rate of both revisions and infections. Ultimately, we set the revision reporting threshold at 80%. This decision balanced the trade-off between the goal of a larger dataset (achieved by a lower threshold) and the goal of a smaller potential bias from missing revisions (achieved by a higher threshold). The selected rate was also the one recommended by the RACat database managers.

Treatment and outcome variables

The treatment was the type of cement employed during arthroplasty surgery, with two possible categories, ALBC and plain cement. The outcome of interest was the revision event, that is, the presence of a revision surgery after the primary surgery, where at least one component was revised (excluding the patella component). The main outcome variable was the time elapsed between surgery and a revision event or the end of follow-up for censored individuals, measured in months. Note that three types of revision events were possible, depending on the nature of the revision: septic revision (revision due to infection), aseptic revision (revision due to a cause other than infection), and all-cause revision. In the remainder of this text, we may use the terms septic revision or simply infection interchangeably.

Other variables

The preoperative characteristics, covariates, or risk factors considered were patient's age, as a continuous variable in years; sex assigned at birth, as a dichotomous variable (woman or man); obesity, diabetes, rheumatoid arthritis, and alcohol abuse, as dichotomous variables (yes or no); smoking status, as a categorical variable (smoker, nonsmoker, former smoker; missing values treated as nonsmoker); body mass index (BMI) as a continuous variable in kg/m2 (missing values imputed with the average stratified by age group); Charlson comorbidity index and Elixhauser index, as continuous variables; hospital category, as a categorical ordinal variable with five categories (between 1 and 5); primary surgery year, as a categorical ordinal variable (from 2011 to 2020); surgery duration, as a continuous variable in minutes (missing values assigned with the average stratified by the hospital. Note that the hospital category classifies hospitals regarding their size and specialization level, and it is a categorization established by CatSalut. Category 1 is for high-technology reference hospitals, while Category 5 is for regional, basic hospitals. Alcohol abuse was defined as in the definition of the Elixhauser index (Lix et al., 2016).

5.2.1 Descriptive analysis of the data

We performed a descriptive analysis of the employed database with the aim of knowing its taxonomy. In addition, we stratified this description by the treatment variable, cement type. Table 5.1 shows the centrality and dispersion measures of the most relevant preoperative characteristics (confounders or risk factors) of the study population, stratified by treatment. Overall, plain cement was used in 9656 (42.4%) cases and ALBC in 13,125 (57.6%) cases. In the ALBC group, gentamicin was used in 12703 (96.78%) cases, tobramycin in 410 (3.12%) cases, and erythromycin in 12 (0.09%) cases. Small but statistically significant differences were found for sex and antibiotic usage (among females, 42.97% received plain cement and 57.03% ALBC, while among males, the percentages were 41.02% and 58.98%, respectively), and for age and antibiotic usage. No significant differences were found regarding the analyzed comorbidities and antibiotic usage, although the small differences observed in the Charlson and Elixhauser indexes were statistically significant. Finally, larger differences were encountered in the variables of smoking status, surgery year, and hospital category. The most significant one is the steady increase in ALBC usage over the years, going from 46.74% in 2011 to 84.16% in 2020 (of all surgeries in each year, respectively).

5.3 Correlational analysis

Having outlined the observational data used in this study, we now proceed to the correlational analysis of the relationship between ALBC and prosthetic survival. In this section, we explain the methods, results, and conclusions of the correlational approach.

5.3.1 Methods of the correlational analysis

For the correlational analysis, we conducted a retrospective cohort study and employed correlational methods. In particular, we employed common survival analysis methods, as other studies in the literature of total knee arthroplasty (Jameson et al., 2019; Bohm et al., 2014). Recall that the question of interest was whether the usage of ALBC had any effect on the survival of the prostheses, and thus the main quantity of interest was prosthetic survival. Other quantities under study were the infection rate and the hazard ratios of the different risk factors. In these lines, we define all these quantities and other related concepts, as well as the statistical estimators employed to approximate them. Let us start by defining the time passed in months between the surgery date and the revision event (or the end of follow-up for censored cases) by the random variable T, and the treatment, ALBC or plain cement, expressed as a binary variable that takes values 1 or 0, by the random variable A. Then, the survival function $S(t)$ is defined as the probability that T is greater than a given time t, $S(t) = P(T > t)$. A widely used nonparametric estimator of this function is the Kaplan-Meier estimator, defined as $\hat{S}(t) = \prod_{t_i \leq t}(1 - \frac{d_i}{n_i})$, where t_i is a time when at least one event occurred, d_i is the number of occurred events at time t_i, and n_i is the number of patients *at risk*, i.e., the number of patients who did not have an event and were not censored just before t_i. By computing this estimator stratified for patients using ALBC and plain cement, we are effectively

Table 5.1: Preoperative characteristics (risk factors) of the study population, stratified by the cement type.

		Plain Cement	ALBC	
Characteristic	Categories			P-value
		n (%)		
Sex	Female	6843.0 (42.97%)	9081.0 (57.03%)	<0.01
	Male	2813.0 (41.02%)	4044.0 (58.98%)	
		mean (SD)		
Age (years)		72.24 (7.61)	71.94 (8.01)	<0.01
Surgery duration		89.82 (14.40)	89.98 (21.21)	n.s.
Charlson Index		0.41 (0.76)	0.46 (0.79)	<0.01
Elixhauser index		1.34 (1.18)	1.41 (1.20)	<0.01
		n (%)		
Obesity	No	8509.0 (42.49%)	11518.0 (57.51%)	n.s.
	Yes	1147.0 (41.65%)	1607.0 (58.35%)	
Diabetes	No	8112.0 (42.68%)	10893.0 (57.32%)	0.04
	Yes	1544.0 (40.89%)	2232.0 (59.11%)	
Rheumatoid arthritis	No	9460.0 (42.47%)	12817.0 (57.53%)	n.s.
	Yes	196.0 (38.89%)	308.0 (61.11%)	
Alcohol abuse	No	9601.0 (42.43%)	13025.0 (57.57%)	n.s.
	Yes	55.0 (35.48%)	100.0 (64.52%)	
Smoking status	Non smoker	7985.0 (43.37%)	10426.0 (56.63%)	<0.01
	Smoker	555.0 (39.67%)	844.0 (60.33%)	
	Former smoker	1116.0 (37.56%)	1855.0 (62.44%)	
		mean (SD)		
Body mass index		31.82 (4.57)	31.76 (4.89)	n.s.

estimating $\hat{S}(t|A)$, and we can observe differences in prosthetic survival between both groups. Nevertheless, note that this estimator does not take into account other potential factors influencing the survival of the prosthesis. For introducing a method that takes into account other factors, let us first define the hazard function, which is the instantaneous rate of events at a given time t, defined as $\lambda(t) = \lim_{\Delta t \to 0} \frac{S(t)-S(t+\Delta t)}{\Delta t\ S(t)}$. And then, we can define the conditional hazard function given the covariates and the treatment, $\lambda(t|X, A)$, and approximate it with a widely used model, the Cox proportional hazards model (Abd ElHafeez et al., 2021), as $\bar{\lambda}(t|X, A) = \beta_0(t) \exp\left([X, A]\beta\right)$. $X \in \mathbb{R}^p$ is the vector of covariates, and $[X, A]$ is the concatenation of such vector and the treatment, β is a vector of parameters, and β_0 is another parameter usually known as the baseline hazard. The Cox proportional hazards model is very popular in survival analysis because it provides hazard ratios of covariates and treatment, which can be interpreted as the influence of confounders and treatment on the risk of the event relative to a reference value. It also makes the assumption that the influence of factors on the risk of the event is time-invariant.

An element that has to be taken into account when performing survival analysis is that of competing events. Competing events are those events that are different from the event of interest and may prevent the event of interest from happening. The Aalen-Johanssen estimator (Aalen and Johansen, 1978) is a matrix version of the Kaplan-Meier estimator that can take into account competing events, while Kaplan-Meier treats them as end-of-follow-up censoring. Given that we had two types of outcome events, i.e., revision for infection and aseptic revision, it could be argued that our scenario was one with competing events and that the Aalen-Johanssen estimator could be more suitable. In order to test if that was the case, we performed a sensitivity analysis, comparing the results of the Aalen-Johanssen and the Kaplan-Meyer estimators. The differences in survival between the estimates of both methods were negligible, and thus, for the sake of simplicity, we decided to treat competing events as censoring and to use the Kaplan-Meier estimator. Note that this is common practice in most studies of the analyzed literature. Note also that this justifies choosing the Cox proportional hazards model as a parametric estimator of the conditional hazard instead of using Fine and Gray's subdistribution method (Fine and Gray, 1999), which is sometimes used in the presence of competing events. Finally, note that death has been considered administrative censoring, which is also a common practice.

Finishing with the statistical aspects, we employed t-tests, χ^2-tests, and Log-rank tests for statistical hypothesis testing, i.e., for determining whether observed differences between measures in means, proportions, and survival curves, respectively, were statistically significant. The confidence level was set to $\alpha = 0.05$.

Finally, we employed the Python programming language (version 3.10) for the analysis of the data and the *lifelines* package for survival-related functions (version 0.27.5).

5.3.2 Results of the correlational analysis

In this section, we present the results of the correlational analysis. In particular, we divide it into the subsection of prosthetic survival and infection rates and the subsection of risk factor analysis.

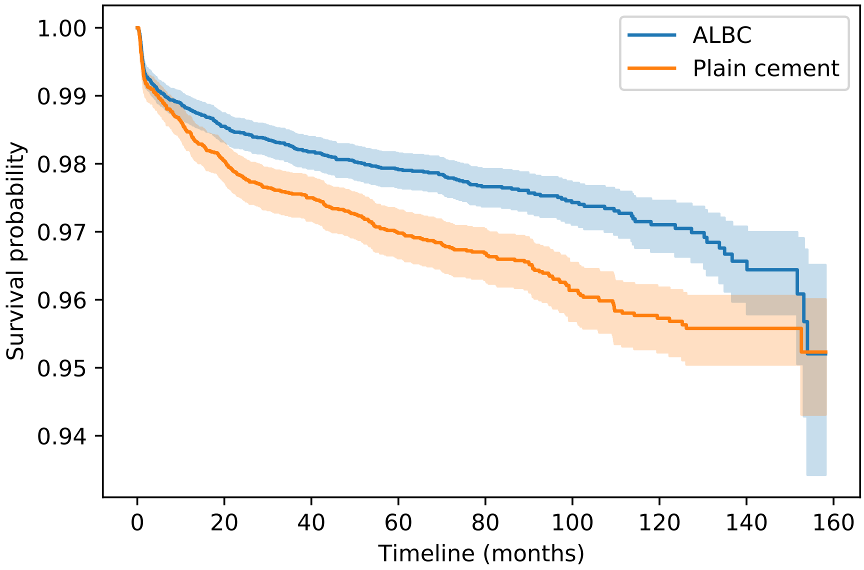


Figure 5.2: Kaplan-Meier curves for revision for infection.

Prosthetic survival and revision rates

In this subsection, we present the results of the estimated prosthetic survival per treatment group, as well as the infection rates for different follow-up times. Recall that the overall follow-up period was 158 months (approximately 13 years). Figure 5.2 shows the Kaplan-Meier survival curves for septic revision, 5.3 for aseptic revision, and 5.4 for all-cause revision, for the whole duration of the follow-up period. The ALBC group showed higher survival values for all endpoints, with the log-rank test showing that the differences were statistically significant in all cases. The survival values (with 95% confidence intervals in brackets) at the end of the study time were 95.2% (93.4%, 96.5%) for ALBC and 95.2% (94.3%, 96%) for plain cement, in the case of septic revision; 90.2% (86.7%, 92.9%) for ALBC and 85% (83.8%, 86.1%) for plain cement, in the case of aseptic revision; and 85.9% (82.3%, 88.8%) for ALBC and 81% (79.6%, 82.2%) for plain cement, in the case of all-cause revision.

Table 5.2 shows the results of infection rates, computed as the fractions of infected prostheses with respect to all prostheses, stratified per treatment, and for shorter follow-up times (3, 6, 12, and 24 months, respectively). For this calculation, only uncensored arthroplasties were employed, i.e., arthroplasties for which no loss of follow-up happened. This reduces the risk of potential bias introduced by censoring. The infection rate was lower for the antibiotic cement group in all cases, with statistically significant differences.

Figure 5.3: Kaplan-Meier curves for aseptic revision.

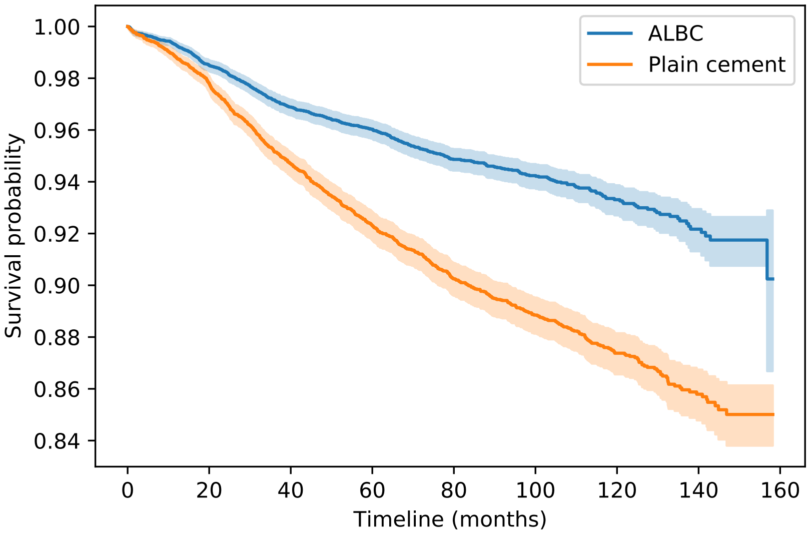


Figure 5.4: Kaplan-Meier curves for all-cause revision.

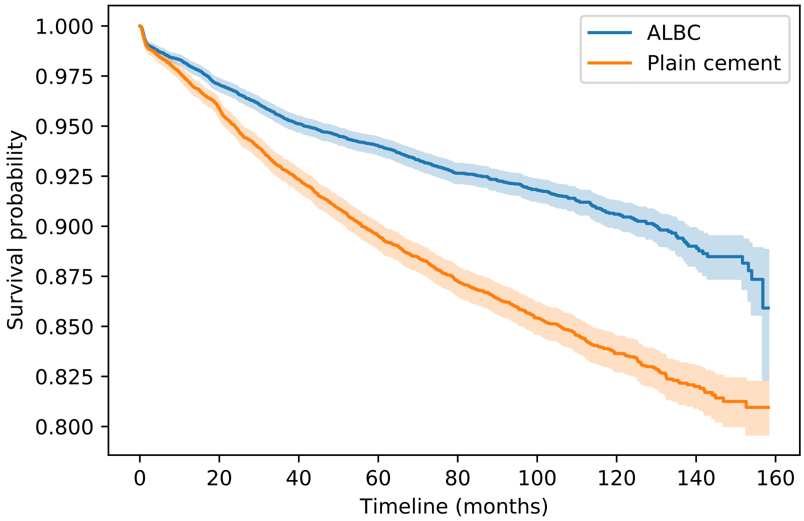


Table 5.2: Infection rates for short follow-up times, stratified by the cement type.

	Cement type		
Follow-up time	Plain cement %	ALBC %	P-value
3 months	0.78	0.52	0.04
6 months	0.98	0.68	0.04
12 months	1.33	0.72	<0.01
24 months	1.79	0.84	<0.01

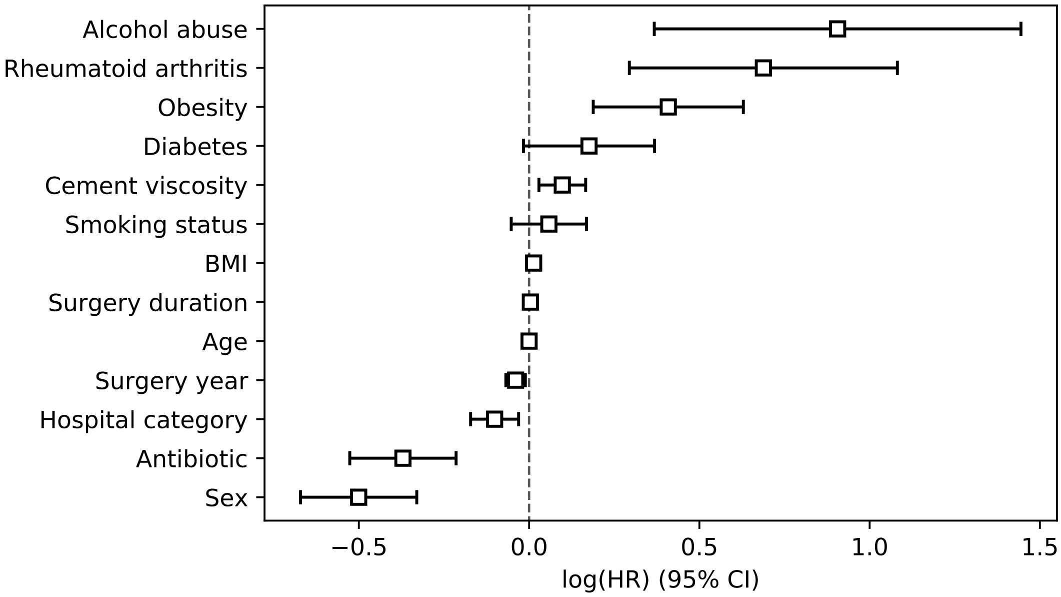


Figure 5.5: Boxplot of the hazard ratios of the Cox proportional hazards model with the preoperative covariates.

Risk factors for revision

Finally, we conducted an analysis of the influence of all risk factors and treatment in the risk of revision, using the Cox proportional hazards model. Figure 5.5 shows a box plot of the hazard ratios (HR) of the different covariates and the treatment for septic revision. Note that a hazard ratio above (respectively, below) 1 indicates that the risk factor increases (respectively, decreases) the risk of revision. Statistical significance is achieved if the confidence interval does not contain 1. The associated numerical information can be found in Table 5.3. ALBC was associated with a protective effect over infection, having a hazard ratio (and confidence intervals) of 0.53 (0.44, 0.63) (i.e., knees with antibiotic cement had 47% less risk of revision, compared to those without antibiotic). Other covariates, such as alcohol abuse, rheumatoid arthritis, obesity, diabetes, and surgery year, were associated with a higher rate of infection (although diabetes was not statistically significant by a small margin). Hospital category and being female showed protective effects, and finally, smoking status, BMI, surgery duration, and age showed small or statistically non-significant effects.

This same analysis was repeated for the endpoints of aseptic revision and all-cause revision. It is worth mentioning that in both cases, ALBC showed a statistically significant protective effect. In the case of the aseptic revision endpoint, the hazard ratio was 0.499 (0.452, 0.552), and in the case of the all-cause revision endpoint, it was 0.549 (0.504, 0.597). The tables with the numerical results can be found in Appendix C.

5.3.3 Discussion and conclusions of the correlational analysis

The most important finding of this study was that ALBC was associated with lower septic and aseptic revision rates after TKA and, thus, with higher prosthetic survival. The suggested mechanism for the protective effect of ALBC against infection is that the initial concentration of antibiotic released when performing the TKA would be enough to prevent bacterial biofilm formation (Belt et al., 2000; Hinarejos et al., 2013; Jämsen et al., 2009).

Table 5.3: Cox proportional hazards model results for revision for infection.

		Hazard Ratio	Event Count	No Event Count	p-value
Totals			658	22123	
Antibiotic	Plain Cement	ref	344	9312	<0.01
	ALBC	0.53 (0.44, 0.63)	314	12811	
Sex	Male	Ref	270	6587	<0.01
	Female	0.61 (0.52, 0.73)	388	15536	
Age		1.00 (0.99, 1.01)	71.54 (8.32)	72.08 (7.83)	n.s.
Hospital category	1	1.11 (1.19,1.03)	129	4408	0.01
	2	Ref	140	4029	
	3	1.22 (1.41,1.06)	208	6447	
	4	0.90 (0.84,0.97)	173	6921	
	5	0.82 (0.71,0.94)	8	318	
Surgery duration		1.00 (1.00, 1.01)	92.71 (31.74)	89.83 (18.09)	<0.01
Alcohol abuse	No	ref	644	21982	<0.01
	Yes	2.31 (1.35, 3.96)	14	141	
Diabetes	No	Ref	525	18480	<0.01
	Yes	1.25 (1.03, 1.51)	133	3643	
Obesity	No	ref	541	19486	<0.01
	Yes	1.61 (1.29, 2.00)	117	2637	
Rheumatoid arthritis	No	Ref	632	21645	<0.01
	Yes	1.92 (1.29, 2.84)	26	478	
Smoking status	Non smoker	ref	499	17912	ref
	Smoker	1.13 (0.83, 1.54)	51	1348	n.s.
	Former smoker	1.10 (0.88, 1.38)	108	2863	n.s.
BMI		1.01 (0.97, 1.03)	32.26 (5.16)	31.77 (4.74)	n.s.

The evidence of the benefits of using ALBC during TKA is inconclusive in the scientific literature, with some studies reaching conclusions in favor of it (Jameson et al., 2019; Jämsen et al., 2009), others against (Namba, Chen, et al., 2009; Tayton et al., 2016) and others finding no differences (Bohm et al., 2014; Hinarejos et al., 2013).

Our results align with those of the study with the largest sample size coming from a single database, the National Joint Registry from England and Wales (Jameson et al., 2019). In that work, the vast majority of analyzed TKAs (93%) belonged to the ALBC group, and that circumstance also happened in other studies whose results are aligned with ours (Jämsen et al., 2009). In our analysis, nevertheless, the distribution between groups was more balanced: 57.6% in the ALBC group and 42.4% in the plain cement group.

Among the studies that found negative or no effect of ALBC, some performed the analysis in a sequential way (Tayton et al., 2016), first doing a univariate analysis and then doing a multivariate analysis with factors that had achieved statistical significance in the first step. On the contrary, our analysis was carried out in a multivariate manner with all the variables in a single step. Some works may have suffered from confounding bias (Parvizi, Cavanaugh, and Diaz-Ledezma, 2013), as patients in the ALBC group had significantly higher values of risk factors, such as diabetes mellitus or ASA grade. This was not the case in our registry. Finally, some other studies used antibiotics such as erythromycin and colistin (Hinarejos et al., 2013), which are less routinely used than gentamicin, which is predominant in our database.

The meta-analyses that have been published on the topic to date have shown, in general, no statistically significant differences in infection rates between groups (King et al., 2018; H.-Q. Li et al., 2022; T. Leta et al., 2024). Nevertheless, in the largest and most recent multi-registry meta-analysis found in the literature (T. Leta et al., 2024), almost half of the registries showed results aligned with ours' (i.e., in favor of ALBC), and two of them achieved statistical significance. In another meta-analysis (H.-Q. Li et al., 2022), no statistically significant differences between groups were found, although it can be highlighted that the two largest studies included did report them, both in favor of ALBC. Most of the largest studies analyzed in that meta-analysis have already been discussed in this section.

A secondary finding in our study was that the ALBC group also showed lower aseptic revision rates and higher prosthetic survival than the plain cement group. These results were aligned with the findings of several works in the literature (Bendich et al., 2020; Bohm et al., 2014; Jameson et al., 2019), with the plausible explanation that ALBC could have acted as a protective factor against subclinical infections misclassified as aseptic revisions. Other studies found no effect or a negative one, suggesting that it could be caused by the worsening of the mechanical properties of the prostheses due to the ALBC. The evidence for that mechanism was found mostly in vitro (Lautenschlager et al., 1976; Moran, Greenwald, and Matejczyk, 1979), and we did not observe it in our data. Some other works indicate that wrongly classified subclinical infections could bias the results of aseptic revision rates (Bozzo et al., 2022; Maathuis et al., 2005).

Previous works have also analyzed the associated risk of preoperative characteristics (risk factors) with prosthetic survival (Jämsen et al., 2009; Rand et al., 2003; Randelli et al., 2010). In our data, sex (being male) was found to be a risk factor for infection after TKA, in line with other works (Hinarejos et al., 2013; Kurtz et al., 2010; Namba, Inacio, and Paxton, 2013). Alcohol abuse, diabetes, obesity, and rheumatoid arthritis were also risk factors for infection. Being a smoker or a former smoker was also associated with a higher risk of infection, although this variable did not achieve statistical significance. The results on all these risk factors were in line with those of the largest meta-analysis carried out to date (Resende et al., 2021), except for the fact that we did not find age to be a protective factor.

Our correlational analysis had limitations. First, RACat data completeness varied for each hospital. We tried to overcome this by analyzing TKAs performed in hospitals that had at least an 80% of prosthetic revision reporting rate, but the limitation was still present. Second, septic revision diagnostic categories of RACat were used to identify PJI. Infections that were not treated surgically, whether superficial or deep, were not identified by this method. Nevertheless, this limitation is shared with other studies, making results comparable in principle. Finally, inputting missing values of some of the confounders or the effect of other unexplored confounders (such as surgical time, individual surgeons, individual hospitals, or others) could have introduced biases in the results.

After this discussion of the results of the correlational analysis, we continue to the causal analysis section.

5.4 Causal analysis

We divide this part again into methods, results, and conclusions.

5.4.1 Methods of the causal analysis

We start this section by providing some definitions. Some of them have already been introduced in the correlational section, but some others are new. Thus, in an ideal *population*, let A be the treatment, antibiotic-loaded bone cement use, or plain cement use, expressed as a binary variable that takes values 1 or 0, respectively; let T be the time passed between surgery and event dates, a continuous time variable measured in months; and let X be a vector of confounders. Then, given a time horizon h and using the *do-calculus* notation, we define the average treatment effect (ATE) of ALBC usage on prosthetic survival as $\psi^h = \mathbb{P}[T > h|do(A = 1)] - \mathbb{P}[T > h|do(A = 0)]$. Under the usual identifiability conditions of positivity, consistency, and exchangeability, ψ^h can be expressed in terms of the observed variables, dropping the *do*-operator, as $\psi^h = \mathbb{P}[T > h|A = 1, X] - \mathbb{P}[T > h|A = 0, X]$. Nevertheless, a problem of survival analysis is that in our data *sample*, we do not always get to measure T_i for each unit, as some units might be censored. Thus, for our sample we define the censoring time C_i as the time at which the i-th unit gets censored, together with Δ_i, a censoring indicator such that $\Delta_i = 1\{T_i > C_i\}$, and $U_i = T_i \wedge C_i$ (with $\wedge$ the logical *and* operator). The goal, then, is to estimate ψ^h using U_i and Δ_i instead of T_i. Similarly, note that one of the secondary goals of this section is to analyze the effect of ALBC on prosthetic survival in specific subgroups of the population, defined by the confounders. Thus, we can define the conditional average treatment effect (CATE) for the j-th confounder as $\pi^h(x^j) = \mathbb{P}[T > h|do(A = 1), X^j = x^j] - \mathbb{P}[T > h|do(A = 0), X^j = x^j]$, and then follow an analogous logic as with the ATE. We follow the notation and definitions of Cui et al., 2023, and thus we refer to that work for a detailed explanation of the implementation of the CATE and ATE estimators. The provided definitions of the ATE and the CATE imply that they represent the difference in survival probability between the ALBC group and the plain cement group (in the case of the ATE, in the whole population; in the case of the CATE, in a particular subgroup defined by the values of a given confounder). The sign of those quantities will indicate whether the treatment with ALBC increases (when positive sign) or decreases (when negative sign) the survival probability of the prosthesis and the value will indicate the size of such effect. Note that in this section, we focus by default on the outcome event of all-cause revision.

