Predicting time-to-event outcomes under different intervention strategies: methods and applications

Prosepe, I.

Chapter 1

Introduction

What if questions

In clinical research, "what if" (i.e. causal) questions have long been a focus of interest. A historical example dates back to 1760, when Daniel Bernoulli was interested in estimating mortality rates in a hypothetical world where smallpox was eradicated [1]. To estimate such mortality rates, Bernoulli would have needed to rely on at least two assumptions: (a) that the rate of death from causes other than smallpox would remain unchanged after eradication, and (b) that deaths from smallpox and deaths from other causes were independent [2]. Such assumptions are not realistic. For instance, assumption (a) implies that people's behavior would be identical in the presence or absence of smallpox. To offer a modern parallel, this is similar to assuming that population behavior during the first wave of Covid-19 was the same as in 2019. In reality, the population behavior differed substantially, not just because of Covid-19 deaths, but also due to changes such as reductions in car accidents (due to lockdowns) and delays in accessing medical care (due to overloading of hospitals). Assumption (b) requires that, at the population level, dying from smallpox and dying from other causes are independent events. However, this independence is highly unlikely in practice. For example, lower socio-economic status can indirectly increase both the risk of both dying from smallpox as well as the risk of dying from other causes, creating a dependence between the two.

However, we are no longer in 1760, and today we often have access to high-quality, detailed data. If Bernoulli had access to computers and the detailed data available now, he could have tried to estimate his target mortality rate by dividing the population into many subgroups of individuals who are more comparable to one another (e.g. in terms of age, exposure to smallpox, and socio-economic status). He could have relaxed assumption (b), requiring it to hold only within each subgroup — a far more realistic approach than assuming it holds across the entire population.

In this thesis, we develop and apply estimation methods for "what if" questions and we for-

malize the assumptions required to ensure the validity of these methods.

What if predictions

In clinical practice, healthcare professionals are frequently faced with questions related to the health prospects of their patients. Typical examples include:

Q1. Given this patient's current health status, what is their risk of death in the next 10 years?
Q2. If this patient receives a transplant now, what is their risk of death in the next 10 years?

One way of answering such questions is to develop clinical prediction models, which provide predictions, i.e. estimates of the probability of specific future outcomes for individual patients based on their characteristics or for clearly defined (subsets of) populations. Importantly, question Q1 refers to outcomes under current clinical practice and can therefore be answered with predictions based on observed data. In contrast, Q2 is a "what-if" question — similarly to the question that first captured Bernoulli's interest in 1760 — as it concerns a hypothetical scenario in which a specific intervention (e.g., a transplant) is applied. These two different type of questions yield to two different types of predictions. Questions like Q1 lead to associational predictions, which estimate what is likely to happen to patients based on their current health status, without accounting for the possible effect of interventions. We will refer to these type of predictions as 'factual predictions'. Questions like Q2 lead to predictions that aim to inform about what will happen to patients with specific characteristics if a particular action or intervention is taken. These type of predictions that aim to quantify the consequences of actions are causal in nature [3]. Estimating such predictions has been referred to as 'counterfactual prediction' and 'prediction under hypothetical interventions'. In this introduction, as well as in the title of this dissertation, we use the term 'prediction under interventions'.

A particularly common type of clinical outcomes involves time-to-event outcomes. These outcomes occur when the aim is to estimate the probability of an event occurring over time, measured from a specific starting point (e.g., diagnosis) to an outcome of interest that occurs some time after prediction (e.g., death, disease progression, or recurrence) [4]. A common challenge in analyzing such outcomes is right-censoring [5], which occurs when the exact time of the event is unknown, but it is known to have occurred after a certain point (the censoring time). This typically arises because follow-up may be incomplete — either due to individuals being lost to follow-up or the study ending before the event occurs. Time-to-event outcomes are the subject of survival analysis and they are the focus of this dissertation. One quantity that will appear at several points in this dissertation is *conditional survival benefit*, defined as a difference in survival with versus without treatment over a specified time horizon for a patient with certain characteristics. Estimating conditional survival benefit requires predictions under interventions, as it involves comparing two "what if" scenarios: one in which treatment is given and one in which it is not. This makes survival benefit an inherently causal quantity.