One of the identifiability conditions mentioned in the previous paragraph is exchangeability, which states that the outcome must be independent of the treatment given the confounders, $T \perp A|X$. This assumption is connected to the DAG of the problem, as the DAG allows us to identify the set of variables that we need to control for. For this reason, we employ DAGs for representing the causal structure of our scenario and the *do*-calculus rules for assessing the identifiability of the query of interest. We used the software CausalFusion (*CausalFusion* 2024) for DAG-related calculations. Note that this software also allows for the inclusion of selection bias nodes. Selection bias happens when some individuals are more likely to be present in the dataset than others based on their particular value or values of some variable or set of variables.

For estimation, we used causal survival forests (CSF) (Cui et al., 2023), a method that employs random forests for heterogeneous treatment effect estimation in survival settings, where outcomes can be right-censored. The method is based on orthogonal estimating equations for robustly adjusting for censoring and confounding (for further reference, see the original work). Several characteristics make it the optimal choice for our particular use case: it is specifically developed for survival problems; it is one of the few methods that explicitly accounts for censoring (by modeling the censoring process) and introduces a robust and data-efficient correction for it; it is based on random forests, which are a powerful tool for modeling expectation functions and are well known for their flexibility, achieved by imposing minimum assumptions on the underlying distributions of the data; and finally, it shows top performance for heterogeneous treatment effect estimation, which allows us to effectively estimate the CATE of each confounder.

We employed the R statistical language (version 4.2.3) for the analysis of the data and the GRF package *GRF* 2024 (version 2.3.0) for the implementation of the CSF. The default settings of the package were used, except for the *mtry* parameter of the random survival forests that model the censoring probabilities and the treatment probabilities, which were set to 3. The default value was the number of features, i.e., in our case, 13. A typical heuristic recommendation is to use the closest integer to the square root of the number of features, which in our case was 3. The use of the default value resulted in some extreme numbers in the censoring and treatment probabilities for medium and long horizon times ($h > 20$ months), and that, in turn, resulted in numerically unstable estimates of the ATE caused by the functional form of the CSF ATE estimator that contains probabilities in denominators. This problem most likely originated from the imbalance between censored and observed cases in our dataset.

Regarding estimators, other alternatives were available, but all of them showed theoretical downsides. One of the simplest options was to use the weighted Kaplan-Meier non-parametric estimator (Zare et al., 2014) to obtain the survival curves under ALBC and plain cement treatments and then compute the ATE by subtracting both curves. Another option was to employ the G-computation formula (Naimi, Stephen R Cole, and Edward H Kennedy, 2017), using a model of choice. None of these methods is doubly robust, which increases the risk of model misspecification-induced bias, nor is it optimized for estimating heterogeneous treatment effects, which increases the risk of bias when computing the CATE. Other doubly-robust estimators for survival analysis exist (J. Wang, 2018), but they would still be biased in the presence of non-random censoring.

Once the methods have been defined, we continue to the section on experiments and results.

5.4.2 Experiments and results of the causal analysis

In this section, we present the experiments and results of our study, the first explicitly causal analysis of the question of interest in the literature to the best of our knowledge. We start by providing the DAG that was obtained in collaboration with experts in the field (knee surgeons). Then, we present the results of the ATE for different time horizons and the results of the CATE for a fixed time horizon. Finally, we describe the experiments and the results of the assessment of the DAG structure and the estimator choice.

DAG of the problem

Figure 5.6 presents a directed acyclic graph (DAG) obtained in collaboration with knee surgeons, representing the qualitative expert knowledge of the causal relationships of the variables of the problem. The treatment is represented by A, and the outcome is represented by T. The rest of the variables are confounders, i.e., they affect both the treatment and the outcome. Causal relationships among confounders are represented too. Due to the case inclusion criteria (explained in section 5.2), the hospital category and the surgery year could have induced selection bias, and this is represented in the DAG by means of a selection bias node (explained in section 5.4.1), and represented by V, a binary variable. Given that the identifiability conditions hold, the causal effect of using ALBC on prosthetic survival, given the confounders, is recoverable from the observational data distribution. We present the proof using the rules of *do*-calculus, computed with the *CausalFusion* tool.

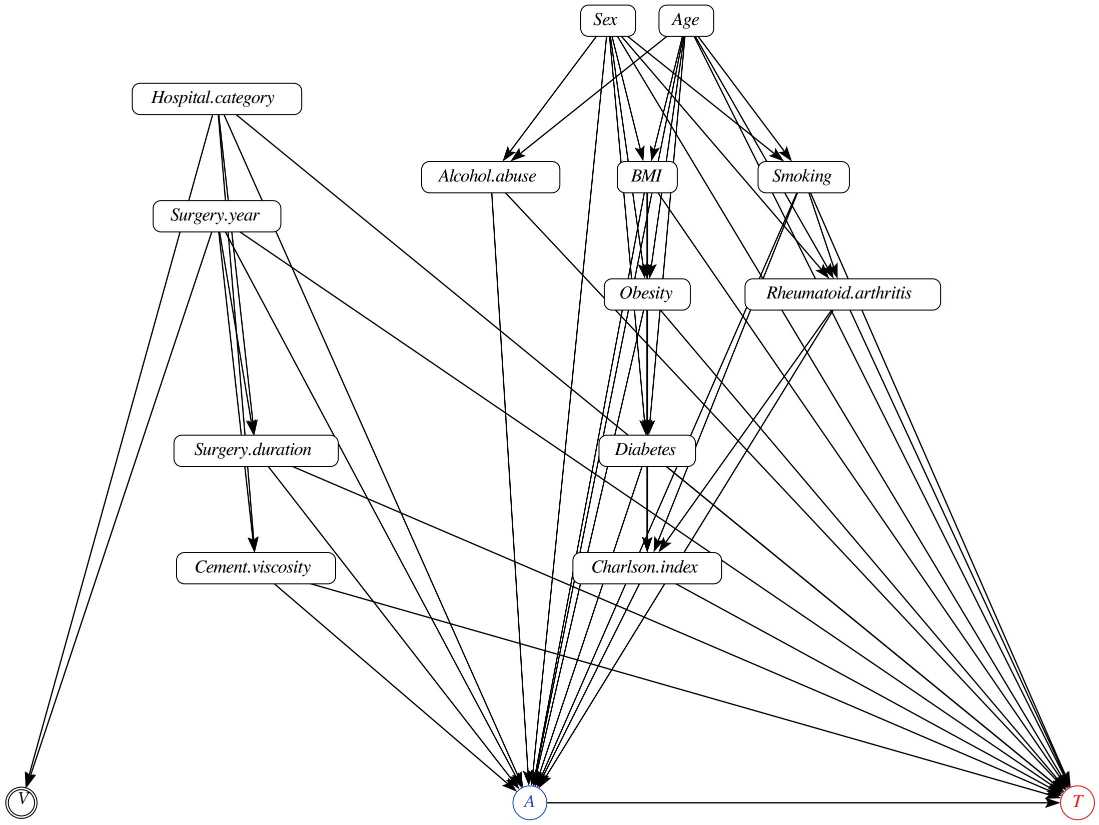


Figure 5.6: Directed acyclic graph of the problem. A is the treatment variable, which is a binary variable indicating the use of ALBC or plain cement. T is the outcome variable, indicating the survival time of the prostheses. The rest of the variables are confounders. After controlling for the confounders, no biasing paths are open, and the treatment effect is estimable based on the observational data.

We show that the causal effect $do(A = 1)$ on T, given X containing the variables {Age, Sex, Hospital category, Surgery year, Cement viscosity, Surgery duration, Obesity, Diabetes, Rheumatoid arthritis, BMI, Charlson index, Smoking, Alcohol abuse}, written as $P(T|do(A = 1), X)$ is recoverable from the distribution of observed variables with selection bias $P(A, T, X|V = 1)$.

Proposition 1. *The causal effect of A on T given $\{X\}$ is recoverable from $P(A, T, X|V = 1)$ and* *is given by the formula*

$$P\left(T|do(A=1),X\right)=P\left(T|A,X,V=1\right)$$

Proof.

$$P\left(T|do(A=1),X\right) \tag{5.1}$$
$$=P\left(T|A,X\right) \tag{5.2}$$
$$=P\left(T|A,X,V=1\right) \tag{5.3}$$

Eq. (5.2) follows from the second rule of do-calculus with the independence $(A\perp T|X)_{G_{\underline{A}}}$ Eq. (5.3) follows from the first rule of do-calculus with the independence $(S\perp T|A,X)$

Finally, we get

$$P\left(T|do(A=1),X\right)=P\left(T|A,X,V=1\right) \tag{5.4}$$

□

Average treatment effect over horizon time

In this part we present the main result of the analysis, i.e., the effect of using ALBC during total knee arthroplasty on prosthetic survival, compared to using plain bone cement. We selected the ATE as the measure to answer this question, which is the difference in survival probability between the group of arthroplasties with ALBC and the group of arthroplasties with plain cement at a particular horizon time. At a horizon of 120 months (10 years), this difference is of 0.08 $(0.0718, 0.0892)$, or 8 percentage points, in favor of the ALBC. This value clearly shows a positive effect of the treatment on prosthetic survival.

Figure 5.7 shows the ATE (green line, values in the left y axis) for different time horizons (x axis). Point estimates are provided with confidence intervals. The effect is small for short time horizons and gradually increases, reaching a 0.05 difference after 50 months, 0.075 around 100 months, and a maximum value above 0.08 around 130 months. This means that using ALBC increases the prosthetic survival probability by more than 8% after 120 months. We consider that the drop that is observed after 130 months is not a real effect but a result of the numerical instabilities of the employed method when the time horizon is too large for the characteristics of the dataset. To justify this, we depict the number of unstable points (red line, values in the right y axis) for each time horizon. An unstable point or estimate is defined as a data point in the dataset where the censoring probability, the treatment propensity, or both are bigger than 0.95 or smaller than 0.05. As can be observed, after 130 months, the number of unstable points increases noticeably.

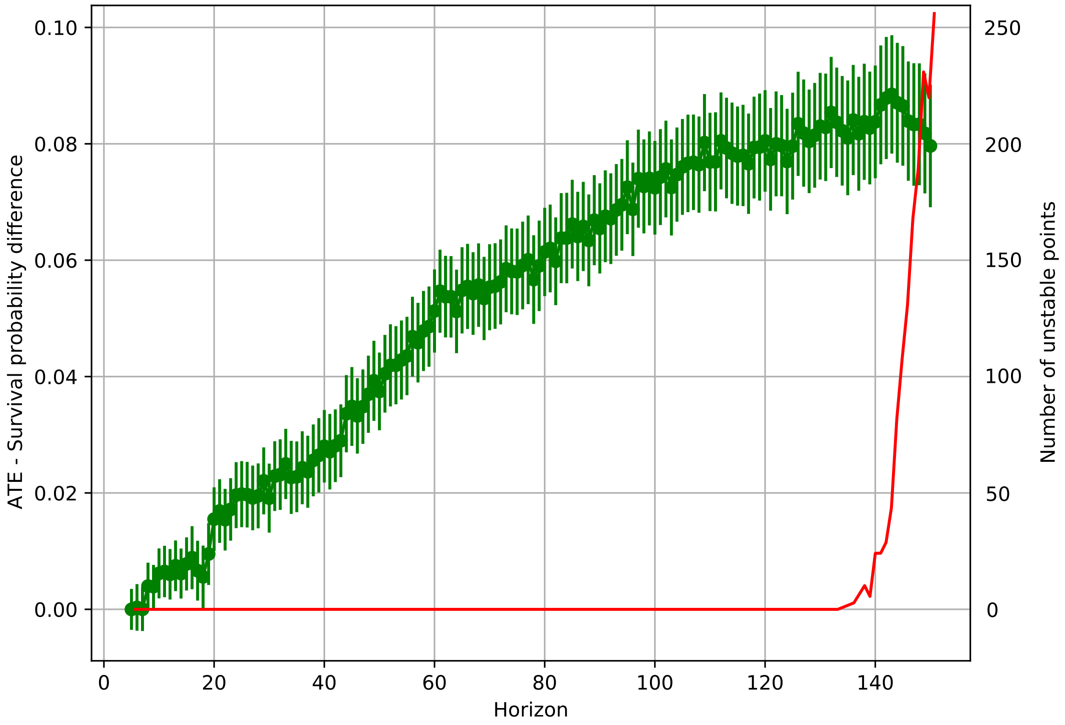


Figure 5.7: In green (left y axis), the average treatment effect, as the difference in prosthetic survival probability between antibiotic-loaded bone cemented arthroplasties and plain cemented arthroplasties, along horizon time. In red, (right y axis) the number of unstable data points over time. An unstable data point is an observation in the dataset that has been estimated to have extreme values of treatment and/or censoring probabilities. Unstable data points produce errors in the ATE.

Conditional average treatment effect (CATE) over horizon time

We also performed a CATE analysis of each confounder. For doing so, we computed the CATE for different values of each confounder and different time horizons up to 120 months. Recall that the CATE is equivalent to the ATE but in a subpopulation defined by the value of a particular confounder. For simplicity, we only report the values of the CATEs for the most relevant confounders at 120 months of horizon time in Table 5.4, but figures with CATE values along the whole horizon range and confounder categories/ranges can be found in Appendix D.

The ALBC treatment had a bigger positive effect on prosthetic survival probability for male patients (0.082) than for female patients (0.076). Similarly, the effect is, in general, bigger for younger patients than for older ones. The treatment effect is also bigger for patients with comorbidities than for those without, in the cases of obesity (0.083 vs. 0.077) and rheumatoid arthritis (0.086 vs. 0.077). The same is observed for factors that are considered to increase the risk of prosthetic revision (in particular for infection), such as smoking (0.076, 0.081, 0.088 for non-smokers, former smokers, and smokers, respectively) and alcohol abuse (0.077 vs. 0.095 for non-abusers and abusers, respectively). In the case of diabetes, no significant difference was observed (0.077 vs. 0.078).

Table 5.4: Table of CATEs of the most relevant confounders, at horizon time $h = 120$ months. Highlighted with green (respectively, red) are CATEs whose confidence intervals are above (respectively, below) and do not overlap with the confidence intervals of the ATE at $h = 120$.

		CATE	95% CI
Sex	Male	0.082	(0.0811, 0.0835)
	Female	0.076	(0.0747, 0.0764)
Obesity	No	0.077	(0.0759, 0.0778)
	Yes	0.083	(0.0817, 0.0841)
Rheumatoid arthritis	No	0.077	(0.0764, 0.0784)
	Yes	0.086	(0.0848, 0.0874)
Diabetis	No	0.077	(0.0765, 0.0784)
	Yes	0.078	(0.0771, 0.0793)
Smoking status	Non-smoker	0.076	(0.0752, 0.0771)
	Former smoker	0.081	(0.0868, 0.0895)
	Smoker	0.088	(0.0804, 0.0825)
Alcohol abuse	No	0.077	(0.0765, 0.0784)
	Yes	0.095	(0.0933, 0.0963)
Viscosity	None	0.055	(0.0547, 0.0562)
	Low	0.055	(0.0546, 0.0563)
	Medium	0.081	(0.0806, 0.0819)
	High	0.130	(0.1284, 0.1317)

In Table 5.4, we have highlighted with colors the cases of the confounders where the CATE differs most from that of the ATE at $h = 120$. In particular, we highlight with green (respectively, red) those CATEs whose confidence intervals are above (respectively, below) and do not overlap with the ones of the ATE. This way, we are effectively highlighting those segments of the population where the treatment is more (respectively, less) effective. Note that we have not included the age in this table to avoid excessive information, but the plot of this confounder can be found in Appendix D.

The relevance of the DAG

In this section, we report the results of an experiment designed to test the importance of the DAG. Suppose that the usage of ALBC or plain cement depended solely on the category of the hospital where the surgery was performed. This would be the case if, for instance, each hospital had a policy for the usage of ALBC or plain cement depending only on its category and not on any other confounder. Such a scenario would be reflected by the DAG in Figure 5.8. As can be seen, all arrows to the treatment have been erased, leaving only the one coming from the hospital category. Under this DAG, in order to obtain an unbiased estimate of the effect of ALBC on prosthetic survival, it is necessary that we control not only for the hospital category, which is the only variable directly affecting the treatment, but also for the surgery year, the surgery duration, and the cement viscosity. This is due to the existing relationships among those variables, which create biasing paths between them, the treatment, and the outcome. Note that without the usage of a DAG, it would be much more difficult to identify the need to control for those variables.

Under this hypothetical DAG scenario, we computed the ATE, and the results can be seen in Figure 5.9. The figure shows that both ATEs differ. In particular, the ATE computed under the hypothetical DAG 5.8 is smaller, i.e., the protective effect of the antibiotic is diminished with respect to the scenario under the real DAG 5.6. This is likely due to the fact that antibiotic protects those patients who have risk factors that increase the likelihood of revision, such as comorbidities, and, as we are taking those variables out of the confounding path, the size of the protective effect of antibiotics diminishes. This highlights even further the importance of using DAGs for coding assumptions of causal problems, as their structure has a direct impact in the magnitude of the estimated effects.

The relevance of the estimating method

Finally, in this section, we report the results of an experiment to assess the importance of the selected estimating method. As mentioned in Section 5.4.1, simpler methods for estimating ATEs and CATEs were available, but all were theoretically inferior (regarding performance) to CSF. We compared, in a specific example scenario, the results of CSF with the results of some of the alternatives to assess if the theoretical superiority of CSF had any practical, observable impact on the estimated values. Figure 5.10 shows the CATE for patients with obesity, computed with CSF and with two other alternative methods: a weighted Kaplan-Meier-based estimator of the CATE and a generalized linear model-based g-computation of the CATE.

The weighted Kaplan-Meier-based estimator of the CATE was obtained using the following steps:

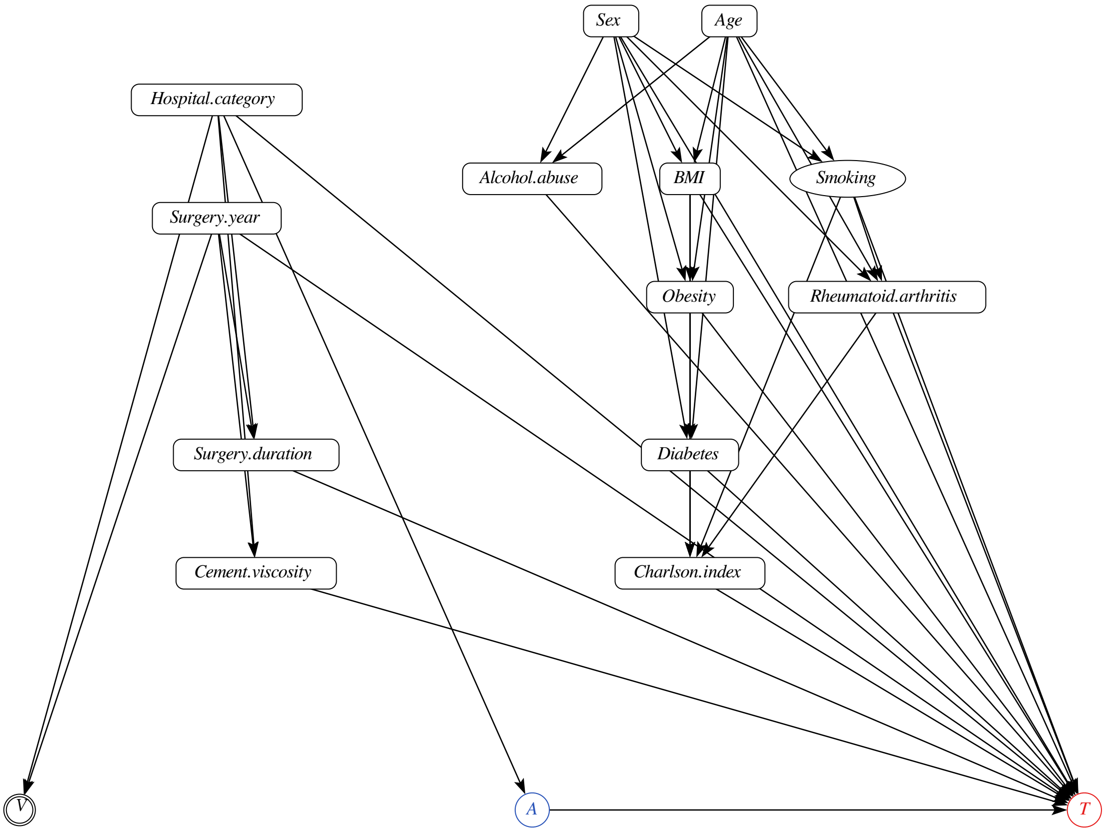


Figure 5.8: Alternative, hypothetical DAG of the problem, where the treatment is only influenced by the hospital category, constructed for assessing the impact of the DAG on the estimated ATE values.

first, confounder-based stabilized weights were computed. Then, weighted Kaplan-Meier estimates of survival were computed for the ALBC and plain cement groups for the subset of patients with obesity. Finally, those estimates were subtracted, obtaining the CATE. On the other hand, the model-based g-computation of the CATE was obtained by fitting a generalized linear model of the outcome, with the treatment and the confounders as predictor variables. Simulated values of survival were obtained for patients with obesity and both ALBC and plain cement treatments. Finally, the values were subtracted, obtaining the CATE.

The results of this experiment can be seen in Figure 5.10. The values obtained with the weighted Kaplan-Meier-based method differ from those obtained with CSF, and the values obtained with the G-computation method are similar to those obtained with CSF but seem to be artificially *softened*. This shows that the theoretical differences between the methods do indeed have a practical impact on the estimated values and justifies our choice of CSF.

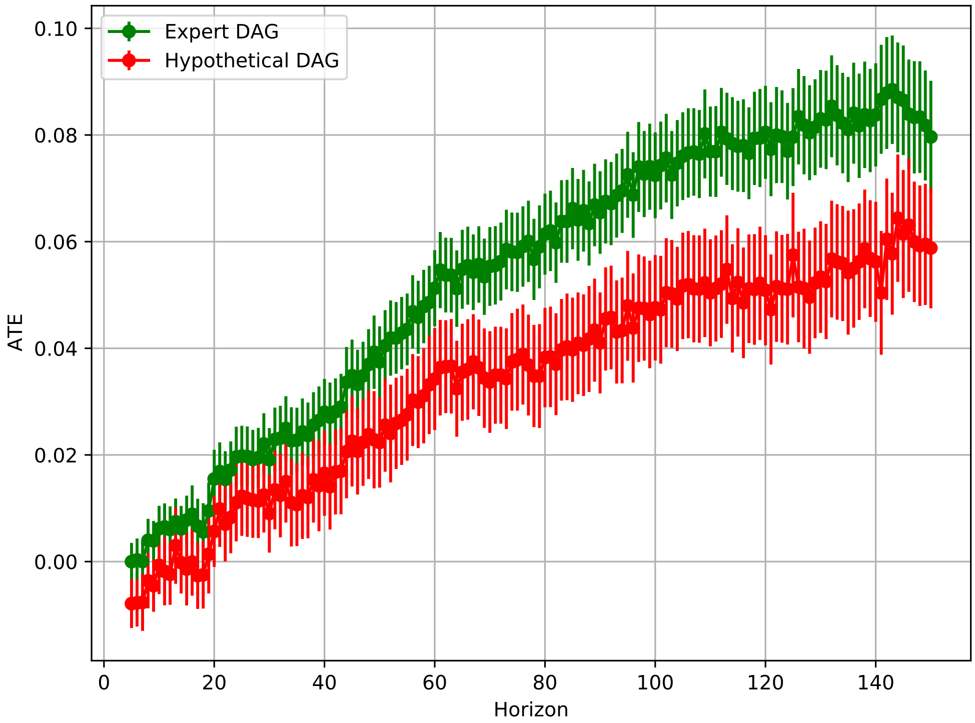


Figure 5.9: ATE estimates under different DAGs. The difference in the estimates shows the impact the DAG has and, thus, the importance of its correct specification.

5.4.3 Discussion and conclusions of the causal analysis

In this subsection, we present the conclusions of our causal analysis of the problem at hand. Recall that we have analyzed the effect of the use of ALBC versus plain cement on prosthetic survival during total knee arthroplasty in the presence of relevant confounders and in specific subgroups of the population. To do so, we have proposed a DAG that encodes expert knowledge about the relationships between the variables of the problem, and we have employed a machine learning-based, top-performing method for estimation. This method perfectly fitted the features of our problem: a survival analysis problem with right censoring and potentially heterogeneous treatment effects. To the best of our knowledge, this has been the first work with these characteristics in the literature. We have also shown through experiments the importance of the chosen methodology, both regarding the DAG and the estimation method. We believe these results contribute to the trend of treating clinical causal problems with causally explicit and tailored methods. In addition, our work constitutes one of the first successful usages of CSF in a real-world problem, together with the work in Inoue, Athey, and Tsugawa, 2023. Finally, we believe that our piece of evidence could

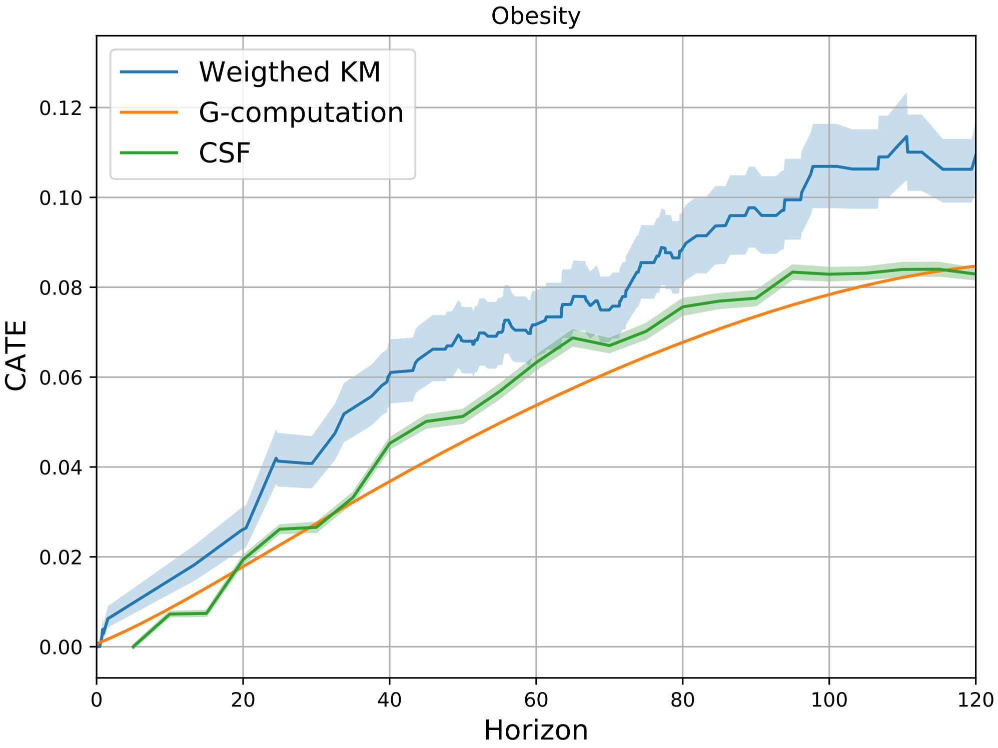


Figure 5.10: CATE for patients with obesity, computed with CSF and with two other alternative methods: a weighted Kaplan-Meier-based estimator and a generalized linear model-based g-computation estimator.

be integrated into future meta-analyses about the use of ALBC for prosthetic survival after total knee arthroplasty, which is a topic that remains open for discussion among the experts in the field.