Prediction under interventions differs fundamentally from factual prediction in the outcome probabilities they target. Factual prediction informs on risks under the current standard of

care — that is, based on how patients were actually treated in observed practice. For instance, in the context of hospitalized Covid-19 patients, a factual model might estimate a patient's 28-day survival probability under the current standard care where, for example, some of these patients are admitted to the ICU while others are not. Such predictions reflect outcomes under existing practices and may not generalize to alternative treatment strategies. For example, a patient predicted to have a low risk of death may be classified as low-risk because, under current practice, they are routinely admitted to the ICU, reducing mortality [6–8]. Using these predictions to guide decisions — such as whether to admit the patient to the ICU — can be misleading, as they do not account for how risk would change if specific interventions were applied or withheld. Prediction under hypothetical interventions, by contrast, explicitly considers how outcomes would differ under alternative intervention strategies. These predictions are essential in clinical settings, where practitioners must often choose between alternative intervention strategies and therefore want to be informed of risk under each available option. This difference between factual prediction and prediction under interventions is reflected in the estimation methods: factual prediction models estimate risk based on observed data without explicitly accounting for how outcomes may change under alternative treatment strategies; prediction under interventions account for such changes.

Accounting for changes in treatment strategies comes at a cost. Estimation methods for prediction under interventions are typically more complex and rely on strong assumptions, such as assumptions (a) and (b). When these assumption do not hold, prediction under interventions is not possible as it targets a risk that is not identifiable. Concerns over the validity of causal estimation - given its reliance on such strong assumptions — led researchers for many years to focus only on factual prediction. However, there is growing recognition that many clinically relevant questions are inherently causal in nature [9], and require the type of risk estimates provided by prediction under different intervention strategies.

This renewed emphasis on causal inference has been made possible by several key advances over the past decade. First, the increasing availability of rich observational data, along with advances in computational power and algorithms capable of analyzing large-scale datasets [10], has enabled researchers to partition target populations into more comparable subgroups. As a result, strong assumptions — such as (b) — can now be relaxed to hold only within these subgroups, making them more plausible in practice. Second, the development of a formal causal language has clarified the meaning of risk estimates under interventions and led to the formalization of the identifiability conditions that are generally required — though context-dependent in their exact form — for estimation. These identifiability conditions are consistency, positivity, conditional exchangeability and — specifically for time-to-event outcomes — conditionally independent censoring [11].

Identifiability conditions

Prediction under interventions aims to estimate what would happen to patients with specific characteristics if a particular action was taken. For instance, for an individual patient, such predictions might quantify the probability of death within 28 days both if the patient receives treatment and if they do not. These predictions rely therefore on one fundamental assump-

tion: the existence of well-defined *counterfactual outcomes* — that is, outcomes that would occur under each possible intervention (e.g., treat or do not treat). We provide here a general overview of the identifiability conditions of a specific intervention:

- **Consistency:** The counterfactual outcome under the given intervention is equal to the observed outcome for patients who actually received that intervention.
- **Positivity:** Patients have a non-zero probability of receiving the intervention.
- **Conditional exchangeability:** Among patients with the same characteristics, patients who receive the intervention are representative of those who do not receive it. This assumption is also known as "no unmeasured confounding".
- **Independent censoring:** Among patients with the same characteristics, patients who are not censored are representative of those who are censored.

Throughout this thesis, these assumptions will recur as we will provide their specific rigorous formulation, each time adapting to the particular setting and estimation method under consideration. Careful specification of these assumptions is essential for understanding what the prediction model captures — and where it may fail.

Aims and research questions

The objective of this dissertation is to develop and apply estimation methods for *prediction under interventions for time-to-event outcomes*. Specifically, the focus is on estimating risks for time-to-event outcomes using observational data in a manner that supports valid prediction under different intervention strategies, potentially conditional on patient-specific covariates. To this end, we formulate the identifiability assumptions required to ensure the validity of these risk estimates. We investigate how such predictions can be estimated to support repeated use over time, particularly in clinical settings where treatment decisions must be revisited at multiple time points. Furthermore, we explore the conditions under which these predictive models can be extended to incorporate not only patient characteristics but also attributes of the treatment itself.