Regarding results, we found that the use of ALBC vs. plain cement had a positive effect on overall prosthetic survival, which is in line with some previous works in the literature. In the long term, this effect was of 8% difference in survival probability. Regarding the CATE, we observed that patients with characteristics that were considered risk factors for prosthetic infection benefited more from using ALBC. This is explained by the fact that the ALBC protects against those risk factors, and thus, patients who have them get a bigger benefit. In particular, we observed bigger benefits for patients with obesity, rheumatoid arthritis, and patients who abuse alcohol, smoke, or used to smoke. This is aligned with works in the literature that identify all these factors as increasing the risk of prosthetic revision. Nevertheless, we could not directly numerically compare our results to those in other works in the literature due to the fundamental conceptual differences between risk factor analysis (usually done through the Cox proportional hazards model as in the correlational section) and CATE analysis.

The main potential limitation of this study is the one inherent to causal inference studies with observational data, i.e., the violation of some of the assumptions made, and in particular, the violation of the *exchangeability* or *no hidden confounder* assumption.

After finishing the conclusions of the causal analysis, we continue to the section on overall conclusions and comparison of approaches.

5.5 Overall conclusions and comparison of approaches

In this chapter, we conducted two separate studies to answer research question **Q3** *Does the use of antibiotic-loaded bone cement during total knee arthroplasty surgery increase the life of knee prostheses, compared with the alternative of using plain cement?* Each study employed the same observational data and two different approaches, a correlational one and a causal one. In this section, we compare the obtained results, and we qualitatively address research question **Q4** *What are the advantages and disadvantages, strengths and weaknesses, of correlational methods and causal inference methods for generating evidence about clinical interventions?* We do so in a dedicated subsection using the conducted correlational and causal studies as a base.

The main result that is quantitatively comparable between approaches is the effect of antibiotic-loaded bone cement on prosthetic survival for the outcome of all-cause revision. For the correlational approach, we can visually observe an approximation of this effect as the distance between curves of Figure 5.4, and for the causal approach, we can directly see it in Figure 5.7. For a time horizon of 120 months, the correlational method shows a difference of around 0.075 percentage points in survival in favor of antibiotic-loaded cement, very similar to the effect estimated by the causal approach ($\tilde{0}$.08 percentage points). In addition, in both cases, the temporal evolution of the estimation seems to follow a similar pattern, increasing steadily with the horizon time. Nevertheless, note that the Kaplan-Meier estimator does not *discount* the effect (either positive or negative) of other factors on prosthetic survival, and thus its estimation is of the overall survival, not of the isolated effect of the antibiotic. Another set of results that are comparable, although indirectly, are risk factors in the correlational approach and CATEs in the causal approach. These results can only be compared indirectly because the quantities they measure (the hazard ratios in the correlational analysis and the conditional ATEs in the causal analysis) are not exactly the same. Results can be found in Figure 5.5 and Table C.2 for the correlational analysis, and in Table 5.4 for the causal analysis. In general, the risk factor analysis shows that comorbidities, age (being older), and sex (being male) increase the risk of revision. Those are exactly the values of the confounders that show a bigger positive effect for antibiotic-loaded bone cement in the CATE analysis. Despite measuring different things, both phenomena have a similar physical interpretation.

One of the main qualitative differences between the results of both approaches is the clear causal interpretation of the results of the causal approach versus the somewhat fuzzier interpretation of the results of the correlational approach. As an example (already mentioned in the previous paragraph), note that the Kaplan-Meier curves show a difference in prosthetic survival between antibiotic and plain cement groups, but as the Kaplan-Meier estimator does not take into account other factors, it is not possible to conclude that the observed differences are caused by the treatment. Admittedly, this can be improved by using the weighted Kaplan-Meier estimator or constructing survival curves from the conditional hazards of the Cox model, but those approaches still do not make causal assumptions explicit. Furthermore, if we would be interested in obtaining estimates of individual treatment effects (i.e., the effect of the treatment for a given particular individual with their covariates), the employed causal estimator is a better option, as it has been optimized exactly for that purpose. The alternative of the correlational approach, i.e., the Cox model, makes stronger implicit assumptions about hazard functions, does not take into account censoring, is not doubly robust, and would perform worse at the aforementioned task.

After this quantitative and qualitative comparison of the results, we assess and compare the approaches in the next section, outlining their strengths and weaknesses.

5.5.1 Critical assessment of the employed methodologies

In this section, we aim to answer research question **Q4** (*What are the advantages and disadvantages, strengths and weaknesses, of correlational methods and causal inference methods for generating evidence about clinical interventions?*) by presenting a direct comparison of both approaches.

One of the main strengths of the correlational approach and methods is that they are very widespread in the related literature. That allows for direct comparison of results and integration of evidence in meta-analyses and facilitates the understanding of conclusions by other peers. Somehow, it establishes a *common language* for the topic. Furthermore, the employed methods (Kaplan-Meier and Cox proportional hazards model) are statistically sound methodologies whose limitations (i.e., sources of bias, underlying assumptions, etc.) are well-identified and known. Note that this is not necessarily the case with all correlational approaches, as, for example, there is some evidence about statistical problems of the more modern but less used Fine and Gray's Subdistribution method (Austin et al., 2022; Bonneville, Wreede, and Putter, 2024). Finally, it must be mentioned that the interpretation of the results of the employed methods regarding survival and hazard ratios is quite straightforward.

Among the weaknesses of the correlational method, besides the already discussed lack of causal interpretation of the results, we highlight the total blindness of the correlational approach to the causal structure of the problem. With this approach, it is implied that all confounders (usually called *risk factors* in the literature) have exactly the same relationship with the treatment and the outcome: they affect both. If the reality has a different causal structure (for instance, with some of the confounders affecting other confounders but not directly the treatment), and if that structure is (at least partially) known by the researchers, that knowledge remains unexploited, and that is a source of bias in the estimates (recall the results of section 5.4.2, about the importance of the DAG).

Regarding the causal approach, a minor but still somewhat important weakness is that, unlike the correlational approach, it is not a *common language* in the literature. This results in a need for longer preambles and contexts, more detailed definitions, and difficulties for direct comparisons of results. In addition, the existence of a relatively large amount of available options for estimators implies more complex decisions, and their implementation and/or use is often more difficult than that of simpler models typically employed in correlational approaches.

On the strengths' side, first and foremost, we find the clear causal meaning of the provided results: antibiotic-loaded bone cement *causes* an increase of 8 percentage points in prosthetic survival under the assumptions made. The usage of the DAG explicitly displays the researchers' beliefs about the causal relations of the variables of the problem and allows for more transparent discussions. Furthermore, the existence of advanced estimators such as the one employed in this work (a doubly robust method that models the censoring process and uses random forests) implies a superior estimation performance of causal methods in comparison with the performance of the typical estimators of the correlational approach.

We already stated that treating causal problems with associational methods constitutes an epistemic limitation, and the evidence gathered in this chapter shows that it is also a practical one, as we have shown that the identification and estimation choices impact the obtained results. We believe that this comparison constitutes empirical proof that, whenever possible, causal methods should be employed to answer causal questions. Correlational-only, causality-free approaches and methods have been the *common language* for this type of study, but we have better options nowadays. Regarding the limitations of this approach-like and methodological comparison, there are two main factors to mention: on the one hand, it is based on a single use case, which may limit the representativeness of the conclusions, and on the other hand, it is qualitative, when a quantitative approach could constitute stronger evidence.

In the next chapter, we delve into the realm of multivalued treatments, and we explore the generalizability of algorithms from binary treatment to multivalued treatment settings.

Chapter 6

Multivalued treatment settings and a neural network-based causal inference algorithm: Hydranet

In this chapter, we provide an answer to research question **Q5**, *Can we generalize advanced causal inference algorithms from binary treatment settings to multivalued treatment settings?*

This work is the final version of previous works presented as oral communication in the NeurIPS 2022 Workshop on Causality for Real-world Impact (”Borja Velasco, Jesus Cerquides, & Josep Lluis Arcos (2022). Multi-valued Treatment Effect Estimation for Health Technology Assessment with a Neural Network.”) and the article ”Borja Velasco-Regulez, & Jesus Cerquides (2023). Hydranet: A Neural Network for the Estimation of Multi-Valued Treatment Effects. *Artificial Intelligence Research and Development (pp 16–27). IOS Press.* DOI: 10.3233/FAIA230655”, published in the proceedings of the Catalan Conference of Artificial Intelligence of 2023, where it won the best paper award of the conference.

6.1 Background

In the previous chapters of this thesis, we have assessed two different health technologies in three different scenarios. All the scenarios had in common one feature: the analyzed treatment of interest was binary, i.e., it could take two values or categories. This is not a coincidence: a big share of the literature on health technology assessment, clinical studies, statistical methods, and causal inference follows a binary treatment logic. This is natural: when aiming at doing a task as complicated as *understanding how the world works*, understanding the effect of an intervention that can take only two values (or, often, the effect of the *presence* or *absence* of the intervention) is the first logical approach to the problem, as it is the most elemental simplification of it. Nevertheless, as true as this is, it is also true that reality is usually much more complex than the simplification we humans make in order to understand it. In our particular case of interest, this means that treatments or interventions in the field of health are more often than not multivalued, i.e., they can take more than just two values. For example, consider how the question about the effects of the COVID-19 vaccine, which is originally binary (vaccine administered or not administered), can become multi-valued if we want to know the effects of the different vaccine brands. Similarly, the question about the effects of antibiotic-loaded bone cement on prosthetic survival becomes multi-valued if we wonder about the effects of different antibiotics.

When searching in the literature for causal inference algorithms for multivalued treatments, we realized two facts: first, there is an important body of work that connects causal inference with machine learning. This is partially due to the fact that a big share of important contributions to the field of causal inference have come from authors working in the field of computer science, and computer science is the base of the current golden era of machine learning and deep learning. Second, that most works, in the form of machine learning-based or deep learning-based algorithms for causal inference, are being developed and tested in binary treatment settings. This is natural, but it is also a limitation that hinders researchers' adoption of these types of algorithms. For these reasons, in this chapter, we aim to answer research question **Q5** *Can we generalize advanced causal inference algorithms from binary treatment settings to multivalued treatment settings?* We start by over-viewing the existing algorithms, and we select a top-performing, representative use case and test its generalizability and performance in a multi-valued treatment setting.

The rest of this chapter is divided as follows. In the next subsection, we provide an overview of neural network or machine learning-based causal inference algorithms. Then, we dedicate a section to formally express our problem and provide definitions. After that, we present the selected binary treatment algorithm, and we derive its multivalued treatment version. Finally, the next parts contain the strategy for generating data for experiments, the results of those experiments, and the discussion and conclusions.

6.1.1 Neural network-based causal inference algorithms

Machine learning and neural networks are becoming a common choice for performing causal analysis tasks (causal inference, causal discovery) due to their power and flexibility for modeling complex functions, especially when the dimensionality of the data is high (Miguel A Hernán and James M Robins, 2020). Several authors have investigated specific network architectures, loss functions, regularization methods, etc., to tackle the task of inferring causal quantities using neural networks. Here, we review them, some of which were already introduced in the state-of-the-art chapter. Yuan, Ding, and Bar-Joseph (2020) propose a convolutional neural network for causal inference through a method that encodes observational data of a causal problem in an image-like matrix. Louizos et al. (2017) introduce a variational auto-encoder architecture for the estimation of treatment effects at the patient level, mapping proxies of unmeasured confounders to latent variables, exploiting the strengths of auto-encoders with latent variables. Yoon, Jordon, and Van Der Schaar (2018) use generative adversarial networks to learn counterfactuals of a causal inference problem. Shalit, Johansson, and Sontag (2017) propose a neural network that learns a representation of the covariates, and that has two different ”heads” or ends, one for each treatment option. The architecture and training strategy ensure a good trade-off between sharing statistical power in the representation layers and learning the effects of each treatment value (binary) in the ”heads.” Finally, Shi, Blei, and Veitch (2019) presents another architecture named Dragonnet, inspired in the previously explained work (the one by Shalit, Johansson, and Sontag (2017)), which includes another ”head” for learning the propensity score. By defining the adequate loss function, it is ensured that the architecture exploits the sufficiency of the propensity score (Rosenbaum and D. B. Rubin, 1983) for adjustment. All these algorithms have been developed for binary treatment settings. Among the few that we found that could handle multivalued treatments, there were significant differences. Kaddour, Y. Zhu, et al. (2021) present a neural net-based algorithm, but for working with specific data morphologies and *structured treatments*, such as graphs, images, texts, etc. Schwab, Linhardt, and Karlen (2019) use support vector machines, and present one of the most complete works with multivalued treatments. Finally, Künzel et al. (2019) use *meta-learners*, i.e., aggregators that combine the outputs of individual algorithms, and we consider that such an approach is fundamentally different from the one presented in this chapter.

The first step to tackle the problem at hand, i.e., the generalization of a neural network-based causal inference algorithm from binary to multivalued treatment settings, is to formally define it. Thus, in the next section, we provide the required definitions of variables and quantities.

6.2 Problem statement

Consider a treatment of interest represented as a discrete random variable $T \in [0..k]$, capable of taking $k+1$ different values. The outcome, denoted by Y, is a continuous random variable in $\mathbb{R}$. Additionally, let the covariates—variables that influence both the treatment and the outcome—be represented by a random vector $X \in \mathbb{R}^j$. Thus, each data point in our observational dataset is represented as a tuple $(Y_i, T_i, X_i),\ i \in [1..N]$. These data points are assumed to be generated independently and identically. This set of data points constitutes the body of observational data. Let the causal effect of the treatment t over the outcome Y be $\mu_t = \mathbb{E}[Y|do(T=t)]$, using Pearl's *do*-calculus notation (Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell, 2016), which denotes intervention. Given that the identifiability conditions hold (positivity, consistency, and "no hidden confounder"), $\mu_t = \mathbb{E}[Y|X=x, T=t]$, which is a quantity that can be inferred from the body of observational data. Along the rest of the section, it is assumed that the identifiability conditions are fulfilled.

Let the conditional outcome be defined as the expectation of the outcome given the treatment and the covariates, $Q(t,x) = \mathbb{E}[Y|t,x]$. Based on Q, a simple estimator $\hat{\mu}_t$ of μ_t as $\hat{\mu}_t = \frac{1}{N}\sum_i Q(t,x_i)$ can be constructed. In the following, the goal will be to approximate Q. Let $\hat{Q}$ be an approximation of Q, and let $\mu_t^{\hat{Q}} = \frac{1}{N}\sum_i \hat{Q}(t,x_i)$ be the estimator of μ_t obtained replacing Q by its estimation $\hat{Q}$. Furthermore, the Generalized Propensity Score (GPS (Cattaneo, 2010)) is expressed as $\mathbf{G}(x) = [g_0(x), g_1(x), \ldots, g_k(x)] \in \mathbb{R}^{k+1}$, with $g_t(x) = P(T=t|x)$.

In a binary treatment setting, under the identifiability conditions, the Average Treatment Effect (ATE) is one of the most common causal quantities of interest, and it is defined as $\psi = \mu_1 - \mu_0$. Given an approximation $\hat{Q}$ of Q, ψ can be easily estimated as $\psi^{\hat{Q}} = \mu_1^{\hat{Q}} - \mu_0^{\hat{Q}}$. In a multi-valued treatment setting, a wider class of causal quantities of interest can be defined, and all the conditional outcomes must be computed together in order to obtain valid estimates of those quantities (Cattaneo, 2010). In this work, such quantities of interest are defined as the pair-wise average differences between the several treatments and a treatment considered the control (note that, in practice, the control treatment does not necessarily mean the absence of treatment). Thus, a vector of ATEs $\boldsymbol{\psi} \in \mathbb{R}^k$, $\boldsymbol{\psi} = [\psi_1, \psi_2, \ldots, \psi_k]$, with $\psi_t = \mu_t - \mu_0$ is defined. These quantities can be approximated in a similar fashion as shown before, the t-th element of the vector being $\psi_t^{\hat{Q}} = \mu_t^{\hat{Q}} - \mu_0^{\hat{Q}}$. Note that if the causal quantity of interest was $\psi_{i,j} = \mu_i - \mu_j$, it could easily be computed based on the previous definition, as $\psi_{i,j} = \psi_i - \psi_j$, due to the linearity of the expectation operator.

This vector of ATEs $\boldsymbol{\psi}$ will be our causal quantity of interest. In the next section, the estimation method provided in Shi, Blei, and Veitch (2019), which has the objective of estimating the ATE in the binary case, is generalized to the estimation of $\boldsymbol{\psi}$ in the multivalued treatment case.

6.3 From Dragonnet to Hydranet

Dragonnet is a high-capacity, end-to-end neural network architecture for estimating binary treatment effects (Shi, Blei, and Veitch, 2019). It was inspired by a previous work (Shalit, Johansson, and Sontag, 2017) and performed better than its predecessor. The architecture and cost functions of Dragonnet provide top-performing results in benchmarking datasets by connecting neural network-based estimation with semiparametric theory and double machine learning, which have already been introduced in the state-of-the-art chapter. In the present chapter, we will re-encounter the concepts of efficient influence curves, score functions, and estimating equations (Edward H. Kennedy, 2016). Similarly, we will speak about double machine learning (Chernozhukov, Chetverikov, Demirer, Duflo, C. Hansen, and Newey, 2017; Chernozhukov, Chetverikov, Demirer, Duflo, and Hansen, 2018), by which estimators show desirable statistical properties in terms of data efficiency and convergence. One of those desirable properties is double robustness, which ensures that the estimator of the causal effect of interest converges to the correct value even when one of the employed models is wrongly specified.

In this section, we present the variation of the architecture, mathematical formulations, and proofs for adapting Dragonnet to a multivalued treatment setting. We call this adaptation Hydranet.

6.3.1 Architecture

The architecture of Hydranet can be seen in Figure 6.1. It consists of two parts: the representation part, formed by the input layer and two hidden layers, and the heads, formed by $k+2$ ends. Out of those, $k+1$ correspond to the conditional outcomes and are formed by two more hidden layers plus the output layer. The remaining head corresponds to the GPS, $\mathbf{G}(x) = [g_0(x), g_1(x), \ldots, g_k(x)] \in \mathbb{R}^{k+1}$, with $g_t(x) = P(T = t|x)$, consisting on just the output layer. All layers are fully connected. Recall that the t-th element of the vector of ATEs is approximated as $\psi_t^{\hat{Q}} = \frac{1}{N}\sum_i \hat{Q}(t, x_i) - \hat{Q}(0, x_i)$.

The baseline objective function has the shape

$$\hat{R}(\theta) = \frac{1}{N}\sum_i \left[(Q^{nn}(t_i, x_i; \theta) - y_i)^2 + \alpha \text{CrossEntropy}(g_t^{nn}(x_i; \theta), t_i)\right] \tag{6.1}$$

where the quadratic term relates to the errors of the potential outcomes' heads and the cross entropy term relates to the errors of the propensity score's head. The model parameters are

$$\hat{\theta} = \arg\min_\theta [\hat{R}(\theta)] \tag{6.2}$$

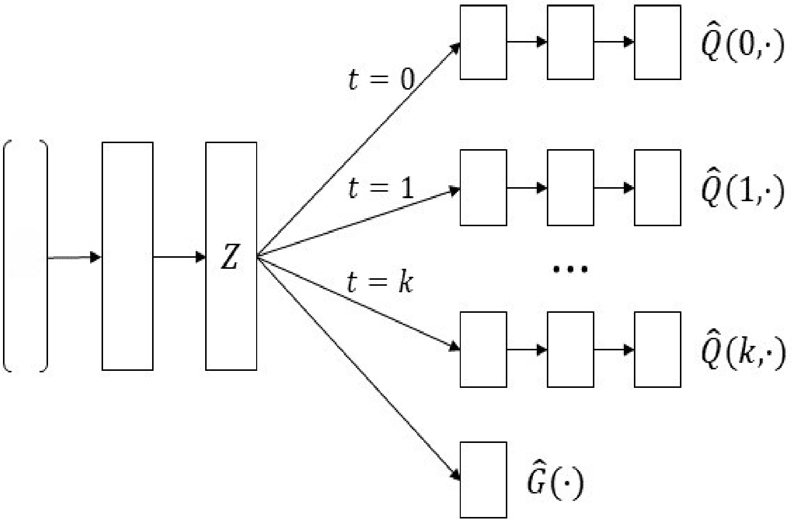


Figure 6.1: Hydranet architecture, where Z is the representation layer, and the $k+2$ heads correspond to the $k+1$ potential outcomes, $\hat{Q}(k,\cdot)$, and the Generalized Propensity Score, $\hat{G}(\cdot)$.

6.3.2 Targeted Regularization

Now, following the reasoning in Shi, Blei, and Veitch (2019), targeted regularization is presented. Targeted regularization is a modification of the objective function that introduces an extra parameter, epsilon. In this setting, $\boldsymbol{\epsilon}$ is a vector in $\mathbb{R}^k$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ...\epsilon_k)$, and the new objective function is

$$\bar{F}(\theta, \boldsymbol{\epsilon}) = \hat{R}(\theta) + \beta \frac{1}{N} \sum_i \gamma_i(y_i, t_i, x_i; \theta, \boldsymbol{\epsilon}), \text{ where} \tag{6.3}$$

$$\gamma_i(y_i, t_i, x_i; \theta, \boldsymbol{\epsilon}) = (y_i - \bar{Q}_i(\theta, \boldsymbol{\epsilon}))^2, \text{ and} \tag{6.4}$$

$$\bar{Q}_i(\theta, \boldsymbol{\epsilon}) = Q^{nn}(t_i, x_i) + \epsilon_1 \left(\frac{\mathbf{I}(T=1)}{g_1^{nn}(x_i)} - \frac{\mathbf{I}(T=0)}{g_0^{nn}(x_i)} \right) + \ldots + \epsilon_k \left(\frac{\mathbf{I}(T=k)}{g_k^{nn}(x_i)} - \frac{\mathbf{I}(T=0)}{g_0^{nn}(x_i)} \right), \tag{6.5}$$

with $\mathbf{I}(T=t)$ the indicator function, and thus the sought model parameters are defined by

$$\hat{\theta}, \hat{\epsilon} = \underset{\theta, \boldsymbol{\epsilon}}{\arg\min} [\hat{R}(\theta) + \beta \frac{1}{N} \sum_i \gamma_i(y_i, t_i, x_i; \theta, \boldsymbol{\epsilon})]. \tag{6.6}$$

The motivation for this modification lies in semiparametric estimation theory and targeted maximum likelihood estimation (TMLE) (Edward H. Kennedy, 2016; Lendle, 2015), both presented in Chapter 2 and in Section 6.3 of the current chapter. Generally, semiparametric estimation theory provides us with conditions that ensure desirable properties (Chernozhukov, Chetverikov, Demirer, Duflo, C. Hansen, Newey, and J. Robins, 2017) of the estimator ψ when they are fulfilled, and TMLE is an efficient method to achieve the fulfillment of those conditions. The conditions are the set of non-parametric estimating equations, defined as

$$\mathbf{0} = \left[\frac{1}{N}\sum_i \varphi_{i,1}, \frac{1}{N}\sum_i \varphi_{i,2}, ...\frac{1}{N}\sum_i \varphi_{i,k}\right], \tag{6.7}$$

and they employ the elements of the vector of efficient influence curves, defined as $\boldsymbol{\varphi} \in \mathbb{R}^k, \boldsymbol{\varphi} = [\varphi_1, \varphi_2, ...\varphi_k]$, with

$$\varphi_{i,t} = Q^{nn}(t, x_i) - Q^{nn}(0, x_i) + \left(\frac{\mathbf{I}(T=t)}{g_t^{nn}(x_i)} - \frac{\mathbf{I}(T=0)}{g_0^{nn}(x_i)}\right)(y_i - Q^{nn}(t, x_i)) - \psi_t. \tag{6.8}$$

Finally, recall that our goal is that the minimization of the modified objective function ensures the fulfillment of the non-parametric estimation equations. This can be expressed mathematically as

$$\mathbf{0} = \nabla\bar{F}\big|_{\hat{\boldsymbol{\epsilon}}} = \left[\frac{\partial\bar{F}}{\partial\epsilon_1}, \frac{\partial\bar{F}}{\partial\epsilon_2}, ...\frac{\partial\bar{F}}{\partial\epsilon_k}\right]\Bigg|_{\hat{\boldsymbol{\epsilon}}} = \left[\frac{\beta}{N}\sum_i \varphi_{i,1}, \frac{\beta}{N}\sum_i \varphi_{i,2}, ...\frac{\beta}{N}\sum_i \varphi_{i,k}\right]. \tag{6.9}$$

This warrants the aforementioned desirable properties of the estimator ψ, i.e., double robustness, fast convergence, and efficiency. Next, we provide the proof of Equation 6.9. The goal is to prove that

$$\frac{\partial\bar{F}}{\partial\epsilon_t}\bigg|_{\hat{\epsilon}_t} = \frac{1}{N}\sum_i \varphi_{i,t}, \qquad \forall\, t \text{ in}[0, k]. \tag{6.10}$$