Thesis outline

Chapter 2 presents a review of Covid-19 prediction models, with the goal of assessing whether the analysis strategies used in published studies are aligned with the intended purpose of the models. This work is motivated by the paper by van Geloven et al. [6], that introduced a prediction estimand framework. The framework provides formal definitions for formulating prediction questions when treatments are initiated after baseline and specifies the appropriate analysis strategy for each type of question. Using this framework, we evaluate the alignment between stated objectives and analysis strategies in the reviewed studies. Additionally, we use national Dutch data to illustrate how predictions can vary depending on

the analysis strategy used — highlighting the potential for misuse of prediction models when they are not developed in accordance with their intended use. This chapter is based on the following publication:

> **I. Prosepe**, R.H.H. Groenwold, R. Knevel, R. Pajouheshnia, N. van Geloven (2022). The Disconnect Between Development and Intended Use of Clinical Prediction Models for Covid-19: A Systematic Review and Real-World Data Illustration. *Frontiers in Epidemiology*, 2. DOI: 10.3389/fepid.2022.899589.

In **Chapter 3** and **Chapter 4**, we explore the application setting of scarce medical resources — focusing on livers from deceased organ donors. Due to their limited availability, donor livers are allocated to patients on a waiting list based on a prioritization rule that determines who is offered a liver when one becomes available. Defining what constitutes a fair or effective rule is challenging, as it involves balancing multiple considerations and criteria. The Eurotransplant region prioritizes patients based on the Model for End-Stage Liver Disease (MELD) score, calculated from serum creatinine, bilirubin, and the international normalized ratio (INR) [12]. In the United States, a modified version — MELD-Na, which includes serum sodium — guides allocation decisions [13]. These prioritization rules are urgency-based, meaning they aim to prioritize the sickest patients first [14]. There is debate about whether urgency-based allocation — which focuses solely on pre-transplant mortality risk and does not consider post-transplant mortality — is the fairest approach [14]. In response, there has been growing interest in incorporating individual expected survival benefit — defined as the predicted number of life-years gained from transplantation — into allocation decisions.

In **Chapter 3**, we estimate the survival benefit among the treated from observational data in the presence of time-varying confounding, in the context of liver transplantation, using methodology developed by Gong and Schaubel [15]. In particular, we quantify survival benefit for patients who received a transplant, with a focus on comparing outcomes between those with and without hepatocellular carcinoma (HCC) who were listed for liver transplantation on the United Network for Organ Sharing (UNOS) waiting list [16]. This work aims at a clinical audience, particularly liver transplant surgeons and policymakers, and was motivated by potential disparities in transplant access between patients with and without HCC. This chapter is based on the following publication:

> B.F.J. Goudsmit, **I. Prosepe**, M.E. Tushuizen, V. Mazzaferro, I.P.J. Alwayn, B. van Hoek, A.E. Braat, H. Putter (2023). Survival benefit from liver transplantation for patients with and without hepatocellular carcinoma. *JHEP Reports*. 5(12). DOI: 10.1016/j.jhepr.2023.100907.

Chapter 4 presents an extension of the methodology developed by Gong and Schaubel [15, 17] that allows for the estimation of survival benefit for *all* patients on the liver transplant waiting list, not just those who received a transplant. We develop an estimation method that enables repeated prediction of survival benefit over time for all listed patients, with the goal of supporting benefit-based prioritization rules on the transplant waiting list. This work addresses several methodological challenges. First, it requires a clear formulation of the estimands and a careful articulation of the assumptions necessary for identification. Second, the methodology

needs to allow for repeated predictions over time. Third, while estimating benefit among the treated involves predictions under intervention for only one of its two components (i.e. the survival without treatment), estimating survival benefit for all patients requires prediction under interventions for both components (i.e. both the survival with and without treatment). Predictions of survival with treatment are complex in the contexts where treatment itself may vary (e.g. different donor characteristics), because this introduces the issue of multiple versions of treatment [18]. We apply the methodology in the context of liver transplantation but this time used data from the Eurotransplant registry [19]. This chapter is based on the following:

I. Prosepe, N. van Geloven, H. de Ferrante, A.E. Braat, H. Putter (2025). Estimating conditional survival benefit for the allocation of scarce resources.