Proof. On one hand, using equations 6.3, 6.4 and 6.5, get

$$\begin{aligned}
\frac{\partial\bar{F}}{\partial\epsilon_t}\bigg|_{\hat{\theta},\hat{\epsilon}_t} &= \frac{\partial}{\partial\epsilon_t}\left(\hat{R}(\theta) + \beta\frac{1}{N}\sum_i \gamma_i(y_i, t_i, x_i; \theta, \boldsymbol{\epsilon})\right)\Bigg|_{\hat{\theta},\hat{\epsilon}_t} \\
&= \frac{\beta}{N}\sum_i \frac{\partial}{\partial\epsilon_t}\gamma_i(\theta, \boldsymbol{\epsilon})\bigg|_{\hat{\theta},\hat{\epsilon}_t} \\
&= \frac{2\beta}{N}\sum_i (y_i - \bar{Q}_i(\theta, \boldsymbol{\epsilon}))\frac{\partial\bar{Q}_i(\theta, \boldsymbol{\epsilon})}{\partial\epsilon_t}\bigg|_{\hat{\theta},\hat{\epsilon}_t} \\
&= \frac{2\beta}{N}\sum_i \left[(y_i - \bar{Q}_i(\theta, \boldsymbol{\epsilon}))\left(\frac{\mathbf{I}(T=t)}{g_t^{nn}(\theta)} - \frac{\mathbf{I}(T=0)}{g_0^{nn}(\theta)}\right)\right]\Bigg|_{\hat{\theta},\hat{\epsilon}_t} \\
&= \frac{2\beta}{N}\sum_i \left[(y_i - \hat{Q}(t, x_i)\left(\frac{\mathbf{I}(T=t)}{\hat{g}_t} - \frac{\mathbf{I}(T=0)}{\hat{g}_0}\right)\right]\text{(evaluate at } \hat{\theta}, \hat{\epsilon}) \\
&= \frac{2\beta}{N}\sum_i \left(\hat{Q}(t, x_i) - \hat{Q}(0, x_i)\right) - \frac{\beta}{N}\sum_i \left(\hat{Q}(t, x_i) - \hat{Q}(0, x_i)\right) + \\
&\quad \frac{\beta}{N}\sum_i \left[(y_i - \hat{Q}(t, x_i))\left(\frac{\mathbf{I}(T=t)}{\hat{g}_t} - \frac{\mathbf{I}(T=0)}{\hat{g}_0}\right)\right]\text{(add and subtract term)} \\
&= \frac{2\beta}{N}\sum_i \left[\hat{Q}(t, x_i) - \hat{Q}(0, x_i) + (y_i - \hat{Q}(t, x_i))\left(\frac{\mathbf{I}(T=t)}{\hat{g}_t} - \frac{\mathbf{I}(T=0)}{\hat{g}_0}\right) - \hat{\psi}_t\right].
\end{aligned}$$

On the other hand, by substituting the definition of the efficient influence curves 6.8 in the set of non-parametric estimation equations 6.7, multiplying by β and particularizing at $\hat{Q}, \hat{g}, \hat{\psi}$ (the functions modeled by the neural network at the optimal point of the parameter space), an expression equal to the one in the last line of the proof is obtained. Thus, the non-parametric estimation equations 6.7 are satisfied, and the proof is complete. □

6.4 Data, metrics and experiments

Evaluation of the performance of causal inference algorithms usually runs into what is known as *the fundamental problem of causal inference*. This refers to the fact that for each individual or patient, we do not get to observe one or some of the potential outcomes, the counterfactuals. Then, because those are needed for computing *ground truth* effects, testing algorithms with real-world data is usually not possible. Thus, synthetic or semi-synthetic data is usually required, where the data-generating process is fully or at least partially under researchers' control, and we can simulate the required information to be able to calculate ground truth effects. Some datasets have been established as *de facto* benchmarks for comparisons. In this chapter, we have tested Hydranet in two datasets, a fully synthetic one and a semi-synthetic one, IHDP (Gross, 1993). We refer to them as the synthetic dataset (or SynD for short) and the IHDP dataset, respectively. In order to generate these datasets, algorithms mimicking different data-generating processes (DGP) have been designed and implemented. For the synthetic dataset, the covariates, treatments, and outcomes have been synthetically generated, taking inspiration from Kaddour, Y. Zhu, et al. (2021). For the IHDP dataset, the covariates are taken from a study with real participants, while the treatments and outcomes are synthetically generated. Those real covariates were collected for a Randomized Controlled Trial (RCT) carried out in 1985 (Gross, 1993) and are routinely used for benchmarking causal inference algorithms, usually following the configuration in Dorie et al. (2018). A similar strategy has been followed in the current work, but adapting the DGP to the present needs (a multi-valued treatment scenario). With both datasets, the number of treatments was set to 5. Additionally, we have also defined the metrics for the performance. In the remainder of this section, we provide a more detailed explanation of the data-generating process (DGP) and the metrics.

6.4.1 Synthetic data generating process

For generating fully synthetic data, DGPs with tunable parameters of bias size B, degree of positivity ρ, dataset size D, and number of confounders NC were designed. The number of treatments was set to 5. The potential covariates are vectors $\mathbf{x} \in \mathbb{R}^{30}$ with each element sampled from a uniform distribution $\mathcal{U}(-1, 1)$. The number of such vectors is equal to the data size parameter D, forming a matrix $\mathbf{X} \in \mathbb{R}^{Dx30}$. The actual confounders, i.e., the variables that participate in the determination of both the treatment and the outcome, are the first *NC* (number of confounders) elements of each covariate vector, thus forming a matrix $\mathbf{C} \in \mathbb{R}^{DxNC}$. The treatment for each data point was obtained in two steps. First, by squaring the confounder vector element-wise and summing the elements, applying a $min - max$ scaler to the range $[0, 4]$ (for 5 treatments), and rounding to the closest integer. Then, in order to fulfill the positivity condition, by drawing the final treatment value from a categorical distribution such that

$$p(t|\mathbf{c}) = \begin{cases} \rho, & \text{if } t = m(\mathbf{c}) \\ \frac{1-\rho}{k-1}, & \text{otherwise} \end{cases}$$

with $m(\cdot)$ the operation defined in the first step and ρ the degree of positivity. Note that with this definition, a value of $\rho = 0.5$ would mean perfect overlap, treatment assigned at random, while a value of $\rho = 1$ would mean the violation of the positivity condition. Finally, for computing the potential outcomes, three outcome functions $(l_a(t, \mathbf{x}), l_b(t, \mathbf{x}), l_c(t, \mathbf{x}))$ were defined, that map a combination of the covariates and the treatment to the output space. The outcome functions have the shape

$$l_a(t, \mathbf{x}) = 30\mathbf{v}_0^T\mathbf{x} + 10\, t^2\, \mathbf{v}_t^T\mathbf{x} + \epsilon$$
$$l_b(t, \mathbf{x}) = 20\mathbf{v}_0^T\mathbf{x} + 5\, B\, t\, \mathbf{v}_t^T\mathbf{x} + \epsilon$$
$$l_c(t, \mathbf{x}) = 10\mathbf{v}_0^T\mathbf{x} + 5log(|B\, t\, \mathbf{v}_t^T\mathbf{x}|) + \epsilon$$

with B the bias parameter, $\mathbf{v}_0$ the baseline effect parameter, defined as $\mathbf{u}_0/||\mathbf{u}_0||$ with $||\cdot||$ the euclidean norm and $\mathbf{u}_0 \sim \mathcal{U}(0, 1)$ a randomly sampled vector ($\mathbf{u}_0 \in \mathbb{R}^{30}$), and $\epsilon \sim \mathcal{N}(0, 1)$. Recall that a potential outcome, denoted $\mathbf{y}^t$, is the outcome that a datapoint would have had, had it been treated with a particular treatment t. The matrix of potential outcomes $\mathbf{Y} \in \mathbb{R}^{Dx5}$ is defined as

$$\mathbf{Y} = [\mathbf{Y}^0, \mathbf{Y}^1, \mathbf{Y}^2, \mathbf{Y}^3, \mathbf{Y}^4] = [l_a(\mathbf{0}, \mathbf{X})^T,\ l_b(\mathbf{1}, \mathbf{X})^T,\ l_c(\mathbf{2}, \mathbf{X})^T,\ l_b(\mathbf{3}, \mathbf{X})^T,\ l_a(\mathbf{4}, \mathbf{X})^T]$$

with **0** = $(0, 0, ...0) \in \mathbb{R}^D$, **1** = $(1, 1, ...1) \in \mathbb{R}^D$, etc.

Several datasets were generated under varying values of the four parameters of interest, bias size $B = [2, 5, 10, 30]$, degree of positivity $\rho = [0.6, 0.7, 0.8, 0.90, 0.95, 0.98]$, dataset size $D = [1000, 2000, 5000, 8000]$ and number of confounders $NC = [2, 5, 10, 18]$, varying one parameter at a time. When kept fixed, the values were set to $B = 20$, $\rho = 0.8$, $D = 2000$ and $NC = 2$.

6.4.2 IHDP data generating process

For generating the IHDP dataset, a similar strategy was followed, but fixing $B = 10$, $\rho = 0.8$, $NC = 2$, with $D = 985$ being the size of the original IHDP covariate set. The treatment assignment function was based on two variables present in the set: mom ethnicity and weeks preterm. Treatment 0 is assigned to individuals with mom ethnicity equalling ”black”, treatment 1 to individuals with mom ethnicity equalling ”white”, treatment 2 to individuals with mom ethnicity equalling ”hispanic”, treatment 3 to individuals with mom ethnicity equalling ”hispanic” and weeks preterm being bigger than 6, and treatment 4 to individuals with mom ethnicity equalling ”black” and weeks preterm smaller than 6. Note that this setting is fictional and has no connection with any real-life situation. Then, the final treatment was sampled from a probability distribution as explained in section 6.4.1. The outcome functions were defined as

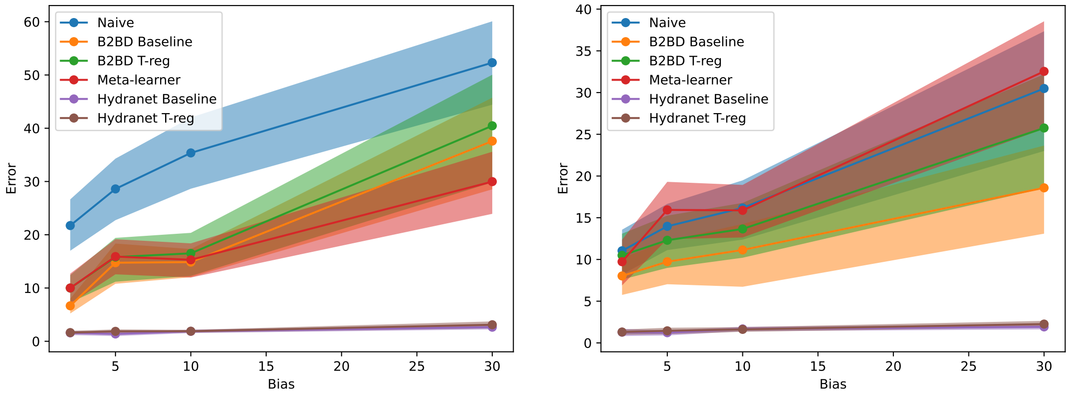
$$\begin{aligned} l_1(t, \mathbf{x}) &= \exp(\mathbf{x}\beta) + B * MB + t^2 + \epsilon \\ l_2(t, \mathbf{x}) &= log(|\mathbf{x}\beta|) + B * MW * t + \epsilon \\ l_3(t, \mathbf{x}) &= \mathbf{x}\beta + B * MH + t^2 + \epsilon \\ l_4(t, \mathbf{x}) &= \exp(\mathbf{x}\beta) + B * WP + t + \epsilon \\ l_5(t, \mathbf{x}) &= log(|\mathbf{x}\beta|) + B * WP * t + \epsilon \end{aligned}$$

where β is a vector of parameters, B is the bias parameter, MB, MW, and MH are the components of the one-hot encoding of mom ethnicity, and WP is weeks preterm.

6.4.3 Metrics

For performance benchmarking purposes, the sum of errors of the vector of ATEs was employed. This is computed as the sum of the absolute values of the differences of all estimated ATE components with respect to their true values, $E = \sum_{t=1}^{k} |\psi_t - \hat{\psi}_t|$. This choice allows us to have a single real number as a final result, making comparisons simpler. All values were computed as averages across 20 dataset realizations to increase the robustness of the results, and 95% confidence intervals were computed with Bootstrapping.

In the case of binary treatment settings, there are *de facto* benchmarking datasets and metrics, i.e., datasets and metrics that are widely used in the literature and thus serve for algorithmic performance comparison purposes. The IHDP dataset and the metrics presented in Dorie et al. (2018) are an example of this. This is not the case in multi-valued treatment settings, where comparators are scarce. Nevertheless, algorithms that can be considered comparable to Hydranet were developed and implemented to benchmark its performance. Thus, in every experiment, the results of the following algorithms are included: 1) **Naive**, a naive estimator of the treatment effect that employs only the observable data, without controlling, and serves to visualize the impact of confounding 2) **B2BD**, back to back Dragonnets, a strategy that uses 4 Dragonnets (with the same setup as in Shi, Blei, and Veitch (2019)), each one estimating one element of the vector of ATEs ψ, 3) **Meta-learner**, a Meta-learner estimator (Künzel et al., 2019) that employs a gradient boosting



(a) Out-sample (b) In-sample

Figure 6.2: Errors of the different algorithms with respect to the bias size parameter.

machine (GBM) model, and finally 4) **Hydranet**, both in its baseline form and with targeted regularization. Note that for the meta-learner, both T-learner and X-learner estimators were tested, and the T-learner was finally selected due to its better performance. Hydranet performed well in all the tested scenarios and outperformed the comparators, both with in-sample (train set) data and with out-sample (test set) data, reaching low or very low error values for different bias sizes, positivity degrees, dataset sizes, and number of confounders. The employed training scheme consisted of a first stage with the ADAM optimizer and a second stage with the Stochastic Gradient Descent (SGD) optimizer, with hyperparameters similar to those in Shi, Blei, and Veitch (2019).

6.4.4 Synthetic data experiments

Figure 6.2 and Table 6.1 show the error of the different algorithms for increasing values of the bias size. As should be expected, the error of the naive algorithm increases with the bias size, and the out-sample error is bigger than the in-sample error. The comparators also suffer from bigger error sizes with the increase of the bias. Hydranet outperforms the comparators and is very robust in the face of an increase in bias. It also shows a similar performance in-sample and out-sample, both for the baseline algorithm as well as the targeted regularization-equipped algorithm.

Figure 6.3 and Table 6.2 show the error of the different algorithms for increasing values of the degree of positivity ρ. Note that here ρ has been expressed in percentage. Again, as expected, due to its definition, all algorithms suffer from increasing error size with the increase of ρ. Hydranet outperforms the comparators both in-sample and out-sample and both in its baseline form as well as with targeted regularization.

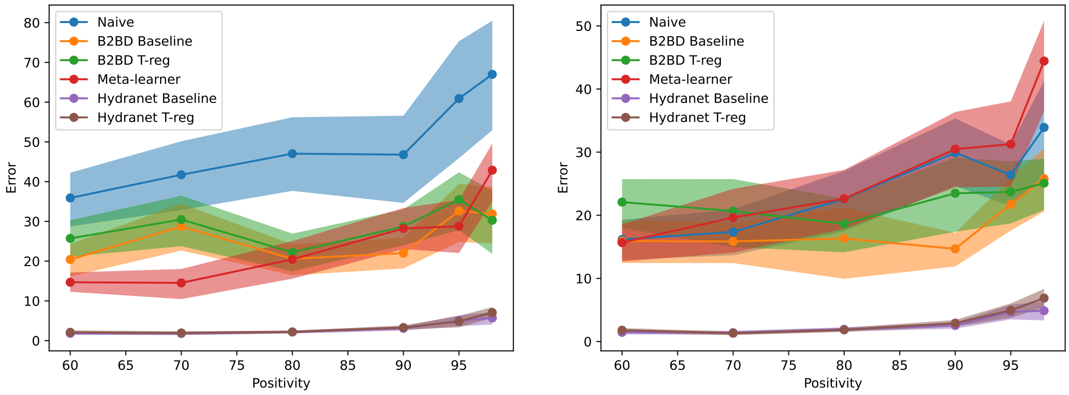
Figure 6.4 and Table 6.3 show the performance of the algorithms for varying dataset sizes. As expected, all algorithms reduce their error with bigger data set sizes, but Hydranet with targeted

Table 6.1: Errors of the different algorithms with respect to the bias size parameter.

Bias	5		10		30	
	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
Naive	28.61 ± 5.78	13.97 ± 2.77	35.37 ± 6.69	16.17 ± 3.55	52.31 ± 7.83	30.49 ± 7.17
B2BD base.	14.75 ± 3.78	9.73 ± 2.6	14.86 ± 2.64	11.14 ± 3.76	37.59 ± 8.6	18.58 ± 5.28
B2BD t-reg.	12.3 ± 4.12	12.3 ± 3.13	13.66 ± 4.05	13.66 ± 3.29	25.76 ± 10.11	25.76 ± 6.83
Meta-learner	15.91 ± 3.3	15.94 ± 3.43	15.3 ± 3.21	15.88 ± 3.15	29.98 ± 5.83	32.54 ± 6.44
Hydranet base.	1.37 ± 0.37	1.22 ± 0.31	1.87 ± 0.32	1.68 ± 0.26	2.65 ± 0.4	1.92 ± 0.31
Hydranet t-reg.	1.45 ± 0.37	1.45 ± 0.38	1.62 ± 0.26	1.62 ± 0.28	2.26 ± 0.65	2.26 ± 0.38

Table 6.2: Errors of the different algorithms with respect to the degree of positivity parameter.

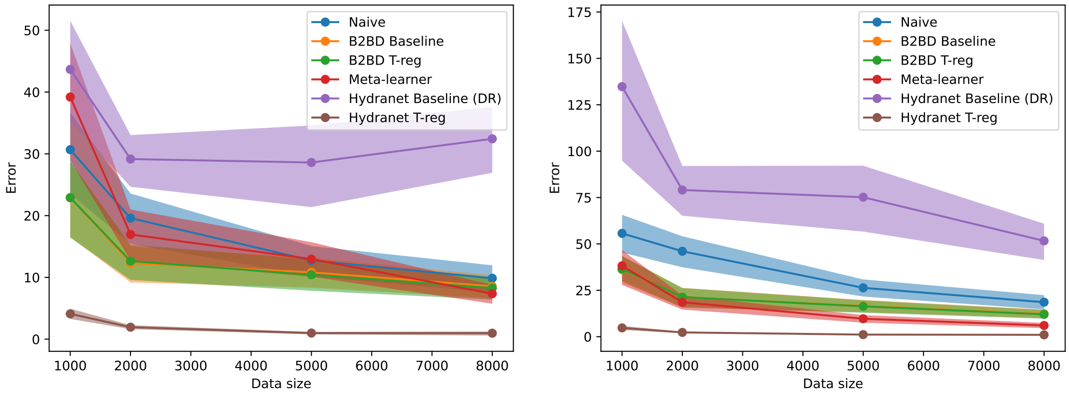
Positivity degree	90		95		98	
	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
Naive	46.79 ± 10.97	29.92 ± 5.35	60.9 ± 14.71	26.39 ± 4.88	67.0 ± 13.77	33.91 ± 7.47
B2BD base.	22.02 ± 3.85	14.68 ± 2.66	32.62 ± 7.33	21.71 ± 4.14	31.87 ± 7.15	25.79 ± 5.02
B2BD t-reg.	23.49 ± 4.52	23.49 ± 5.87	23.7 ± 7.27	23.7 ± 4.9	25.08 ± 7.98	25.08 ± 4.06
Meta-learner	28.21 ± 5.17	30.48 ± 5.96	28.77 ± 6.67	31.26 ± 6.76	42.88 ± 7.3	44.43 ± 7.17
Hydranet base.	3.09 ± 0.53	2.54 ± 0.48	5.01 ± 1.35	4.69 ± 1.17	5.7 ± 1.57	4.88 ± 1.48
Hydranet t-reg.	2.91 ± 0.53	2.91 ± 0.53	4.93 ± 1.43	4.93 ± 1.13	6.88 ± 1.37	6.88 ± 1.46



(a) Out-sample

(b) In-sample

Figure 6.3: Errors of the different algorithms with respect to the degree of positivity parameter.



(a) In sample

(b) Out sample

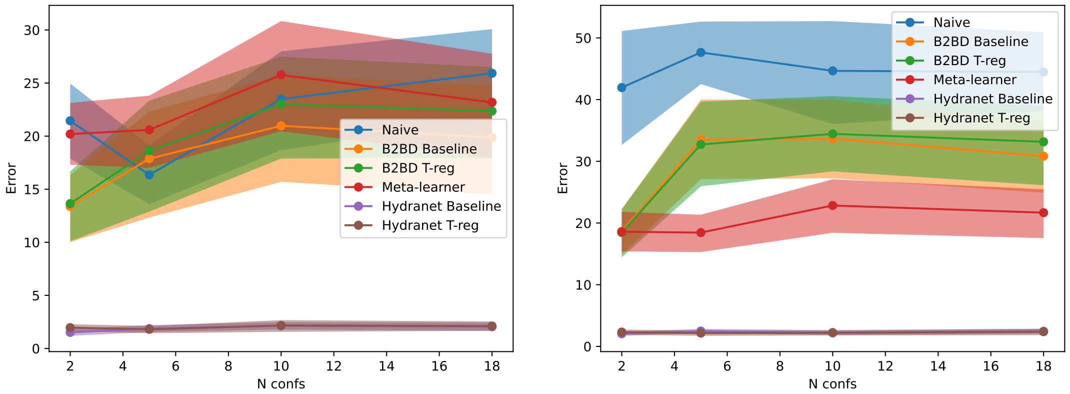
Figure 6.4: Errors of the different algorithms with respect to the dataset size parameter.

regularization outperforms the rest and shows a smaller error even for small dataset sizes, proving its (data) efficiency. It must be highlighted that in this experiment, the estimations of the baseline Hydranet were plugged into a doubly robust estimator, the Augmented Inverse Probability of Treatment Weighted (A-IPTW) estimator. The resulting estimations of that strategy are biased, unlike those of Hydranet with targeted regularization, which proves the utility of the targeted regularization loss function for achieving double robustness.

Figure 6.5 and Table 6.4 show the performance of the algorithms for varying numbers of confounders. Hydranet outperforms the comparators. There is no clear pattern in the impact of the increase in the number of confounders, probably due to the design of the DGP.

Table 6.3: Errors of the different algorithms with respect to the dataset size parameter.

Dataset size	2000		5000		8000	
	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
Naive	46.03 ± 8.36	19.59 ± 4.07	26.33 ± 4.61	12.75 ± 2.44	18.59 ± 3.83	9.86 ± 2.21
B2BD base.	21.23 ± 5.29	12.3 ± 2.92	16.59 ± 3.31	10.82 ± 2.44	12.32 ± 2.31	8.55 ± 1.86
Meta-learner	18.55 ± 3.82	16.95 ± 4.13	9.67 ± 2.06	12.97 ± 2.8	6.04 ± 1.43	7.35 ± 1.53
B2BD t-reg.	12.63 ± 5.19	12.63 ± 2.86	10.42 ± 3.33	10.42 ± 2.53	8.27 ± 2.33	8.27 ± 1.81
Hydranet base. (DR)	79.05 ± 13.41	29.15 ± 4.18	75.16 ± 17.76	28.6 ± 6.61	51.64 ± 9.8	32.43 ± 5.33
Hydranet t-reg.	1.94 ± 0.42	1.94 ± 0.32	0.99 ± 0.22	0.99 ± 0.2	0.97 ± 0.36	0.97 ± 0.34



(a) In sample

(b) Out sample

Figure 6.5: Errors of the different algorithms with respect to the number of confounders parameter.

Table 6.4: Errors of the different algorithms with respect to the number of confounders parameter.

N. confounders	5		10		18	
	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
Naive	47.65 ± 5.07	16.35 ± 2.76	44.66 ± 8.34	23.49 ± 4.65	44.48 ± 6.38	25.92 ± 4.28
B2BD base.	33.48 ± 6.45	17.87 ± 4.99	33.67 ± 6.38	20.97 ± 5.03	30.84 ± 5.81	19.86 ± 5.12
B2BD t-reg.	18.62 ± 6.86	18.62 ± 5.23	23.01 ± 6.1	23.01 ± 4.79	22.36 ± 6.43	22.36 ± 4.3
Meta-learner	18.45 ± 3.02	20.59 ± 3.4	22.84 ± 4.34	25.77 ± 5.19	21.68 ± 3.95	23.18 ± 4.91
Hydranet base.	2.49 ± 0.35	1.88 ± 0.33	2.25 ± 0.42	2.13 ± 0.35	2.46 ± 0.41	2.04 ± 0.39
Hydranet t-reg.	1.8 ± 0.44	1.8 ± 0.34	2.17 ± 0.4	2.17 ± 0.55	2.11 ± 0.52	2.11 ± 0.44

6.4.5 IHDP data experiments

Table 6.5 shows the error of the different algorithms with the IHDP dataset. Similarly to what happens with synthetic data, Hydranet (both baseline and targeted regularization) outperforms the

Table 6.5: Performance of the different algorithms with the IHDP dataset.

	Out-Sample	In-Sample
Naive	14.81 ± 0.95	17.51 ± 2.03
B2BD base.	26.35 ± 2.46	26.73 ± 3.1
B2BD t-reg.	27.57 ± 2.42	26.05 ± 2.84
Meta-learner	13.53 ± 1.22	13.7 ± 1.2
Hydranet base.	3.22 ± 0.73	3.33 ± 0.83
Hydranet t-reg.	2.87 ± 0.57	2.91 ± 0.68

comparators. The targeted regularization algorithm has a slightly smaller error than the baseline algorithm. These results prove the efficacy of Hydranet with semi-synthetic data, showing its suitability for real-world scenarios.

6.5 Discussion

In this chapter, we generalized a top-performing, neural network-based algorithm for ATE estimation from a binary treatment setting to a 5-valued treatment setting. We developed and implemented synthetic and semi-synthetic DGPs for algorithmic benchmarking purposes in multivalued settings, and we designed comparator algorithms to evaluate the performance of Hydranet. We have shown that Hydranet performs well under different bias sizes and degrees of positivity, and we provide both theoretical and empirical evidence about the benefits of developing targeted regularization-equipped Hydranet. In addition, the algorithm's good performance with semi-synthetic data is demonstrated.

The main limitations of this work are twofold: on one hand, only a 5-valued treatment scenario has been tested. It is a line of future work to adapt the algorithm and perform experiments for k-valued scenarios. On the other hand, competitor algorithms of Hydranet have been constructed *ad-hoc* due to the scarcity of benchmarking data in the literature. In one of the few potential comparison candidates, Schwab, Linhardt, and Karlen (2019), some experiments are performed in multivalued treatment settings, with TARNet being the best method. TARNet was shown to be outperformed by Dragonnet in binary treatment settings in Shi, Blei, and Veitch (2019), and thus, we presumed that, as an extension of Dragonnet, Hydranet would also outperform TARNet in multivalued treatment scenarios. Nevertheless, this has not been tested empirically.

Regarding research question **Q5** (*Can we generalize advanced causal inference algorithms from binary treatment settings to multivalued treatment settings?*), we can state that despite the direct generalizability of neural network-based algorithms for ATE estimation from binary to k-valued treatment settings is a common claim in the literature, this work shows that it has its own challenges and that the behavior of the algorithms in each particular scenario requires its own interpretations. As far as we know, the work in this chapter is opening ground on the proposal of benchmarking results for neural network-based ATE estimation in multivalued treatment scenarios.

The next chapter is the last of this thesis and presents the overall conclusions and the lines of future work.

Chapter 7

Conclusions and future work

In this chapter, we review the key lessons learned from this thesis, highlight its main contributions and related publications, and outline the primary directions for future research.

7.1 General conclusions

In this thesis, we have spoken about causality in the particular domain of generating evidence about the effects of health technologies and interventions on health outcomes. The overall motivation was to determine whether causal inference should be the framework of choice for *producing* that evidence in such context. In that sense, this thesis, with its real-world use cases and analyses, aimed to be a piece of evidence itself in favor of the positive hypothesis, i.e., the hypothesis stating that causal inference should indeed be, in general, and when possible, the choice for generating evidence on effects of interventions in the domain of human health and health technology assessment. We consider the goal fulfilled.

We have tackled real-world use cases of evidence generation on several health outcomes using real-world data, and we have done so both with correlational and causal methods. Thus, we have shown an association between the timing of the COVID-19 vaccine administration and alterations in the menstrual cycle. This discovery, obtained employing associational methods, has later been confirmed by other works, i.e., evidence in favor of our conclusions has continued to accumulate. We have also shown that the COVID-19 vaccine protects from the COVID-19-increased risk of diabetes mellitus in a causal analysis that is first-of-its-kind in the literature about that topic. Finally, we have shown a beneficial effect of antibiotic-loaded bone cement on the survival of knee prostheses, adding our piece of evidence to the literature in a question that is still considered open. We have done so using both a correlational and a causal approach and employing data from the largest arthroplasty registry of southern Europe published to date.

Throughout the execution of the aforementioned studies, we have employed correlational and causal approaches and methods, and we have qualitatively and critically assessed and compared them. In general, we have noted the limitations of correlational methods. One of the most important ones is epistemic, i.e., it refers to the type of knowledge that the correlational approach provides. Strictly speaking, that approach can only find correlations, not causation. Nevertheless, the true nature of the questions of our studies was causal, and we claim that so is the case with most studies of this type. In fact, more often than not, other correlational studies from the related literature used causal vocabulary and implied, more or less explicitly, causality in their conclusions. Interestingly enough, there is evidence showing that people do infer causality from statements of association (Gershman and Ullman, 2023). This whole combination of facts can lead to confusion. On the contrary, the causal approach and causal inference methods allowed us to make our assumptions explicit and provided guarantees that if those held, the observed correlations would indeed be causal relationships. Furthermore, as we saw in Chapter 5 during the direct comparison between approaches, the implications of the mentioned epistemic differences were also practical, with observable differences in the obtained results. The weakness of the causal approach is that it requires us to make assumptions, and some of them are empirically untestable from the employed observational data, but we believe that this is still better than making them implicitly, as the correlational approach requires.

In addition, we have also shown that it is possible to extend neural network-based, top-performing causal inference algorithms from binary treatment settings to multivalued treatment settings, further paving the way for the adoption of these methods for real-world problems.

Overall, according to the experience acquired throughout this thesis, we claim that the causal approach is an improvement with respect to the correlational approach for the generation of evidence on the effects of health interventions. This improvement is big in the theoretical and epistemic aspects and somewhere between big and marginal in the practical aspect, depending on the use case and the overall study design. We forecast that the causal approach, as any improvement, will continue growing, eventually becoming commonplace. This does not necessarily mean completely substituting or making the correlational approach disappear, but it does mean that the quality of the evidence generated with the former will be considered higher than that generated with the latter. As an example to illustrate our point, note that something similar happens already with cohort studies and case reports, as the former are considered to produce higher quality evidence than the latter, without resulting in the disappearance of case report studies. Recall that, at the end of the day, this whole endeavor is *just* about obtaining better evidence, and whatever works better will eventually become apparent and prevail.

7.2 Contributions

In this section, we briefly review the main contributions of this thesis, connecting them with the research questions.

- Chapter 3:
  - Contribution 1: We provide the first piece of evidence in the literature correlating the COVID-19 vaccine administration time and vaccine-induced menstrual cycle disorders, and by so doing, we answered research question **Q1** (*Do the vaccine against COVID-19 and the vaccination time have any effect on the menstrual cycle?*)

    We used data from a menstrual cycle tracking smartphone application (Lunar App (APP Lunar 2024)) to find a correlation between vaccination timing (in particular, the phase of the menstrual cycle) and alterations in the cycle. We discovered that individuals vaccinated during the luteal phase suffered fewer changes in their cycles than those vaccinated during the follicular phase. Larger posterior studies have confirmed our findings. This could translate into recommendations for menstruating individuals about when to get vaccinated. This study is a successful example of multi-institution collaboration and citizen science, as the data used was collected through a smartphone application for purposes other than research.
- Chapter 4:
  - Contribution 2: We provide one of the first pieces of evidence of the protective effect of the COVID-19 vaccine against infection-induced increased risk of diabetes mellitus. Furthermore, our work also constitutes the first explicit causal analysis of the topic in the literature. This contribution answered research question **Q2** (*Does the vaccine against COVID-19 have any effect on the risk of developing diabetes?*)

    We emulated a target trial for analyzing the effect of the COVID-19 vaccine on the risk of diabetes mellitus. For that purpose, we employed vaccination, infection, diagnostics, and other data from the whole population of Catalonia, with around 7.5 million individuals in our cohort. We developed a DAG of the problem, including time-varying confounding, and we employed a parametric implementation of G-formula with random forests for estimating the effects of getting 1, 2, or 3 doses of the COVID-19 vaccine on the cumulative hazard of diabetes onset. We discovered that the vaccine has a protective effect in front of the risk of diabetes onset.
  - Contribution 3: We are the first to integrate random forests for modeling purposes in the parametric NICE implementation of the G-formula.

    We modified an existing software package (*pygformula* 2024) that implements the parametric noniterative conditional expectation (NICE) G-formula. The original package employed generalized linear models for the modeling steps, and we changed those to random forests. The goal was to use more flexible models that would impose fewer assumptions on the distributions of the modeled data. In addition, this change substantially increased the efficiency (in time and memory) of the overall algorithm.

- Chapter 5:
  - Contribution 4: We provide evidence of the positive effect of antibiotic-loaded bone cement on the survival of knee prostheses, using data from the largest arthroplasty registry in southern Europe with published results to date. By so doing, we provide the first answer to research question **Q3** (*Does the use antibiotic-loaded bone cement during total knee arthroplasty surgery increase the life of knee prostheses, compared with the alternative of using plain cement?*)

    We performed a correlational study of the relationship between the type of bone cement and knee prosthetic survival, employing classical methods for this task (Kaplan-Meier estimator and Cox proportional hazards model). The used data came from RACat, a population-based knee arthroplasty registry from Catalonia. We observed a positive correlation between antibiotic-loaded bone cement and prosthetic survival. Our results can be integrated into future meta-analyses of a topic that is considered open in the literature.
  - Contribution 5: We provide the first explicit causal analysis of the previous topic, employing a tailored, random forest-based causal survival analysis algorithm for estimation. In this manner, we provide our second answer to research question **Q3** (*Does the use antibiotic-loaded bone cement during total knee arthroplasty surgery increase the life of knee prostheses, compared with the alternative of using plain cement?*)

    We also performed a causal analysis of the effect of antibiotic-loaded bone cement on prosthetic survival, the first of its kind to the best of our knowledge. To do so, we developed a DAG with experts in the matter, and we selected the most advanced estimator in the literature for the task of causal survival analysis. Such an estimator was based on random forests, was doubly robust with a correction for the censoring process, and showed top performance for the estimation of individual treatment effects. Our use is one of the first examples of application to a real-world problem.
  - Contribution 6: Assessment and comparison of the employed correlational and causal methods. This contribution answered research question **Q4** (*What are the advantages and disadvantages, strengths and weaknesses, of correlational methods and causal inference methods for generating evidence about clinical interventions?*)

    We performed a comparison between the causal and the correlational approaches, using the studies about antibiotic cement type and prosthetic survival as a base. We conclude that the benefits of the causal approach, both theoretical and practical, outweigh the disadvantages, and that the limitations of the correlational approach are enough to avoid its use when possible.

- Chapter 6:
  - Contribution 7: We generalize a top-performing neural network-based causal inference algorithm from a binary to a multivalued treatment setting. This contribution addresses research question **Q5** (*Can we generalize advanced causal inference algorithms from binary treatment settings to multivalued treatment settings?*)

    The literature on advanced causal inference algorithms, especially those based on machine learning and neural networks, is overwhelmingly dominated by scenarios of binary treatments. This is logical but also a limitation. We open ground on the empirical generalization and assessment of these types of algorithms to multivalued treatment settings by providing a use case with a top-performing algorithm (originally developed by Shi, Blei, and Veitch, 2019).

In a more general and broad sense, a significant part of the industrial contribution of this thesis lies in the knowledge gained. AQuAS, the Agency of Health Quality and Assessment of Catalonia, now possesses a more profound know-how on causal inference, which is crucial for one of its core missions, health technology assessment. In addition, the algorithms developed for this thesis remain in AQuAS' repositories and will be reused in the future.

7.3 Publication list

In this section, we list and briefly explain the publications of this thesis. This list includes work that has already been published, that is under the peer review process, or that will be sent to a journal soon. We also include two non-peer-reviewed AQuAS reports.

- Published work

  1. Borja Velasco-Regulez, Jose L. Fernandez-Marquez, Nerea Luqui, Jesus Cerquides, Josep Analia Fukelman, & Josep Perelló (2022). Is the phase of the menstrual cycle relevant when getting the covid-19 vaccine? *American Journal of Obstetrics and Gynecology*, 227, 913-915. DOI: 10.1016/j.ajog.2022.07.052

     Journal research letter in the American Journal of Obstetrics and Gynecology about the correlation between the administration time of the COVID-19 vaccine and the changes in the menstrual cycle. This publication is connected to the work presented in Chapter 3 of this thesis.

2. Borja Velasco, Jesus Cerquides, & Josep Lluis Arcos (2022). ”Multi-valued Treatment Effect Estimation for Health Technology Assessment with a Neural Network.” *NeurIPS 2022 Workshop on Causality for Real-world Impact.*

   Poster and virtual talk in the NeurIPS 2022 Workshop on Causality for Real-world Impact. This work was the early stage, work in progress of the content of Chapter 6 of this thesis, about a neural network-based multivalued treatment causal inference algorithm. Back then, no comparator algorithm was developed, and the data-generating process was more basic. The algorithm's performance was worse than that of its final version.

3. Borja Velasco-Regulez, & Jesus Cerquides (2023). Hydranet: A Neural Network for the Estimation of Multi-Valued Treatment Effects. *Artificial Intelligence Research and Development (pp 16–27). IOS Press.* DOI: 10.3233/FAIA230655

   Proceedings of the Catalan Conference of Artificial Intelligence of 2023. ***Best paper award*** **of the conference**. At this conference, we presented a nearly final version of the work about the Hydranet algorithm, which can be found in Chapter 6 of this thesis. Back then, the employed data-generating process did not have a tunable knob for the degree of positivity, and the performance of the algorithm was slightly worst than in its final version due to the training strategy.

4. Sergi Gil-Gonzalez, Borja Velasco-Regúlez, Jesus Cerquides, et al. (2023). ¿El cemento con antibiótico reduce el riesgo de infección protésica en artroplastia primaria total de rodilla? Análisis del registro catalán de artroplastias. *10º Congreso de la AEA-SEROD.*

   Oral poster communication in the 10^{th} congress of the Spanish Arthroplasty Association (AEA) and the Spanish Knee Association (SEROD). ***Best oral poster communication award*** **of the conference**. This work contained an early version of a correlational analysis of antibiotic-loaded bone cement and prosthetic survival, present in Chapter 5 of this thesis.

5. Gil-Gonzalez Sergi, Velasco-Regúlez Borja, Cerquides Jesus, et al. (2024). Antibiotic-loaded bone cement is associated with a reduction of the risk of revision of total knee arthroplasty: Analysis of the Catalan Arthroplasty Register. *Knee Surgery, Sports Traumatology, Arthroscopy*. DOI: 10.1002/ksa.12361

   Journal article published in the Knee Surgery, Sports Traumatology, Arthroscopy (KSSTA) journal. This article presents a correlational analysis of the relationship between the use of antibiotic-loaded bone cement and prosthetic survival. This work has been presented in Section 5.3 of Chapter 5 of this thesis.

- Work under review or in progress

1. Borja Velasco-Regúlez, Sergi Gil-Gonzalez, Jesus Cerquides. Causal analysis of the effect of antibiotic-loaded bone cement on knee prosthesis survival. *Sent to the Journal of Healthcare Informatics Research* - Under review.

   Article sent to the Journal of Healthcare Informatics Research. This article contains a causal analysis of the effect of using antibiotic-loaded bone cement on prosthetic survival. This work has been presented in Section 5.4 of Chapter 5 of this thesis.

2. Borja Velasco-Regulez, & Jesus Cerquides. Hydranet: A Neural Network for the Estimation of Multi-Valued Treatment Effects. *Sent to Artificial Intelligence Communications.* - Under review.

   The article was sent to the Artificial Intelligence Communications journal and contains the final version of the work of Hydranet, as it appears in Chapter 6 of this thesis.

3. Article about the effect of the vaccine against COVID-19 on the risk of diabetes onset. This work is currently in progress and will be sent to the European Journal of Epidemiology. It can be found in Chapter 4 of this thesis.

- Non-peer-reviewed work

1. Pérez-Troncoso, Daniel, Borja Velasco-Regulez, Jessica Ruiz-Baena, Silvia Ballesta, Gemma Llauradó Cabot, Rosa Maria Vivanco-Hidalgo, Juan J. Chillarón, and Elisenda Climent. ”La incidència de diabetis mellitus de tipus 1 durant la pandèmia de COVID-19 a Catalunya.” (2023).

   Study about the incidence of Type 1 diabetes (T1D) during the COVID-19 pandemic in Catalonia. Using historical data from 2010-2019, we employed a Poisson regression model for estimating the expected incidence for 2020-2021, which was then compared to the actual incidence computed with data from a population-based registry. Results showed no significant increase in 2020 but a 28% rise in 2021, particularly among women and patients under 18 years old. Further research was warranted to explore potential biological or social causes of the rise and their health implications.

2. Ruiz, Jessica, Laura Llinàs Mallol, Roland Pastells-Peiró, Daniel Pérez-Troncoso, Borja Velasco-Regulez, Agata Carreño, and Rosa Maria Vivanco-Hidalgo. ”Guia per a la generació d'evidència amb dades del món real en l'avaluació de tecnologies sanitàries.” (2023).

   Methodological guide for the generation of evidence with real-world data in the domain of health technology assessment. This guide describes how The Agency for Health Quality and Assessment of Catalonia (AQuAS) manages access to health data for its reuse and conducts HTA, aiming to improve public health. Standardized processes for producing high-quality research are provided.

7.4 Future work

In this section, we present the lines of future work for this thesis. We divide this part into four subsections. In the first one, we introduce the ideas for a causal analysis of the effect of the COVID-19 vaccine on the menstrual cycle. The second one is about an analysis of different COVID-19 scenarios in the assessment of the effect of the vaccine on the risk of diabetes. The third discusses a target trial for determining the effect of antibiotic-loaded bone cement on prosthetic survival, and finally, the last one proposes a more realistic data-generating process for multivalued treatments.

7.4.1 Analysis of the effect of the COVID-19 vaccine on the menstrual cycle with causal methods

The research that we conducted analyzing the association between vaccination timing and effects on the menstrual cycle could be expanded and improved by using causal methods. We would analyze the effect of two different interventions: on the one hand, the vaccine as a binary treatment (vaccinated or not vaccinated), and on the other hand, the vaccine as a treatment with three categories: not vaccinated, vaccinated during the luteal phase, and vaccinated during the follicular phase. We would need to develop a DAG of the problem, with the help of gynecologists and other experts, to include all the potentially involved variables and their causal relationships.

To be able to conduct such a study, we would need to gather more data, as the database employed originally contained limitations such as the lack of information about confounders. Thus, using citizen science again, the research team would have to design the data requirements and implement them in the Lunar App smartphone application. We could also devise and introduce reward strategies to encourage the registration of the required information. The study would then become prospective. This has benefits, such as having a bigger control over the collected data, but also important inconveniences, such as the higher costs and the longer data collection times.

7.4.2 Analysis of different COVID-19 infection scenarios in the assessment of the effect of the vaccine on the risk of diabetes mellitus

We plan two important interventions for this line of work, which discusses the effect of the COVID-19 vaccine on the risk of diabetes mellitus. First, we plan to collect finer data on the COVID-19 infection status variable, going from a binary variable (infection yes or no) to a variable with at least three categories: no infection, mild infection, or severe infection. Then, we will repeat the analysis, as explained in Chapter 4, to see whether this change has an impact on the obtained estimates. In addition, we plan to intervene not only the treatment variable (the number of administered vaccine doses), but also the COVID-19 infection status variable, using the simulation capabilities of the G-formula. Thus, we will simulate several scenarios with different percentages of infection and severity rates and observe potential variations of the effects of the vaccine on the risk of diabetes onset.

Besides these changes in the data and the intervention scenarios, we also plan to use other alternative G-methods for estimation. The options are structural nested models with G-estimation (Vansteelandt and Joffe, 2014) and marginal structural models with inverse probability of treatment weighting (James M. Robins, M. Á. Hernán, and Brumback, 2000). These alternatives make different assumptions on different aspects of the estimation problem: distribution of the effects across the population, distribution of the covariates, etc. Computing estimates with more than one approach can increase the robustness of the results.

7.4.3 Target trial for determining the effect of antibiotic-loaded bone cement on prosthetic survival

We plan to design the protocol of a target trial for measuring the effect of antibiotic-loaded bone cement on prosthetic survival and emulate such a trial with the available observational data. Having already developed a DAG for the problem and having employed a machine learning-based, doubly-robust causal survival method for estimation, the only step left that could remove potential sources of bias (in particular time biases) is to conduct a trial emulation. Thus, we plan to use this framework, together with the cloning-censoring-weighting method, and tackle the challenges of using the weights with the causal survival forests algorithm.

In addition, we also plan to estimate the same quantities as in the correlational approach, i.e., survival curves and conditional hazard functions and ratios, to be able to make a more direct comparison between the results of the causal and the correlational approaches.

7.4.4 A more realistic data generating process for the evaluation of algorithmic performance in scenarios with multivalued treatments: multivalued RealCause

As we explained in Chapters 2 and 6 of this thesis, an important challenge for testing the performance of causal inference algorithms is to have access to the appropriate benchmarking data. Because with real-world data, we do not usually have access to *ground truth* causal effects, we need to synthetically or semi-synthetically generate it for performance testing. And then, in general, we have no guarantee about the realism (i.e., the similarity to real-world data) of such generated data.

In Chapter 6, we developed two well-designed, fully synthetic, and semi-synthetic data-generating processes for testing Hydranet, our neural network-based, multivalued treatment causal inference algorithm. Afterward, we found a work that proposed a method that we considered very interesting for generating more realistic data for that purpose, and we extended it from binary to multivalued treatment settings. The method is called *RealCause* and is presented in Neal, Huang, and Raghupathi, 2021. It consists of using real-world, observational data for fitting generative models and then using those models for generating data, including counterfactuals (which gives us access to ground truth causal effects). The method employs the TARNet neural network architecture (explained in Chapters 2 and 6) to parameterize the generative models. In addition, the authors of that work provide evidence that the data outputted by the generative models is statistically indistinguishable from the real data under statistical tests of distribution distances. We already adapted the algorithm for multivalued treatments, and we are gathering multivalued real-world data for fitting the generative models. Our plan is to test Hydranet with this new, more realistic data. Finally, we also plan to compare Hydranet to potentially stronger competitor algorithms than the ones we employed so far. In particular, an algorithm named *Perfect Match* (Schwab, Linhardt, and Karlen, 2019) is the best candidate, as it is one of the few ones in the literature that presented an extension to multivalued treatment settings.

7.5 Final conclusions

The core objective of the work behind this thesis has been to explore the application of advanced causal inference algorithms in the domain of healthcare and to compare them with traditional correlational algorithms. This goal was partly motivated by the growing interest in causality observed in the literature over recent years. Causal inference methods are not new, but some of the foundational algorithms of the field were published at a time when the causal inference *label* did not have the cohesive power that it has nowadays. Still, the distinction between what is a causal approach and what is not is sometimes fuzzy in the literature, especially in the applied one. Thus, one of the secondary but still important tasks of this thesis has been to demarcate those boundaries very clearly. We say that an approach to a question is causal when it explicitly acknowledges the causal nature of the question and the provided answer, when it states that it is based on common assumptions (consistency, positivity, and no hidden confounder or exchangeability), and when it shows that, under those and potentially other assumptions, the obtained estimation of the quantity of interest has a causal interpretation, i.e., it is a *causal effect*. We believe this distinction will become more and more clear in the near future. We also believe that causal methods will continue their way toward becoming the standard methodology for analyzing the effects of interventions with observational data. Statistics courses in universities have recently started including causal inference as a matter of study, and some causal inference methods have proved their performance against the standard of randomized controlled trials (S. V. Wang, Schneeweiss, and Initiative, 2023).

To conclude, we hope that the ideas included in this thesis contribute to the aforementioned trends, assist other researchers in their work, and open new grounds for research. All of this aims to further advance causal inference methods that generate better evidence about interventions in the field of human health, a topic crucial to all of us.

Appendix A

Full, written DAG of the effect of the COVID-19 vaccine on the risk of diabetes mellitus onset

The list below shows all the existing relationships between all variables in the DAG of our problem. Note that the time has been ”collapsed” into a single stage, and thus instead of three nodes for the treatment or for a time-varying confounder, we have only one.

BMI $->$ Diastolic blood pressure
BMI $->$ Systolic blood pressure
BMI $->$ Adjusted comorbidity index
BMI $->$ COVID-19 infection
BMI $->$ Blood glucose
BMI $->$ Abdominal perimeter

Birth year $->$ BMI
Birth year $->$ Diastolic blood pressure
Birth year $->$ Systolic blood pressure
Birth year $->$ Adjusted comorbidity index
Birth year $->$ Cholesterol
Birth year $->$ Blood glucose
Birth year $->$ Smoking
Birth year $->$ Socioeconomic status indicator
Birth year $->$ Abdominal perimeter

Diastolic blood pressure $->$ Adjusted comorbidity index

Systolic blood pressure $->$ Adjusted comorbidity index

Adjusted comorbidity index $->$ COVID-19 infection

Cholesterol $->$ Adjusted comorbidity index

Country of origin $->$ BMI
Country of origin $->$ Diastolic blood pressure
Country of origin $->$ Systolic blood pressure
Country of origin $->$ Adjusted comorbidity index
Country of origin $->$ Cholesterol
Country of origin $->$ Blood glucose
Country of origin $->$ Smoking
Country of origin $->$ Socioeconomic status indicator
Country of origin $->$ Abdominal perimeter

Blood glucose $->$ Adjusted comorbidity index
Blood glucose $->$ COVID-19 infection

Smoking $->$ BMI
Smoking $->$ Diastolic blood pressure
Smoking $->$ Systolic blood pressure
Smoking $->$ Adjusted comorbidity index
Smoking $->$ COVID-19 infection
Smoking $->$ Abdominal perimeter

Socioeconomic status indicator $->$ BMI
Socioeconomic status indicator $->$ Diastolic blood pressure
Socioeconomic status indicator $->$ Systolic blood pressure
Socioeconomic status indicator $->$ Cholesterol
Socioeconomic status indicator $->$ COVID-19 infection
Socioeconomic status indicator $->$ Blood glucose
Socioeconomic status indicator $->$ Smoking

Appendix B

Models of covariates and outcome for the G-formula for the problem of the effect of the COVID-19 vaccine on the risk of diabetes mellitus onset

Covariate models and outcome model employed in the parametric NICE G-formula. $t-1$ sub-index indicates a lagged value of a variable.

Covariate models

Systolic blood pressure $\sim$ Systolic blood pressure$_{t-1}$ + BMI$_{t-1}$ + Birth year + Country of origin + Smoking$_{t-1}$ + Socioeconomic status indicator + Vaccine$_{t-1}$ + time

Smoking $\sim$ Smoking$_{t-1}$ + Birth year + Country of origin + Socioeconomic status indicator + Vaccine$_{t-1}$ + time

Cholesterol $\sim$ Cholesterol$_{t-1}$ + Birth year + Country of origin + Socioeconomic status indicator + Vaccine$_{t-1}$ + time

Abdominal perimeter $\sim$ Abdominal perimeter$_{t-1}$ + BMI$_{t-1}$ + Birth year + Country of origin + Smoking$_{t-1}$ + Vaccine$_{t-1}$ + time

Diastolic blood pressure $\sim$ Diastolic blood pressure$_{t-1}$ + BMI$_{t-1}$ + Birth year + Country of origin + Smoking$_{t-1}$ + Socioeconomic status indicator + Vaccine$_{t-1}$ + time

BMI $\sim$ BMI$_{t-1}$ + Birth year + Country of origin + Smoking$_{t-1}$ + Socioeconomic status indicator + Vaccine$_{t-1}$ + time

Blood glucose $\sim$ Blood glucose$_{t-1}$ + BMI$_{t-1}$ + Birth year + Country of origin + Socioeconomic status indicator + Vaccine$_{t-1}$ + time

COVID-19 infection $\sim$ COVID-19 infection$_{t-1}$ + BMI$_{t-1}$ + Birth year + Country of origin + Blood glucose$_{t-1}$ + Smoking$_{t-1}$ + Socioeconomic status indicator + Vaccine$_{t-1}$ + time

Adjusted comorbidity index $\sim$ Adjusted comorbidity index$_{t-1}$ + BMI$_{t-1}$ + Birth year + Diastolic blood pressure$_{t-1}$ + Systolic blood pressure$_{t-1}$ + Cholesterol$_{t-1}$ + Country of origin + Blood glucose$_{t-1}$ + Smoking$_{t-1}$ + Vaccine$_{t-1}$ + time

Vaccine $\sim$ Vaccine$_{t-1}$ + Systolic blood pressure$_{t-1}$ + Smoking$_{t-1}$ + Cholesterol$_{t-1}$ + Abdominal perimeter$_{t-1}$ + Diastolic blood pressure$_{t-1}$ + BMI$_{t-1}$ + Blood glucose$_{t-1}$ + COVID-19 infection$_{t-1}$ + Adjusted comorbidity index$_{t-1}$ + time

Outcome model

DM $\sim$ Area of residence + Country of origin + Sex + Birth year + Socioeconomic status indicator + Systolic blood pressure + Smoking + Cholesterol + Abdominal perimeter + Diastolic blood pressure + BMI + Blood glucose + COVID-19 infection + Adjusted comorbidity index + Vaccine + time

Appendix C

Results of the Cox proportional hazards model for the events of aseptic revision and all-cause revision in the problem of the effect of antibiotic-loaded bone cement on knee prosthetic survival

Table C.1: Cox proportional hazards model results for aseptic revision.

		Hazard Ratio	Event Count	No Event Count	p-value
Totals			1670	21111	
Antibiotic	Plain Cement	ref	1007	8649	<0.001
	ALBC	0.499 (0.452, 0.552)	663	12462	
Sex	Male	ref	497	6360	n.s.
	Female	1.063 (0.949, 1.190)	1173	14751	
Age		0.949 (0.943, 0.955)	68.80 (7.88)	72.33 (7.78)	<0.001
Cement viscosity	High	ref	946	12007	n.s.
	Low	1.028 (0.945,1.119)	209	4298	
	Medium	1.043 (0.918,1.184)	505	4695	
	Not informed	1.014 (0.972,1.058)	10	111	
Hospital category	1	0.954 (0.996,0.915)	290	4247	0.032
	2	ref	306	3863	
	3	0.910 (0.992,0.837)	472	6183	
	4	1.048 (1.004,1.093)	586	6508	
	5	1.098 (1.008,1.195)	16	310	
Surgery duration		1.003 (1.001, 1.005)	91.36 (19.84)	89.80 (18.52)	0.008
Alcohol abuse	No	ref	1651	20975	0.02
	Yes	1.726 (1.091, 2.729)	19	136	
Diabetes	No	ref	1426	17579	n.s.
	Yes	0.935 (0.815, 1.072)	244	3532	
Obesity	No	ref	1452	18575	n.s.
	Yes	1.132 (0.971, 1.319)	218	2536	
Rheumatoid arthritis	No	ref	1640	20637	n.s.
	Yes	0.755 (0.525, 1.084)	30	474	
Smoking status	Non smoker	ref	1334	17077	ref
	Smoker	0.821 (0.671, 1.005)	114	1285	0.056
	Former smoker	1.087 (0.935, 1.264)	222	2749	0.276
BMI		0.979 (0.968, 0.990)	31.85 (4.67)	31.78 (4.76)	<0.001

Table C.2: Cox proportional hazards model results for all-cause revision.

		Hazard Ratio	Event Count	No Event Count	p-value
Totals			2328	20453	
Antibiotic	Plain Cement	ref	1351	8305	<0.001
	ALBC	0.549 (0.504, 0.597)	977	12148	
Sex	Male	ref	767	6090	0.027
	Female	0.899 (0.819, 0.988)	1561	14363	
Age		0.963 (0.958, 0.968)	69.58 (8.10)	72.35 (7.76)	<0.001
Cement viscosity	High	ref	1308	11645	0.052
	Low	1.073 (1.000,1.153)	281	4226	
	Medium	1.112 (1.000,1.239)	715	4485	
	Not informed	1.036 (1.000,1.074)	24	97	
Hospital category	1	0.994 (1.031,0.959)	419	4118	n.s.
	2	ref	446	3723	
	3	0.988 (1.063,0.919)	680	5975	
	4	1.006 (0.970,1.043)	759	6335	
	5	1.012 (0.941,1.088)	24	302	
Surgery duration		1.003 (1.002, 1.005)	91.74 (23.81)	89.70 (17.93)	<0.001
Alcohol abuse	No	ref	2295	20331	<0.001
	Yes	1.973 (1.391, 2.797)	33	122	
Diabetes	No	ref	1951	17054	n.s.
	Yes	1.008 (0.902, 1.127)	377	3399	
Obesity	No	ref	1993	18034	0.001
	Yes	1.239 (1.093, 1.404)	335	2419	
Rheumatoid arthritis	No	ref	2272	20005	n.s.
	Yes	1.071 (0.821, 1.397)	56	448	
Smoking status	Non smoker	ref	1833	16578	n.s.
	Smoker	0.896 (0.757, 1.060)	165	1234	
	Former smoker	1.102 (0.973, 1.249)	330	2641	
BMI		0.988 (0.979, 0.998)	31.97 (4.81)	31.76 (4.75)	0.013

Appendix D

Plots of CATEs of all confounders in the problem of the effect of antibiotic-loaded bone cement on knee prosthetic survival

Figures D.1 to D.14 present the CATE plots for the different confounders. For continuous confounders, the CATE is represented as a contour map, with values coded by colors in a color bar. The x axis contains the time horizon and the y axis the confounder values. For categorical confounders, we depict as many lines as categories, the x axis containing the time horizon and the y axis the CATE value. Recall the notes about interpretation of the CATEs: 1) all CATEs represent the *increase* (if positive) or *decrease* (if negative) of the survival probability of prostheses, as a consequence of the usage of antibiotic-loaded bone cement, for given time horizons and for given subpopulations based on confounders; and 2) all CATEs are expressed as fractions of 1.

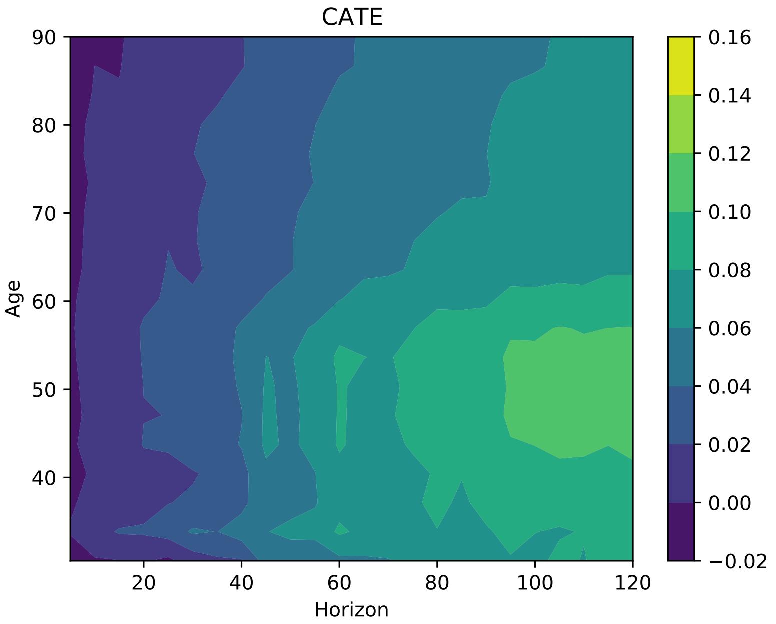


Figure D.1: CATE for age. The impact of age on the CATE can be seen for medium and long time horizons, with younger patients getting a bigger benefit in prosthetic survival from using antibiotic-loaded bone cement. Thus, above 60 months important differences in the CATE between patients below 60 years and above 60 years can be observed, the former having a CATE of 7.5%-10% and the latter of 5%-7.5%. Values above 90 years might not be reliable due to the small sample size.

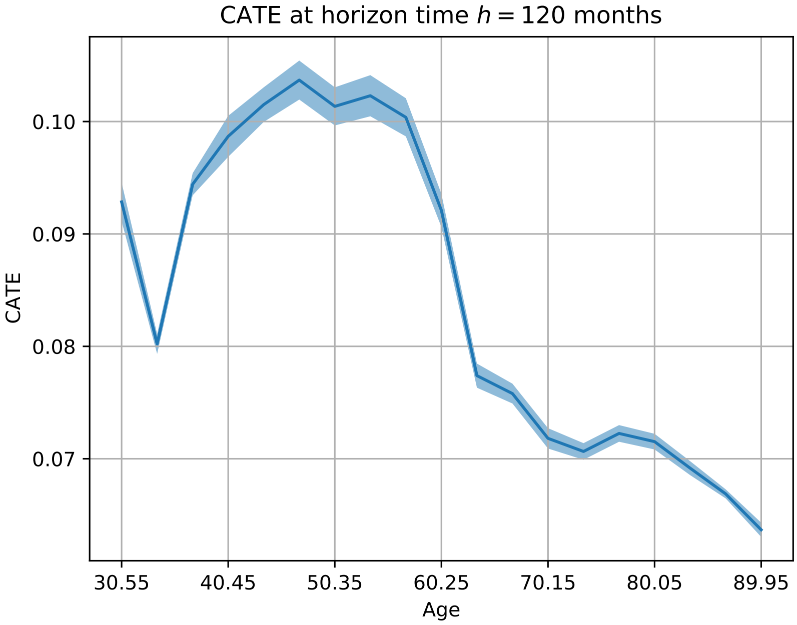


Figure D.2: CATE for age, at the specific value of horizon time of $h = 120$ months.

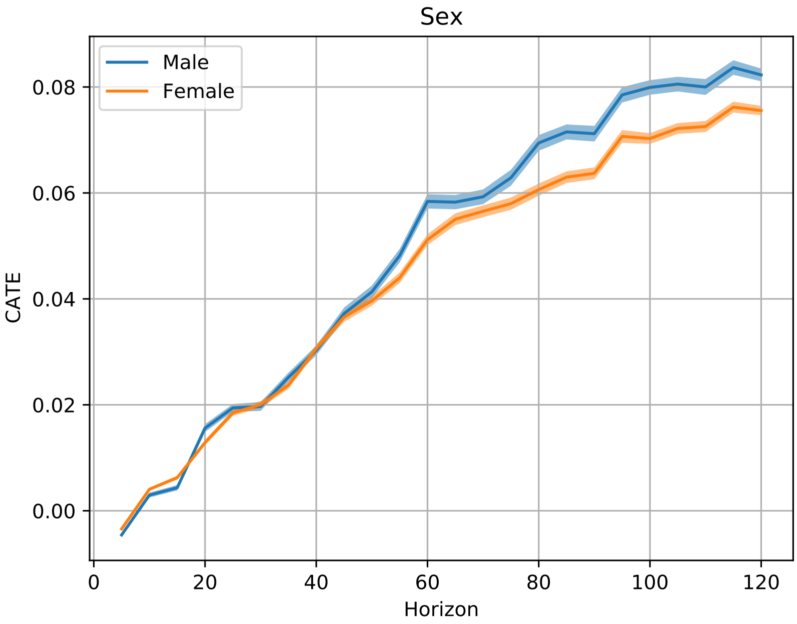


Figure D.3: CATE for sex. Sex does not have an impact on short-term horizons. A small effect becomes observable from 50 months on but remains smaller than 1 percentage point difference for the rest of the study time.

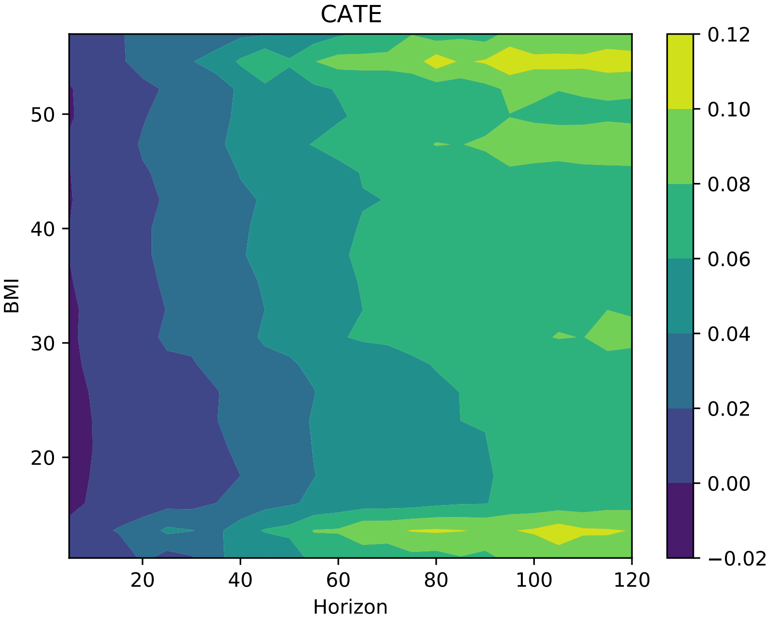


Figure D.4: CATE for BMI. Patients with bigger values of BMI (above 40) benefit more from antibiotic-loaded bone cement, especially for longer horizon times. Differences range between 8%-12% for a BMI value above 40 and between 6%-8% below that value.

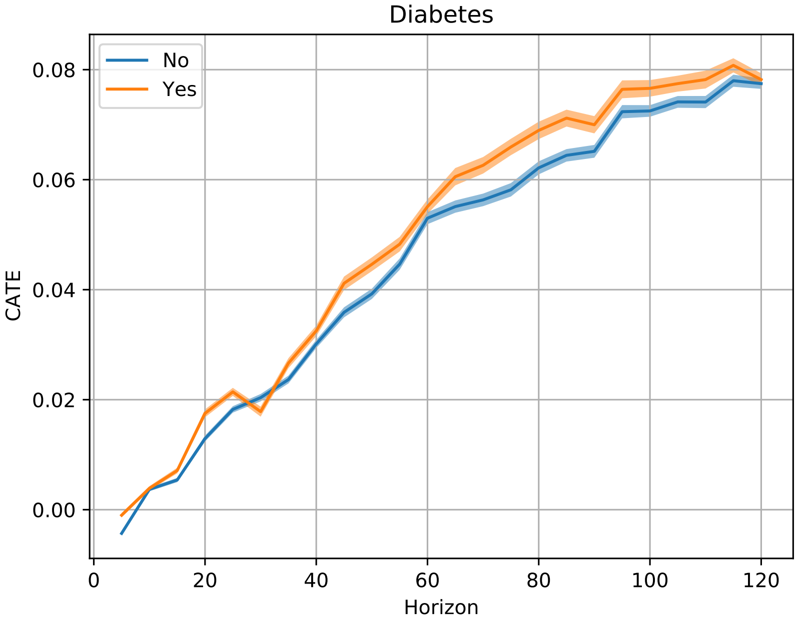


Figure D.5: There are almost no differences for short time horizons. Then some small differences appear, remaining always below 1 porcentual point. Patients with the disease benefit more from antibiotic-loaded bone cement use.

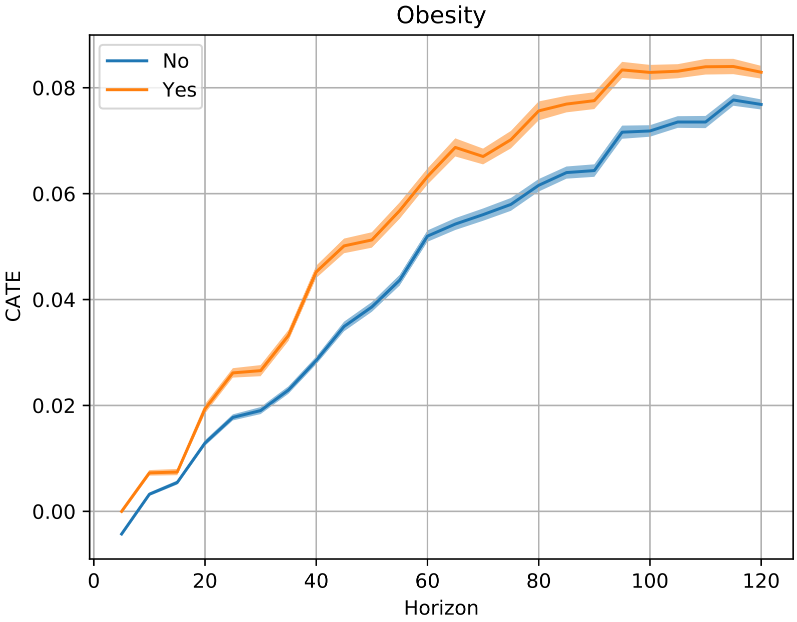


Figure D.6: CATE for obesity. Small differences for short time horizons that then amplify, but always remaining slightly above 1 percentage point. Patients with the disease benefit more from antibiotic-loaded bone cement use.

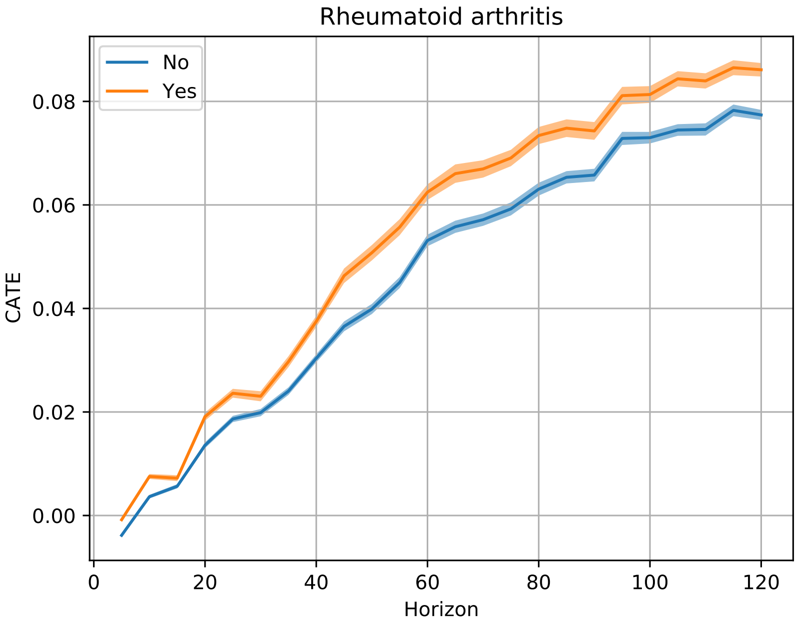


Figure D.7: CATE for rheumatoid arthritis. Small differences for short time horizons that then amplify, but always remaining around 1 percentage point. Patients with the disease benefit more from antibiotic-loaded bone cement use.

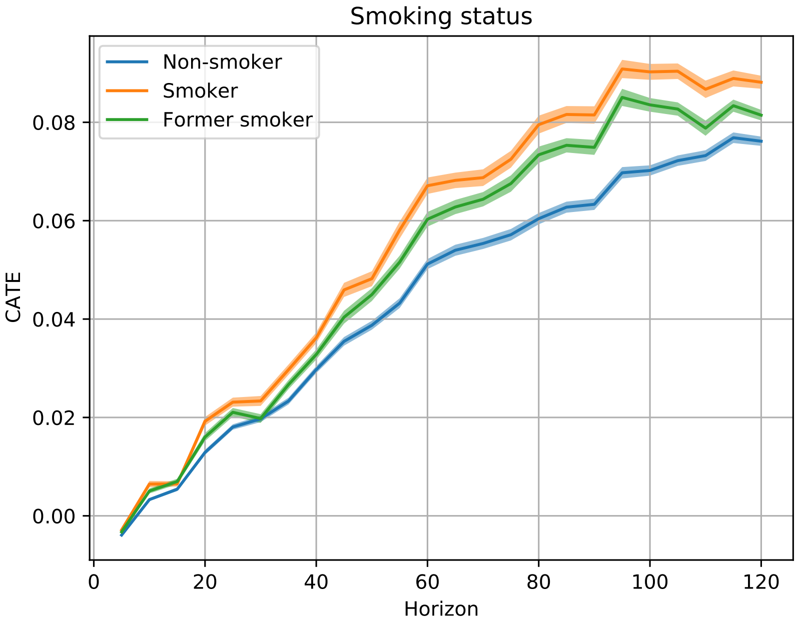


Figure D.8: CATE for smoking status. 0 for a non-smoker, 1 for a former smoker, and 2 for a smoker at the time of surgery. Patients who smoke or used to smoke benefit more from the use of antibiotic-loaded bone cement, with differences increasing with horizon time.

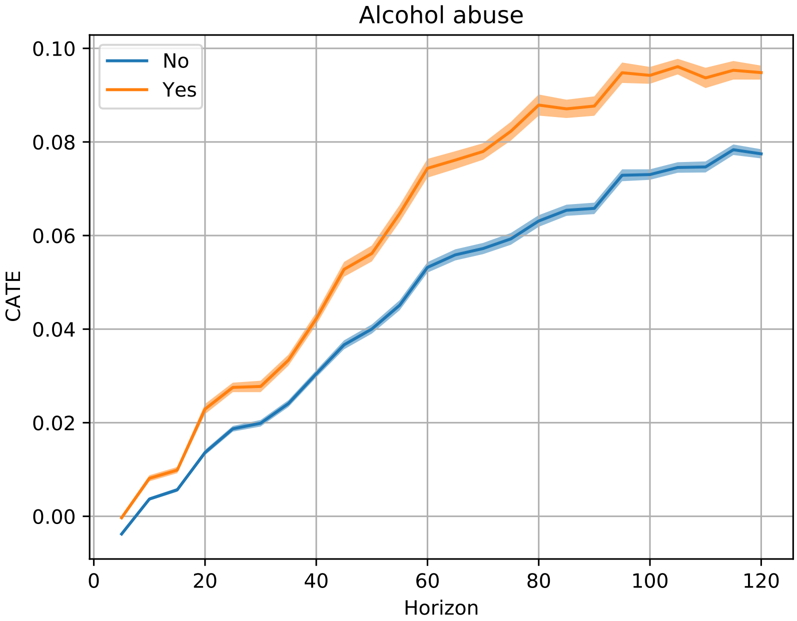


Figure D.9: CATE for alcohol abuse. Patients who abuse alcohol benefit more from the use of antibiotic-loaded bone cement, differences increase with horizon time and reach values above 2 porcentual points.

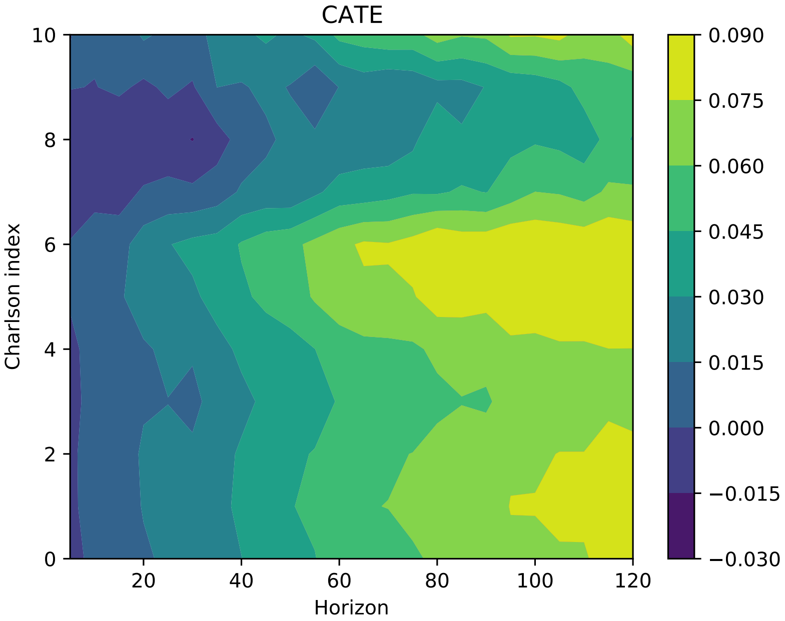


Figure D.10: CATE of the Charlson index.

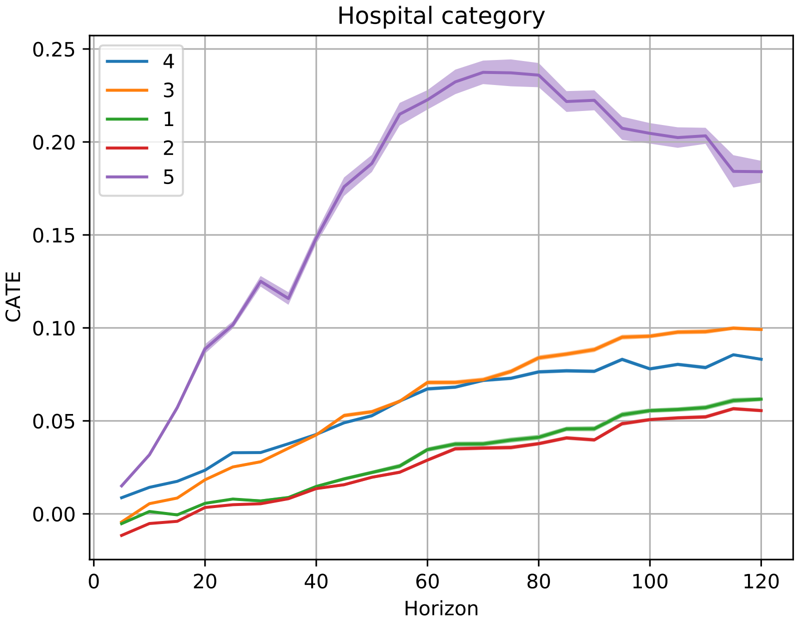


Figure D.11: CATE of the hospital category.

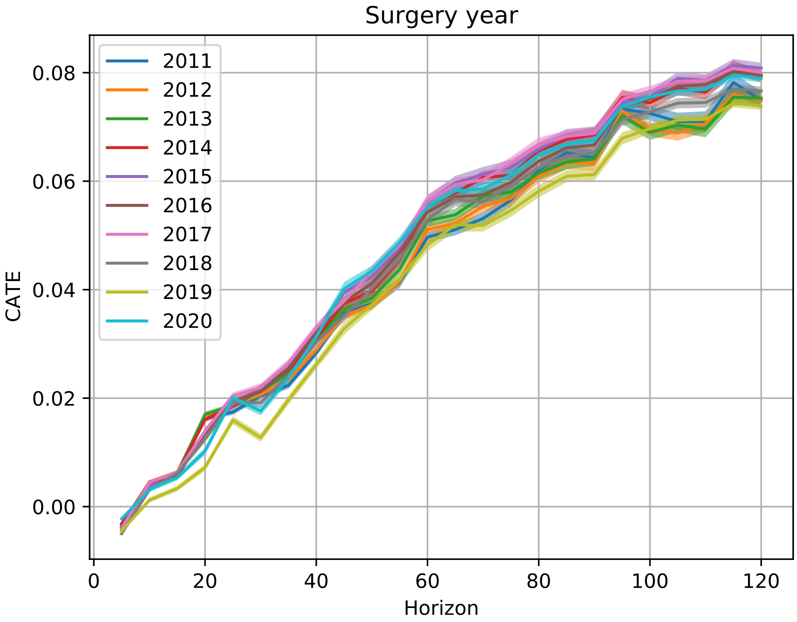


Figure D.12: CATE of the surgery year.

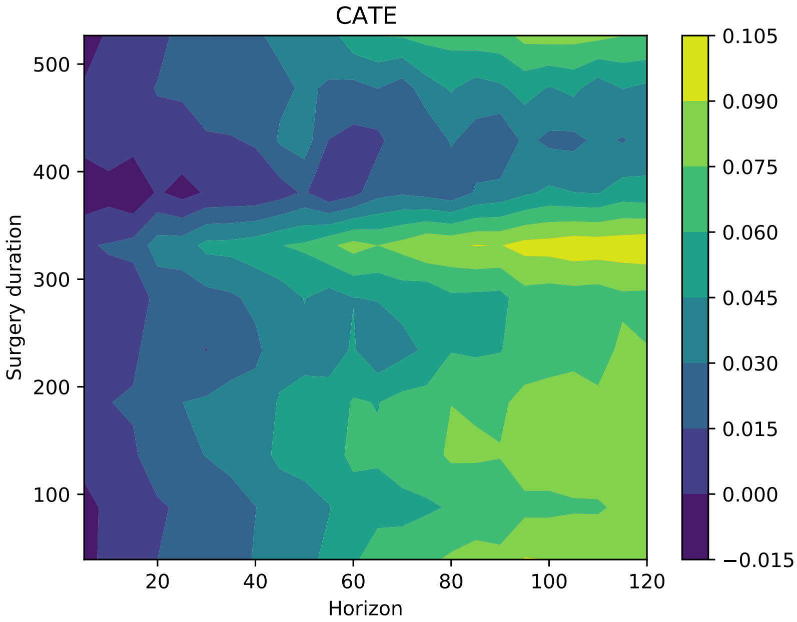


Figure D.13: CATE of the surgery duration.

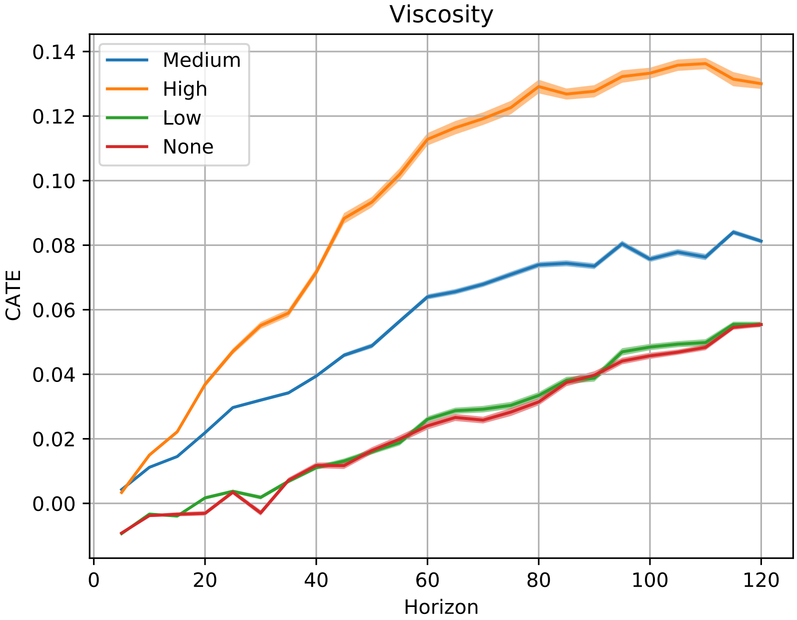


Figure D.14: CATE of the cement viscosity.

Bibliography

Aalen, Odd O. and Søren Johansen (1978). “An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations”. In: *Scandinavian Journal of Statistics* 5.3, pp. 141–150 (cit. on p. 56).

Abd ElHafeez, Samar, Graziella D’Arrigo, Daniela Leonardis, Maria Fusaro, Giovanni Tripepi, and Stefanos Roumeliotis (2021). “Methods to Analyze Time-to-Event Data: The Cox Regression Analysis”. en. In: *Oxidative Medicine and Cellular Longevity* 2021.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/1302811, p. 1302811. DOI: 10.1155/2021/1302811 (cit. on p. 56).

Akobeng, A K (2005). “Principles of evidence based medicine”. en. In: *Archives of Disease in Childhood* 90.8, pp. 837–840. DOI: 10.1136/adc.2005.071761 (cit. on p. 35).

Alsudais, Ali S, Raghad S Alkanani, Abdulaziz B Fathi, Saleh S Almuntashiri, Jafar N Jamjoom, Mustafa A Alzhrani, Alaa Althubaiti, and Suhaib Radi (2023). “Autoimmune diabetes mellitus after COVID-19 vaccination in adult population: a systematic review of case reports”. en. In: *BMC Endocrine Disorders* 23.1, p. 164. DOI: 10.1186/s12902-023-01424-0 (cit. on p. 22).

Amer, Samar A. et al. (2024). “Exploring the reported adverse effects of COVID-19 vaccines among vaccinated Arab populations: a multi-national survey study”. en. In: *Scientific Reports* 14.1. Publisher: Nature Publishing Group, p. 4785. DOI: 10.1038/s41598-024-54886-0 (cit. on pp. 21, 26).

Angrist, Joshua D. and Alan B. Krueger (1991). “Does Compulsory School Attendance Affect Schooling and Earnings?*”. In: *The Quarterly Journal of Economics* 106.4, pp. 979–1014. DOI: 10.2307/2937954 (cit. on p. 11).

APP Lunar (2024). https://www.lunarcomunidad.com/ [Accessed: 14/9/2024] (cit. on pp. 27, 95).

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). “Generalized random forests”. en. In: *The Annals of Statistics* 47.2. DOI: 10.1214/18-AOS1709 (cit. on p. 18).

Austin, Peter C. (2018). “Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes”. en. In: *Statistics in Medicine* 37.11, pp. 1874–1894. DOI: 10.1002/sim.7615 (cit. on p. 20).

Austin, Peter C., Hein Putter, Douglas S. Lee, and Ewout W. Steyerberg (2022). “Estimation of the Absolute Risk of Cardiovascular Disease and Other Events: Issues With the Use of Multiple

Fine-Gray Subdistribution Hazard Models". en. In: *Circulation: Cardiovascular Quality and Outcomes* 15.2, e008368. DOI: 10.1161/CIRCOUTCOMES.121.008368 (cit. on p. 74).

Balagopalan, Aparna, Ioana Baldini, Leo Anthony Celi, Judy Gichoya, Liam G. McCoy, Tristan Naumann, Uri Shalit, Mihaela van der Schaar, and Kiri L. Wagstaff (2024). "Machine learning for healthcare that matters: Reorienting from technical novelty to equitable impact". en. In: *PLOS Digital Health* 3.4. Publisher: Public Library of Science, e0000474. DOI: 10.1371/journal.pdig.0000474 (cit. on p. 19).

Belt, H.D. van de, D. Neut, W. Schenk, J.R. Horn, H.C.D. Mei, and H.J. Busscher (2000). "Gentamicin release from polymethylmethacrylate bone cements and Staphylococcus aureus biofilm formation". In: *Acta Orthopaedica Scandinavica* 71.6, pp. 625–629. DOI: 10.1080/000164700317362280 (cit. on p. 60).

Bendich, I., N. Zhang, J.J. Barry, D.T. Ward, M.A. Whooley, and A.C. Kuo (2020). "Antibiotic-laden bone cement use and revision risk after primary total knee arthroplasty in US veterans". In: *Journal of Bone and Joint Surgery* 102.22, pp. 1939–1947. DOI: 10.2106/JBJS.20.00102 (cit. on pp. 23, 61).

Berrevoets, Jeroen, Krzysztof Kacprzyk, Zhaozhi Qian, and Mihaela van der Schaar (2024). *Causal Deep Learning*. en. arXiv:2303.02186 [cs] (cit. on p. 11).

Bica, Ioana, James Jordon, and Mihaela van der Schaar (2020). *Estimating the Effects of Continuous-valued Interventions using Generative Adversarial Networks*. en. arXiv:2002.12326 [cs, stat] (cit. on p. 20).

Bohm, Eric, Naisu Zhu, Jing Gu, Nicole de Guia, Cassandra Linton, Tammy Anderson, David Paton, and Michael Dunbar (2014). "Does Adding Antibiotics to Cement Reduce the Need for Early Revision in Total Knee Arthroplasty?" en. In: *Clinical Orthopaedics & Related Research* 472.1, pp. 162–168. DOI: 10.1007/s11999-013-3186-1 (cit. on pp. 23, 51, 54, 61).

Bonneville, Edouard F, Liesbeth C de Wreede, and Hein Putter (2024). "Why you should avoid using multiple Fine–Gray models: insights from (attempts at) simulating proportional subdistribution hazards data". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 187.3, pp. 580–593. DOI: 10.1093/jrsssa/qnae056 (cit. on p. 74).

Bozzo, A., S. Ekhtiari, K. Madden, M. Bhandari, M. Ghert, and V. et al. Khanna (2022). "Incidence and predictors of prosthetic joint infection following primary total knee arthroplasty: a 15-year population-based cohort study". In: *The Journal of Arthroplasty* 37.2, pp. 367–372. DOI: 10.1016/j.arth.2021.10.006 (cit. on p. 61).

Brachman, Philip S. (1996). "Epidemiology". en. In: *Medical Microbiology. 4th edition*. University of Texas Medical Branch at Galveston (cit. on p. 10).

Bradburn, M J, T G Clark, S B Love, and D G Altman (2003). "Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods". en. In: *British Journal of Cancer* 89.3, pp. 431–436. DOI: 10.1038/sj.bjc.6601119 (cit. on p. 13).

Brost, Brian, Rishabh Mehrotra, and Tristan Jehan (2020). “The Music Streaming Sessions Dataset”. en. In: *arXiv:1901.09851 [cs]*. arXiv: 1901.09851 (cit. on p. 20).

Cain, Lauren E., James M. Robins, Emilie Lanoy, Roger Logan, Dominique Costagliola, and Miguel A. Hernán (2010). “When to Start Treatment? A Systematic Approach to the Comparison of Dynamic Regimes Using Observational Data”. en. In: *The International Journal of Biostatistics* 6.2. Publisher: De Gruyter. DOI: 10.2202/1557-4679.1212 (cit. on p. 16).

Card, David and Alan B Krueger (1994). “American economic association”. In: *The American Economic Review* 84.4, pp. 772–793 (cit. on p. 11).

Cattaneo, Matias D. (2010). “Efficient semiparametric estimation of multi-valued treatment effects under ignorability”. en. In: *Journal of Econometrics* 155.2, pp. 138–154. DOI: 10.1016/j.jeconom.2009.09.023 (cit. on pp. 19, 80).

Cattaneo, Matias D., David M. Drukker, and Ashley D. Holland (2013). “Estimation of Multivalued Treatment Effects under Conditional Independence”. en. In: *The Stata Journal: Promoting communications on statistics and Stata* 13.3, pp. 407–450. DOI: 10.1177/1536867X1301300301 (cit. on p. 19).

CausalFusion (2024) (cit. on p. 63).

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, and Hansen (2018). “Double/debiased machine learning for treatment and structural parameters”. In: *The Econometrics Journal* 21.1, pp. C1–C68. DOI: 10.1111/ectj.12097 (cit. on pp. 18, 81).

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey (2017). “Double/Debiased/Neyman Machine Learning of Treatment Effects”. In: *American Economic Review* 107.5, pp. 261–65. DOI: 10.1257/aer.p20171038 (cit. on pp. 18, 81).

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2017). “Double/Debiased Machine Learning for Treatment and Causal Parameters”. en. In: *arXiv:1608.00060 [econ, stat]*. arXiv: 1608.00060 (cit. on p. 82).

Chiu, Yu-Han, Lan Wen, Sean McGrath, Roger Logan, Issa J Dahabreh, and Miguel A Hernán (2023). “Evaluating Model Specification When Using the Parametric G-Formula in the Presence of Censoring”. In: *American Journal of Epidemiology* 192.11, pp. 1887–1895. DOI: 10.1093/aje/kwad143 (cit. on p. 40).

Clark, T G, M J Bradburn, S B Love, and D G Altman (2003). “Survival Analysis Part I: Basic concepts and first analyses”. en. In: *British Journal of Cancer* 89.2, pp. 232–238. DOI: 10.1038/sj.bjc.6601118 (cit. on p. 13).

Cole, Stephen R. and Miguel A. Hernán (2008). “Constructing Inverse Probability Weights for Marginal Structural Models”. In: *American Journal of Epidemiology* 168.6, pp. 656–664. DOI: 10.1093/aje/kwn164 (cit. on p. 16).

Critchley, Hilary O.D. et al. (2020). “Menstruation: science and society”. In: *American Journal of Obstetrics and Gynecology* 223.5, pp. 624–664. DOI: https://doi.org/10.1016/j.ajog.2020.06.004 (cit. on pp. 26, 27).

Cui, Yifan, Michael R. Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu (2023). *Estimating heterogeneous treatment effects with right-censored data via causal survival forests*. en. arXiv:2001.09887 [cs, stat] (cit. on pp. 17–19, 63).

Curth, Alicia, David Svensson, and James Weatherall (2021). “Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation”. en. In: (cit. on p. 21).

Dorie, Vincent, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone (2018). “Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition”. en. In: *arXiv:1707.02641 [stat]*. arXiv: 1707.02641 (cit. on pp. 84, 86).

Dunne, N., J. Hill, P. McAfee, K. Todd, R. Kirkpatrick, and M. et al. Tunney (2007). “In vitro study of the efficacy of acrylic bone cement loaded with supplementary amounts of gentamicin: effect on mechanical properties, antibiotic release, and biofilm formation”. In: *Acta Orthopaedica* 78.6, pp. 774–785. DOI: 10.1080/17453670710014545 (cit. on p. 50).

Edelman, Alison, Emily R. Boniface, Eleonora Benhar, Leo Han, Kristen A. Matteson, Carlotta Favaro, Jack T. Pearson, and Blair G. Darney (2022). “Association Between Menstrual Cycle Length and Coronavirus Disease 2019 (COVID-19) Vaccination: A U.S. Cohort”. en-US. In: *Obstetrics & Gynecology* 139.4, p. 481. DOI: 10.1097/AOG.0000000000004695 (cit. on pp. 21, 27, 29, 30).

Edelman, Alison, Emily R. Boniface, Victoria Male, Sharon Cameron, Eleonora Benhar, Leo Han, Kristen A. Matteson, Agathe van Lamsweerde, Jack T. Pearson, and Blair G. Darney (2024). “Timing of Coronavirus Disease 2019 (COVID-19) Vaccination and Effects on Menstrual Cycle Changes”. eng. In: *Obstetrics and Gynecology* 143.4, pp. 585–594. DOI: 10.1097/AOG.0000000000005550 (cit. on pp. 21, 31).

Elm, E. von, D.G. Altman, M. Egger, S.J. Pocock, P.C. Gøtzsche, and J.P. Vandenbroucke (2008). “The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies”. In: *Journal of Clinical Epidemiology* 61.4, pp. 344–349. DOI: 10.1016/j.jclinepi.2007.11.008 (cit. on pp. 28, 52).

ElSayed, Nuha A. et al. (2023). “2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes—2023”. In: *Diabetes Care* 46.Suppl 1, S19–S40. DOI: 10.2337/dc23-S002 (cit. on p. 38).

Engesæter, Lars B, Birgitte Espehaug, Stein Atle Lie, Ove Furnes, and Leif Ivar Havelin (2006). “Does cement increase the risk of infection in primary total hip arthroplasty? Revision rates in 56,275 cemented and uncemented primary THAs followed for 0–16 years in the Norwegian Arthroplasty Register”. en. In: *Acta Orthopaedica* 77.3, pp. 351–358. DOI: 10.1080/17453670610046253 (cit. on p. 50).

Esposti, Roberto (2017). "The heterogeneous farm-level impact of the 2005 CAP-first pillar reform: A multivalued treatment effect estimation". en. In: *Agricultural Economics* 48.3, pp. 373–386. DOI: 10.1111/agec.12340 (cit. on p. 20).

Feuerriegel, Stefan, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar (2024). "Causal machine learning for predicting treatment outcomes". en. In: *Nature Medicine* 30.4. Publisher: Nature Publishing Group, pp. 958–968. DOI: 10.1038/s41591-024-02902-1 (cit. on p. 19).

Fine, Jason P. and Robert J. Gray (1999). "A Proportional Hazards Model for the Subdistribution of a Competing Risk". In: *Journal of the American Statistical Association* 94.446, pp. 496–509 (cit. on p. 56).

Frank, R., M. Cross, and C. Della Valle (2015). "Periprosthetic joint infection: modern aspects of prevention, diagnosis, and treatment". In: *Journal of Knee Surgery* 28.2, pp. 105–112. DOI: 10.1055/s-0034-1396015 (cit. on p. 50).

Frank, Rachel, Michael Cross, and Craig Della Valle (2014). "Periprosthetic Joint Infection: Modern Aspects of Prevention, Diagnosis, and Treatment". en. In: *Journal of Knee Surgery* 28.02, pp. 105–112. DOI: 10.1055/s-0034-1396015 (cit. on p. 23).

Franklin, Jessica M., Sebastian Schneeweiss, Jennifer M. Polinski, and Jeremy A. Rassen (2014). "Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases". en. In: *Computational Statistics & Data Analysis* 72, pp. 219–226. DOI: 10.1016/j.csda.2013.10.018 (cit. on p. 20).

Gaber, Charles E., Kent A. Hanson, Sodam Kim, Jennifer L. Lund, Todd A. Lee, and Eleanor J. Murray (2024). "The Clone-Censor-Weight Method in Pharmacoepidemiologic Research: Foundations and Methodological Implementation". en. In: *Current Epidemiology Reports* 11.3, pp. 164–174. DOI: 10.1007/s40471-024-00346-2 (cit. on pp. 16, 36).

Garrido, Melissa M., Jessica Lum, and Steven D. Pizer (2021). "¡span style="font-variant:small-caps;"¿Vector-based¡/span¿ kernel weighting: A simple estimator for improving precision and bias of average treatment effects in multiple treatment settings". en. In: *Statistics in Medicine* 40.5, pp. 1204–1223. DOI: 10.1002/sim.8836 (cit. on p. 20).

Garvin, K.L. and B.S. Konigsberg (2011). "Infection following total knee arthroplasty: prevention and management". In: *The Journal of Bone & Joint Surgery* 93.12, pp. 1167–1175. DOI: 10.2106/00004623-201106150-00012 (cit. on p. 50).

Gershman, Samuel J. and Tomer D. Ullman (2023). "Causal implicatures from correlational statements". en. In: *PLOS ONE* 18.5. Publisher: Public Library of Science, e0286067. DOI: 10.1371/journal.pone.0286067 (cit. on pp. 2, 32, 51, 94).

Gerstman, B. Burt (2023). "There is no single gold standard study design (RCTs are not the gold standard)". In: *Expert Opinion on Drug Safety* 22.4. Publisher: Taylor & Francis, pp. 267–270. DOI: 10.1080/14740338.2023.2203488 (cit. on p. 2).

Graham, Bryan S. and Cristine Campos de Xavier Pinto (2022). “Semiparametrically efficient estimation of the average linear regression function”. en. In: *Journal of Econometrics* 226.1, pp. 115–138. DOI: 10.1016/j.jeconom.2021.07.008 (cit. on p. 20).

GRF (2024) (cit. on p. 64).

Gross, Ruth T. (1993). “Infant Health and Development Program (IHDP): Enhancing the Outcomes of Low Birth Weight, Premature Infants in the United States, 1985-1988”. In: DOI: 10.3886/ICPSR09795.v1 (cit. on p. 84).

Guyatt, Gordon, Andrew D. Oxman, Elie A. Akl, Regina Kunz, Gunn Vist, Jan Brozek, Susan Norris, Yngve Falck-Ytter, Paul Glasziou, and Hans deBeer (2011). “GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables”. en. In: *Journal of Clinical Epidemiology* 64.4, pp. 383–394. DOI: 10.1016/j.jclinepi.2010.04.026 (cit. on p. 13).

Hariton, Eduardo and Joseph J Locascio (2018). “Randomised controlled trials – the gold standard for effectiveness research”. en. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 125.13. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1471-0528.15199, pp. 1716–1716. DOI: 10.1111/1471-0528.15199 (cit. on p. 2).

Harrell, Frank E. (2001). “Parametric Survival Models”. In: *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer New York, pp. 413–442. DOI: 10.1007/978-1-4757-3462-1_17 (cit. on p. 14).

He, Yan-Fei, Jing Ouyang, Xiao-Dong Hu, Ni Wu, Zhi-Gang Jiang, Ning Bian, and Jie Wang (2023). “Correlation between COVID-19 vaccination and diabetes mellitus: A systematic review”. en. In: *World Journal of Diabetes* 14.6, pp. 892–918. DOI: 10.4239/wjd.v14.i6.892 (cit. on pp. 22, 34).

Hernán, Miguel A and James M Robins (2020). *Causal Inference: What If*. en (cit. on pp. 40, 50, 79).

Hernán, Miguel A. (2018). “The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data”. In: *American Journal of Public Health* 108.5, pp. 616–619. DOI: 10.2105/AJPH.2018.304337 (cit. on p. 51).

Hernán, Miguel A. and James 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, pp. 758–764. DOI: 10.1093/aje/kwv254 (cit. on pp. 11, 35).

Hernán, Miguel A., Wei Wang, and David E. Leaf (2022). “Target Trial Emulation: A Framework for Causal Inference From Observational Data”. In: *JAMA* 328.24, pp. 2446–2447. DOI: 10.1001/jama.2022.21383 (cit. on p. 36).

Hinarejos, Pedro, Pau Guirro, Joan Leal, Ferran Montserrat, Xavier Pelfort, M.L. Sorli, J.P. Horcajada, and Lluis Puig (2013). “The Use of Erythromycin and Colistin-Loaded Cement in Total Knee Arthroplasty Does Not Reduce the Incidence of Infection: A Prospective Randomized Study in 3000 Knees”. en. In: *Journal of Bone and Joint Surgery* 95.9, pp. 769–774. DOI: 10.2106/JBJS.L.00901 (cit. on pp. 23, 50, 60–62).

Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960 (cit. on p. 14).

Hong, Guanglei (2012). "Marginal mean weighting through stratification: A generalized method for evaluating multivalued and multiple treatments with nonexperimental data." en. In: *Psychological Methods* 17.1, pp. 44–60. DOI: 10.1037/a0024918 (cit. on p. 20).

Hoskins, T., J.K. Shah, J. Patel, C. Mazzei, D. Goyette, and E. et al. Poletick (2020). "The cost-effectiveness of antibiotic-loaded bone cement versus plain bone cement following total and partial knee and hip arthroplasty". In: *Journal of Orthopaedics* 20, pp. 217–220. DOI: 10.1016/j.jor.2020.01.029 (cit. on p. 50).

Hosseini, Mohammad-Salar, Farid Jahanshahlou, Mohammad Amin Akbarzadeh, Mahdi Zarei, and Yosra Vaez-Gharamaleki (2024). "Formulating research questions for evidence-based studies". In: *Journal of Medicine, Surgery, and Public Health* 2, p. 100046. DOI: https://doi.org/10.1016/j.glmedi.2023.100046 (cit. on p. 35).

Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62.2, pp. 467–475 (cit. on p. 11).

Imbens, Guido W. and Donald B. Rubin (2010). "Rubin Causal Model". en. In: *Microeconometrics*. Ed. by Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan UK, pp. 229–241. DOI: 10.1057/9780230280816_28 (cit. on pp. 4, 14).

Inoue, Kosuke, Susan Athey, and Yusuke Tsugawa (2023). "Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management". en. In: *International Journal of Epidemiology* 52.4, pp. 1243–1256. DOI: 10.1093/ije/dyad037 (cit. on p. 71).

Irwin, Alan (2024). *Citizen Science: A Study of People, Expertise and Sustainable Development*. en (cit. on p. 26).

Jameson, Simon S., Asaad Asaad, Marina Diament, Adetatyo Kasim, Theophile Bigirumurame, Paul Baker, James Mason, Paul Partington, and Mike Reed (2019). "Antibiotic-loaded bone cement is associated with a lower risk of revision following primary cemented total knee arthroplasty: an analysis of 731 214 cases using National Joint Registry data". en. In: *The Bone & Joint Journal* 101-B.11, pp. 1331–1347. DOI: 10.1302/0301-620X.101B11.BJJ-2019-0196.R1 (cit. on pp. 23, 51, 54, 61).

Jämsen, Esa, Heini Huhtala, Timo Puolakka, and Teemu Moilanen (2009). "Risk Factors for Infection After Knee Arthroplasty: A Register-Based Analysis of 43,149 Cases". en. In: *The Journal of Bone and Joint Surgery-American Volume* 91.1, pp. 38–47. DOI: 10.2106/JBJS.G.01686 (cit. on pp. 23, 60–62).

Kaddour, Jean, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva (2022). *Causal Machine Learning: A Survey and Open Problems*. en. arXiv:2206.15475 [cs, stat] (cit. on p. 19).

Kaddour, Jean, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva (2021). "Causal Effect Inference for Structured Treatments". en. In: (cit. on pp. 20, 79, 84).

Kennedy, Edward H. (2016). “Semiparametric Theory and Empirical Processes in Causal Inference”. In: *Statistical Causal Inferences and Their Applications in Public Health Research*. Ed. by Hua He, Pan Wu, and Ding-Geng (Din) Chen. Cham: Springer International Publishing, pp. 141–167. DOI: 10.1007/978-3-319-41259-7_8 (cit. on pp. 18, 81, 82).

King, J.D., D.H. Hamilton, C.A. Jacobs, and S.T. Duncan (2018). “The hidden cost of commercial antibiotic-loaded bone cement: a systematic review of clinical results and cost implications following total knee arthroplasty”. In: *The Journal of Arthroplasty* 33.12, pp. 3789–3792. DOI: 10.1016/j.arth.2018.08.009 (cit. on pp. 23, 50, 61).

Kuang, Kun, Yunzhe Li, Bo Li, Peng Cui, Hongxia Yang, Jianrong Tao, and Fei Wu (2021). “Continuous treatment effect estimation via generative adversarial de-confounding”. en. In: *Data Mining and Knowledge Discovery* 35.6, pp. 2467–2497. DOI: 10.1007/s10618-021-00797-x (cit. on p. 20).

Kuehne, Felicitas, Marjan Arvandi, Lisa M. Hess, Douglas E. Faries, Raffaella Matteucci Gothe, Holger Gothe, Julie Beyrer, Alain Gustave Zeimet, Igor Stojkov, Nikolai Mühlberger, Willi Oberaigner, Christian Marth, and Uwe Siebert (2022). “Causal analyses with target trial emulation for real-world evidence removed large self-inflicted biases: systematic bias assessment of ovarian cancer treatment effectiveness”. en. In: *Journal of Clinical Epidemiology* 152, pp. 269–280. DOI: 10.1016/j.jclinepi.2022.10.005 (cit. on pp. 11, 36, 47).

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu (2019). “Metalearners for estimating heterogeneous treatment effects using machine learning”. In: *Proceedings of the National Academy of Sciences* 116.10. _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1804597116, pp. 4156–4165. DOI: 10.1073/pnas.1804597116 (cit. on pp. 20, 79, 86).

Kurtz, S.M., K.L. Ong, E. Lau, K.J. Bozic, D. Berry, and J. Parvizi (2010). “Prosthetic joint infection risk after TKA in the Medicare population”. In: *Clinical Orthopaedics & Related Research* 468.1, pp. 52–56. DOI: 10.1007/s11999-009-1013-5 (cit. on pp. 23, 62).

Kwan, Alan C., Joseph E. Ebinger, Patrick Botting, Jesse Navarrette, Brian Claggett, and Susan Cheng (2023). “Association of COVID-19 Vaccination With Risk for Incident Diabetes After COVID-19 Infection”. In: *JAMA Network Open* 6.2, e2255965–e2255965. DOI: 10.1001/jamanetworkopen.2022.55965 (cit. on pp. 22, 34, 35).

Laird, Yvonne, Leah Marks, Ben J Smith, Pippy Walker, Kate Garvey, Kim Jose, Sean O’Rourke, Katherine Pontifex, Karen Wardle, and Samantha Rowbotham (2023). “Harnessing citizen science in health promotion: perspectives of policy and practice stakeholders in Australia”. en. In: *Health Promotion International* 38.5, daad101. DOI: 10.1093/heapro/daad101 (cit. on p. 27).

Lampe, Kristian, Marjukka Mäkelä, Marcial Velasco Garrido, Heidi Anttila, Ilona Autti-Rämö, Nicholas J. Hicks, Björn Hofmann, Juha Koivisto, Regina Kunz, Pia Kärki, and et al. (2009). “The HTA Core Model: A novel method for producing and reporting health technology assessments”. In: *International Journal of Technology Assessment in Health Care* 25.S2, pp. 9–20. DOI: 10.1017/S0266462309990638 (cit. on p. 4).

Lautenschlager, E.P., J.J. Jacobs, G.W. Marshall, and P.R. Meyer (1976). “Mechanical properties of bone cements containing large doses of antibiotic powders”. In: *Journal of Biomedical Materials Research* 10, pp. 929–938. DOI: 10.1002/jbm.820100610 (cit. on p. 61).

Lee, Hae Sang and Jin Soon Hwang (2019). “Genetic Aspects of type 1 diabetes”. en. In: *Annals of Pediatric Endocrinology & Metabolism* 24.3, pp. 143–148. DOI: 10.6065/apem.2019.24.3.143 (cit. on p. 38).

Lee, Ying-Ying (2018). “Efficient propensity score regression estimators of multivalued treatment effects for the treated”. en. In: *Journal of Econometrics* 204.2, pp. 207–222. DOI: 10.1016/j.jeconom.2018.02.002 (cit. on p. 20).

Lendle, Samuel David (2015). “Targeted Minimum Loss Based Estimation: Applications and Extensions in Causal Inference and Big Data”. en. PhD thesis. UC Berkeley (cit. on p. 82).

Leta, T.H., S.A. Lie, A.M. Fenstad, S.H.L. Lygre, M. Lindberg-Larsen, and A.B. et al. Pedersen (2024). “Periprosthetic Joint Infection After Total Knee Arthroplasty With or Without Antibiotic Bone Cement”. In: *JAMA Network Open* 7.5, e2412898. DOI: 10.1001/jamanetworkopen.2024.12898 (cit. on pp. 23, 61).

Leta, Tesfaye H et al. (2021). “Antibiotic-Loaded Bone Cement in Prevention of Periprosthetic Joint Infections in Primary Total Knee Arthroplasty: A Register-based Multicentre Randomised Controlled Non-inferiority Trial (ALBA trial)”. en. In: *Open access* (cit. on pp. 6, 50).

Li, Ang and Judea Pearl (2022). *Probabilities of Causation with Nonbinary Treatment and Effect*. en. arXiv:2208.09568 [cs] (cit. on p. 20).

Li, Hao-Qian, Peng-Cui Li, Xiao-Chun Wei, and Jun-Jun Shi (2022). “Effectiveness of antibiotics loaded bone cement in primary total knee arthroplasty: A systematic review and meta-analysis”. en. In: *Orthopaedics & Traumatology: Surgery & Research* 108.5, p. 103295. DOI: 10.1016/j.otsr.2022.103295 (cit. on pp. 23, 50, 61).

Li, Meng, Shengqi Chen, Yunfeng Lai, Zuanji Liang, Jiaqi Wang, Junnan Shi, Haojie Lin, Dongning Yao, Hao Hu, and Carolina Oi Lam Ung (2021). “Integrating Real-World Evidence in the Regulatory Decision-Making Process: A Systematic Analysis of Experiences in the US, EU, and China Using a Logic Model”. In: *Frontiers in Medicine* 8. DOI: 10.3389/fmed.2021.669509 (cit. on p. 46).

Linden, Ariel and Paul R Yarnold (2016). “Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments: Machine learning and multi-valued treatments”. en. In: *Journal of Evaluation in Clinical Practice* 22.6, pp. 875–885. DOI: 10.1111/jep.12610 (cit. on p. 20).

Lix, L., M. Smith, M. Pitz, R. Ahmed, H. Quon, and J. et al. Griffith (2016). *Cancer data linkage in Manitoba: expanding the infrastructure for research*. Tech. rep. Winnipeg, MB: Manitoba Centre for Health Policy (cit. on p. 53).

Lopez, Michael J. and Roee Gutman (2017). “Estimation of causal effects with multiple treatments: a review and new ideas”. en. In: *Statistical Science* 32.3. arXiv:1701.05132 [stat]. DOI: 10.1214/17-STS612 (cit. on p. 20).

Louizos, Christos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling (2017). “Causal effect inference with deep latent-variable models”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., pp. 6449–6459 (cit. on pp. 17, 79).

Maathuis, P.G.M., D. Neut, H.J. Busscher, H.C. van der Mei, and J.R. van Horn (2005). “Perioperative contamination in primary total hip arthroplasty”. In: *Clinical Orthopaedics and Related Research* 433, pp. 136–139. DOI: 10.1097/01.blo.0000149997.14631.0c (cit. on p. 61).

Mackie, J. L. (1980). *The Cement of the Universe: A Study of Causation*. Oxford University Press. DOI: 10.1093/0198246420.001.0001 (cit. on p. 1).

McGrath, Sean, Victoria Lin, Zilu Zhang, Lucia C. Petito, Roger W. Logan, Miguel A. Hernán, and Jessica G. Young (2020). “gfoRmula: An R Package for Estimating the Effects of Sustained Treatment Strategies via the Parametric g-formula”. In: *Patterns* 1.3, p. 100008. DOI: https://doi.org/10.1016/j.patter.2020.100008 (cit. on pp. 16, 47).

McGrath, Sean, Jessica G. Young, and Miguel A. Hernán (2022). “Revisiting the g-null Paradox”. en-US. In: *Epidemiology* 33.1, p. 114. DOI: 10.1097/EDE.0000000000001431 (cit. on p. 46).

Mihm, M., S. Gangooly, and S. Muttukrishna (2011). “The normal menstrual cycle in women”. In: *Animal Reproduction Science*. Special Issue: Reproductive Cycles of Animals 124.3, pp. 229–236. DOI: 10.1016/j.anireprosci.2010.08.030 (cit. on pp. 5, 27, 29).

Monterde, David, Emili Vela, and Montse Clèries (2016). “Los grupos de morbilidad ajustados: nuevo agrupador de morbilidad poblacional de utilidad en el ámbito de la atención primaria”. In: *Atención Primaria* 48.10, pp. 674–682. DOI: https://doi.org/10.1016/j.aprim.2016.06.003 (cit. on pp. 37, 43).

Moran, J.M., A.S. Greenwald, and M.B. Matejczyk (1979). “Effect of gentamicin on shear and interface strengths of bone cement”. In: *Clinical Orthopaedics and Related Research* 141, pp. 96–101 (cit. on p. 61).

Murray, Eleanor J, Ellen C Caniglia, and Lucia C Petito (2021). “Causal survival analysis: A guide to estimating intention-to-treat and per-protocol effects from randomized clinical trials with non-adherence”. en. In: *Research Methods in Medicine & Health Sciences* 2.1, pp. 39–49. DOI: 10.1177/2632084320961043 (cit. on p. 16).

Naimi, Ashley I, Stephen R Cole, and Edward H Kennedy (2017). “An introduction to g methods”. In: *International Journal of Epidemiology* 46.2, pp. 756–762. DOI: 10.1093/ije/dyw323 (cit. on p. 64).

Namba, R.S., Y. Chen, E.W. Paxton, T. Slipchenko, and D.C. Fithian (2009). “Outcomes of routine use of antibiotic-loaded cement in primary total knee arthroplasty”. In: *The Journal of Arthroplasty* 24.6 Suppl, pp. 44–47. DOI: 10.1016/j.arth.2009.05.007 (cit. on pp. 23, 61).

Namba, R.S., M.C.S. Inacio, and E.W. Paxton (2013). “Risk factors associated with deep surgical site infections after primary total knee arthroplasty: an analysis of 56,216 knees”. In: *Journal of Bone and Joint Surgery* 95.9, pp. 775–782. DOI: 10.2106/JBJS.L.00211 (cit. on pp. 23, 50, 62).

Nazir, Maheen, Shumaila Asghar, Muhammad Ali Rathore, Asima Shahzad, Anum Shahid, Alishba Ashraf Khan, Asmara Malik, Tehniat Fakhar, Hafsa Kausar, and Jahanzeb Malik (2022). “Menstrual abnormalities after COVID-19 vaccines: A systematic review”. en. In: *Vacunas (English Edition)* 23. Publisher: Elsevier, S77–S87. DOI: 10.1016/j.vacune.2022.10.019 (cit. on pp. 22, 27, 31).

Neal, Brady, Chin-Wei Huang, and Sunand Raghupathi (2021). *RealCause: Realistic Causal Inference Benchmarking*. en. arXiv:2011.15007 [cs, stat] (cit. on pp. 21, 102).

Nieto-Gómez, P., C. Castaño-Amores, A. Rodríguez-Delgado, and R. Álvarez-Sánchez (2024). “Analysis of oncological drugs authorised in Spain in the last decade: association between clinical benefit and reimbursement”. en. In: *The European Journal of Health Economics* 25.2, pp. 257–267. DOI: 10.1007/s10198-023-01584-9 (cit. on p. 46).

Organization, World Health (2004). *ICD-10 : international statistical classification of diseases and related health problems : tenth revision* (cit. on p. 36).

Parvizi, Javad, Priscilla Ku Cavanaugh, and Claudio Diaz-Ledezma (2013). “Periprosthetic Knee Infection: Ten Strategies That Work”. en. In: *Knee Surgery & Related Research* 25.4, pp. 155–164. DOI: 10.5792/ksrr.2013.25.4.155 (cit. on pp. 50, 61).

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Inglés. Cambridge, U.K. ; New York (cit. on p. 4).

Pearl, Judea (1995). “Causal diagrams for empirical research”. In: *Biometrika* 82.4, pp. 669–688. DOI: 10.1093/biomet/82.4.669 (cit. on pp. 4, 10, 14).

— (2009a). “Causal inference in statistics: An overview”. In: *Statistics Surveys* 3.none, pp. 96–146. DOI: 10.1214/09-SS057 (cit. on p. 3).

— (2009b). *Causality*. Cambridge University Press (cit. on pp. 11, 14).

— (2012). “*Do*-Calculus Revisited”. In: *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. Ed. by Nando de Freitas and Kevin Murphy. Corvallis, OR: AUAI Press, pp. 4–11 (cit. on p. 14).

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell (2016). *Causal Inference in Statistics. A Primer*. John Wiley and Sons Ltd, United States (cit. on pp. 50, 80).

Pepe, Margaret Sullivan and Thomas R. Fleming (2018). “Weighted Kaplan-Meier Statistics: Large Sample and Optimality Considerations”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.2, pp. 341–352. DOI: 10.1111/j.2517-6161.1991.tb01827.x (cit. on p. 14).

Petersen, Irene, Ian Douglas, and Heather Whitaker (2016). “Self controlled case series methods: an alternative to standard epidemiological study designs”. In: *BMJ* 354. DOI: 10.1136/bmj.i4515 (cit. on pp. 13, 27).

pygformula (2024). https://github.com/CausalInference/pygformula (cit. on pp. 41, 47, 95).

Rajsfus, Bia Francis, Ronaldo Mohana-Borges, and Diego Allonso (2023). “Diabetogenic viruses: linking viruses to diabetes mellitus”. In: *Heliyon* 9.4, e15021. DOI: https://doi.org/10.1016/j.heliyon.2023.e15021 (cit. on p. 34).

Ramaiyer, Malini, Malak El Sabeh, Jiafeng Zhu, Amanda Shea, Dorry Segev, Gayane Yenokyan, and Mostafa A. Borahay (2024). “The association of COVID-19 vaccination and menstrual health: A period-tracking app-based cohort study”. In: *Vaccine: X* 19, p. 100501. DOI: 10.1016/j.jvacx.2024.100501 (cit. on pp. 22, 31).

Rand, J.A., R.T. Trousdale, D.M. Ilstrup, and W.S. Harmsen (2003). “Factors affecting the durability of primary total knee prostheses”. In: *The Journal of Bone and Joint Surgery-American Volume* 85.2, pp. 259–265. DOI: 10.2106/00004623-200302000-00012 (cit. on p. 62).

Randelli, P., F.R. Evola, P. Cabitza, L. Polli, M. Denti, and L. Vaienti (2010). “Prophylactic use of antibiotic-loaded bone cement in primary total knee replacement”. In: *Knee Surgery, Sports Traumatology, Arthroscopy* 18.2, pp. 181–186. DOI: 10.1007/s00167-009-0921-y (cit. on pp. 23, 50, 62).

Resende, V.A.C., A.C. Neto, C. Nunes, R. Andrade, J. Espregueira-Mendes, and S. Lopes (2021). “Higher age, female gender, osteoarthritis and blood transfusion protect against periprosthetic joint infection in total hip or knee arthroplasties: a systematic review and meta-analysis”. In: *Knee Surgery, Sports Traumatology, Arthroscopy* 29.1, pp. 8–43. DOI: 10.1007/s00167-018-5231-9 (cit. on pp. 23, 62).

Riesch, Hauke and Clive Potter (2014). “Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions”. en. In: *Public Understanding of Science* 23.1, pp. 107–120. DOI: 10.1177/0963662513497324 (cit. on p. 26).

Robins, James (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. DOI: https://doi.org/10.1016/0270-0255(86)90088-6 (cit. on pp. 10, 40).

Robins, James M., Miguel Ángel Hernán, and Babette Brumback (2000). “Marginal Structural Models and Causal Inference in Epidemiology:” en. In: *Epidemiology* 11.5, pp. 550–560. DOI: 10.1097/00001648-200009000-00011 (cit. on p. 101).

Rosenbaum, Paul R. and Donald B. Rubin (1983). “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70.1, pp. 41–55. DOI: 10.1093/biomet/70.1.41 (cit. on pp. 15, 18, 79).

Rubin, Donald (1972). “ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN EXPERIMENTAL AND OBSERVATIONAL STUDIES”. In: *ETS Research Bulletin Series* 1972.2, pp. i–31. DOI: https://doi.org/10.1002/j.2333-8504.1972.tb00631.x (cit. on pp. 4, 14).

Salmi, Heli, Santtu Heinonen, Johanna Hästbacka, Mitja Lääperi, Paula Rautiainen, Päivi J Miettinen, Olli Vapalahti, Jussi Hepojoki, and Mikael Knip (2022). “New-onset type 1 diabetes in Finnish children during the COVID-19 pandemic”. In: *Archives of Disease in Childhood* 107.2.

Publisher: BMJ Publishing Group Ltd _eprint: https://adc.bmj.com/content/107/2/180.full.pdf, pp. 180–185. DOI: 10.1136/archdischild-2020-321220 (cit. on pp. 22, 34).

Schnabel, Tobias, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims (2016). “Recommendations as Treatments: Debiasing Learning and Evaluation”. en. In: p. 10 (cit. on p. 20).

Schölkopf, B., F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio (2021). “Toward Causal Representation Learning”. In: *Proceedings of the IEEE* 109.5. *equal contribution, pp. 612–634. DOI: 10.1109/JPROC.2021.3058954 (cit. on p. 11).

Schuler, Megan S. and Sherri Rose (2017). “Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies”. In: *American Journal of Epidemiology* 185.1, pp. 65–73. DOI: 10.1093/aje/kww165 (cit. on p. 18).

Schulz, Kenneth F. and David A. Grimes (2002). “Case-control studies: research in reverse”. English. In: *The Lancet* 359.9304. Publisher: Elsevier, pp. 431–434. DOI: 10.1016/S0140-6736(02)07605-5 (cit. on p. 12).

Schwab, Patrick, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen (2020). “Learning Counterfactual Representations for Estimating Individual Dose-Response Curves”. en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04. arXiv:1902.00981 [cs, stat], pp. 5612–5619. DOI: 10.1609/aaai.v34i04.6014 (cit. on p. 20).

Schwab, Patrick, Lorenz Linhardt, and Walter Karlen (2019). *Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks*. en. arXiv:1810.00656 [cs, stat] (cit. on pp. 20, 79, 91, 102).

Shalit, Uri, Fredrik D. Johansson, and David Sontag (2017). *Estimating individual treatment effect: generalization bounds and algorithms* (cit. on pp. 17, 79, 81).

Shi, Claudia, David M. Blei, and Victor Veitch (2019). “Adapting Neural Networks for the Estimation of Treatment Effects”. en. In: *arXiv:1906.02120 [cs, stat]*. arXiv: 1906.02120 (cit. on pp. 18, 79–82, 86, 87, 91, 97).

Smaardijk, Veerle R., Rana Jajou, Agnes Kant, and Florence P. A. M. van Hunsel (2024). “Menstrual disorders following COVID-19 vaccination: a review using a systematic search”. English. In: *Frontiers in Drug Safety and Regulation* 4. Publisher: Frontiers. DOI: 10.3389/fdsfr.2024.1338466 (cit. on pp. 22, 31).

Snow, John (1855). *On the mode of communication of cholera*. eng. London : John Churchill (cit. on p. 9).

Spieth, Peter Markus, Anne Sophie Kubasch, Ana Isabel Penzlin, Ben Min-Woo Illigens, Kristian Barlinn, and Timo Siepmann (2016). “Randomized controlled trials – a matter of design”. In: *Neuropsychiatric Disease and Treatment* 12, pp. 1341–1349. DOI: 10.2147/NDT.S101938 (cit. on p. 26).

Stolberg, Harald O., Geoffrey Norman, and Isabelle Trop (2004). “Randomized Controlled Trials”. In: *American Journal of Roentgenology* 183.6. PMID: 15547188, pp. 1539–1544. DOI: 10.2214/ajr.183.6.01831539 (cit. on p. 12).

Tawfik, Gehad Mohamed, Kadek Agus Surya Dila, Muawia Yousif Fadlelmola Mohamed, Dao Ngoc Hien Tam, Nguyen Dang Kien, Ali Mahmoud Ahmed, and Nguyen Tien Huy (2019). “A step by step guide for conducting a systematic review and meta-analysis with simulation data”. In: *Tropical Medicine and Health* 47.1, p. 46. DOI: 10.1186/s41182-019-0165-6 (cit. on p. 13).

Taylor, Kurt et al. (2024). “Incidence of diabetes after SARS-CoV-2 infection in England and the implications of COVID-19 vaccination: a retrospective cohort study of 16 million people”. en. In: *The Lancet Diabetes & Endocrinology* 12.8, pp. 558–568. DOI: 10.1016/S2213-8587(24)00159-1 (cit. on pp. 22, 34, 35).

Tayton, E.R., C. Frampton, G.J. Hooper, and S.W. Young (2016). “The impact of patient and surgical factors on the rate of infection after primary total knee arthroplasty: an analysis of 64,566 joints from the New Zealand Joint Registry”. In: *The Bone & Joint Journal* 98-B, pp. 334–340. DOI: 10.1302/0301-620X.98B3.36775 (cit. on p. 61).

Troncoso, Daniel Pérez, Borja Velasco Regúlez, Jessica Ruiz Baena, Rosa María Vivanco Hidalgo, Juan José Chillarón Jordán, Elisenda Climent Biescas, Silvia Ballesta Purroy, Gemma Llauradó Cabot, and Carles Forné Izquierdo (2022). “La incidència de diabetis mellitus de tipus 1 durant la pandèmia de COVID-19 a Catalunya”. ca. In: (cit. on p. 34).

Uysal, S. Derya (2015). “Doubly Robust Estimation of Causal Effects with Multivalued Treatments: An Application to the Returns to Schooling: DOUBLY ROBUST ESTIMATION OF CAUSAL EFFECTS”. en. In: *Journal of Applied Econometrics* 30.5, pp. 763–786. DOI: 10.1002/jae.2386 (cit. on p. 20).

Estrategia de vacunación COVID-19 (2024). https://www.vacunacovid.gob.es/ (cit. on pp. 37, 43).

Vansteelandt, Stijn and Marshall Joffe (2014). “Structural Nested Models and G-estimation: The Partially Realized Promise”. In: *Statistical Science* 29.4, pp. 707–731. DOI: 10.1214/14-STS493 (cit. on pp. 16, 101).

Velasco-Regulez, Borja, Jose L Fernandez-Marquez, Nerea Luqui, Jesus Cerquides, Josep Lluis Arcos, Analia Fukelman, and Josep Perelló (2022). “Is the phase of the menstrual cycle relevant when getting the covid-19 vaccine?” In: *American Journal of Obstetrics & Gynecology* (cit. on p. 31).

Vohland, Katrin, Anne Land-Zandstra, Luigi Ceccaroni, Rob Lemmens, Josep Perelló, Marisa Ponti, Roeland Samson, and Katherin Wagenknecht, eds. (2021). *The Science of Citizen Science*. en. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-58278-4 (cit. on p. 27).

Wander, Pandora L., Elliott Lowy, Lauren A. Beste, Luis Tulloch-Palomino, Anna Korpak, Alexander C. Peterson, Steven E. Kahn, and Edward J. Boyko (2022). “The Incidence of Diabetes Among 2,808,106 Veterans With and Without Recent SARS-CoV-2 Infection”. In: *Diabetes Care* 45.4, pp. 782–788. DOI: 10.2337/dc21-1686 (cit. on pp. 22, 34).

Wang, Jixian (2018). “A simple, doubly robust, efficient estimator for survival functions using pseudo observations”. en. In: *Pharmaceutical Statistics* 17.1, pp. 38–48. DOI: 10.1002/pst.1834 (cit. on p. 64).

Wang, Shirley V., Sebastian Schneeweiss, and RCT-DUPLICATE Initiative (2023). “Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials”. In: *JAMA* 329.16, pp. 1376–1385. DOI: 10.1001/jama.2023.4221 (cit. on pp. 11, 47, 103).

Wang, Xiaofeng and Michael W. Kattan (2020). “Cohort Studies: Design, Analysis, and Reporting”. In: *Chest*. An Overview of Study Design and Statistical Considerations 158.1, Supplement, S72–S78. DOI: 10.1016/j.chest.2020.03.014 (cit. on p. 12).

Wen, Lan, Jessica G. Young, James M. Robins, and Miguel A. Hernán (2021). “Parametric g-formula implementations for causal survival analyses”. en. In: *Biometrics* 77.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13321, pp. 740–753. DOI: 10.1111/biom.13321 (cit. on pp. 16, 40).

Wise, Jacqui (2024). “Covid-19: Two rare vaccine side effects detected in large global study”. en. In: *BMJ* 384. Publisher: British Medical Journal Publishing Group Section: News, q488. DOI: 10.1136/bmj.q488 (cit. on pp. 21, 26).

Xie, Yan and Ziyad Al-Aly (2022). “Risks and burdens of incident diabetes in long COVID: a cohort study”. In: *The Lancet Diabetes & Endocrinology* 10.5, pp. 311–321. DOI: https://doi.org/10.1016/S2213-8587(22)00044-4 (cit. on pp. 22, 34).

Xiong, Xi et al. (2023). “Incidence of diabetes following COVID-19 vaccination and SARS-CoV-2 infection in Hong Kong: A population-based cohort study”. en. In: *PLOS Medicine* 20.7. Ed. by Amitabh Bipin Suthar, e1004274. DOI: 10.1371/journal.pmed.1004274 (cit. on pp. 22, 35).

Yoon, Jinsung, James Jordon, and Mihaela Van Der Schaar (2018). “GANITE: Estimation of individualized treatment effects using generative adversarial nets”. In: *International Conference on Learning Representations* (cit. on pp. 17, 79).

Young, Jessica G., Lauren E. Cain, James M. Robins, Eilis J. O’Reilly, and Miguel A. Hernán (2011). “Comparative Effectiveness of Dynamic Treatment Regimes: An Application of the Parametric G-Formula”. en. In: *Statistics in Biosciences* 3.1, pp. 119–143. DOI: 10.1007/s12561-011-9040-7 (cit. on p. 40).

Yuan, Ye, Xueying Ding, and Ziv Bar-Joseph (2020). *Causal inference using deep neural networks*. en. arXiv:2011.12508 [cs, stat] (cit. on pp. 17, 79).

Zare, Ali, Mahmood Mahmoodi, Kazem Mohammad, Hojjat Zeraati, Mostafa Hosseini, and Kourosh Holakouie Naieni (2014). “A Comparison between Kaplan-Meier and Weighted Kaplan-Meier Methods of Five-Year Survival Estimation of Patients with Gastric Cancer”. en. In: (cit. on p. 64).

Zhao, Lili, Huong Tran, Malcolm Risk, and Girish Nair (2024). "Risk of Blood Clots After COVID-19 Vaccination and Infection: A Risk-Benefit Analysis". In: *Research Square*, rs.3.rs–4378029. DOI: 10.21203/rs.3.rs-4378029/v1 (cit. on p. 26).

Zhu, Jie and Blanca Gallego (2022). "Causal inference for observational longitudinal studies using deep survival models". In: *Journal of Biomedical Informatics* 131, p. 104119. DOI: 10.1016/j.jbi.2022.104119 (cit. on p. 16).