In **Chapter 5**, we shift the focus to a new clinical setting that introduces new challenges. The aim of this work was to develop and validate a dynamic prediction model for major bleeding or mortality in infants with severe thrombocytopenia (platelet count $< 50 \times 10^9/L$), under two intervention strategies: receiving a platelet transfusion within 6 hours versus no transfusion for 3 days. We are motivated by existing clinical uncertainty around whether prophylactic transfusion effectively reduces bleeding risk or may, in some cases, cause more harm than benefit. The statistical challenges in this setting include the dynamic nature of the predictions, the fact that treatment is repeated over time, and the need to accommodate more complex treatment strategies. To address these challenges, we combine clone–censor–weighting [20] with landmarking [4, 21]. The prediction model is developed using data from fourteen neonatal intensive care units across the Netherlands, Sweden, and Germany, and externally validated on data from seven Dutch centers. Validation is performed using recently developed evaluation methods specifically designed for evaluating predictions under interventions [22]. This chapter is based on the following:

H. van der Staaij, **I. Prosepe**, C. Caram-Deelder, R.H. Keogh, E. Deschmann, C. Dame, W. Onland, S.A. Prins, F. Cassel, E.J. d'Haens, E. van Westering-Kroon, P. Andriessen, S.L. Vrancken, C.V. Hulzebos, D.C. Vijlbrief, S.F. Fustolo-Gunnink, K. Fijnvandraat, E. Lopriore, J.G. van der Bom, N. van Geloven (2025). Individualized Prediction of Platelet Transfusion Outcomes in Preterm Infants With Severe Thrombocytopenia. *JAMA*, published online September 15 2025. DOI: `10.1001/jama.2025.14194`

In **Chapter 6**, we turn our interest to multistate models, which are survival analysis models for settings where individuals may experience multiple different events over time [23–25]. This work is intended for users of multistate models who want to expand the types of questions they can address, to include "what if" questions, framed as interventions that alter specific transition rates. We propose an estimator able to estimate the causal effect of different treatment delay strategies. In particular, we estimate the impact of strategies such as awaiting natural recovery for 3 months, on the marginal probability of recovery. Rather than a comparison between treatment or no treatment, our focus here is on the timing of intervention. We apply the proposed methodology to estimate the effect of treatment delay on a cohort of 1896 couples with unexplained subfertility who seek intrauterine insemination. This chapter is based on the following publication:

I. Prosepe, S. le Cessie, H. Putter, N. van Geloven (2025). Causal multistate models to evaluate treatment delay. *Statistics in Medicine*, 44(7). DOI: 10.1002/sim.70061.

In **Chapter 7**, we address the challenge of keeping clinical prediction models up to date over time. Clinical models can quickly become outdated when new treatments are introduced. While they can be dynamically updated [26–29], this often requires waiting for sufficient new data to accumulate [30, 31] —a process that takes time. To address this, we propose *interventional updates* which integrate external evidence (e.g. from clinical trials) about the effectiveness of new treatments to generate updated predictions under upcoming interventions. We illustrate our methods using electronic health records of hospitalized Covid-19 patients, collected in four Dutch hospitals in 2020 and 2021. This chapter is based on the following:

I. Prosepe, A. Kaal, K. Diaz-Ordaz, K. Tanner, C.M.V. Stark, J.M. Smit, E. Leegwater, C. van Nieuwkoop, M.S. Arbous, N. van Geloven (2025). Interventional updating of time-to-event prediction models.

Finally, **Chapter 8** summarizes the main conclusions and challenges discussed throughout the previous chapters and outlines potential directions for future research.

References

1. Bernoulli D. Essai d'une Nouvelle Analyse de La Mortalité Causée Par La Petite Vérole, et Des Avantages de l'inoculation Pour La Prevenir. 1760;Mémoires de l'Académie Royal des Sciences:1–45.
2. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The Analysis of Failure Times in the Presence of Competing Risks. *Biometrics.* 1978;34(4):541.
3. Gillies D. An Action-Related Theory of Causality. *The British Journal for the Philosophy of Science.* 2005;56(4):823–842.
4. van Houwelingen H, Putter H. *Dynamic Prediction in Clinical Survival Analysis*. Boca Raton: CRC Press 2011.
5. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and HealthNew York Berlin Heidelberg: Springersecond edition, corrected third printing ed. 2005.
6. van Geloven N, Swanson SA, Ramspek CL, *et al.* Prediction Meets Causal Inference: The Role of Treatment in Clinical Prediction Models. *European Journal of Epidemiology.* 2020;35(7):619–630.
7. Sperrin M, McMillan B. Prediction Models for Covid-19 Outcomes. *BMJ.* 2020:m3777.
8. Prosepe I, Groenwold RHH, Knevel R, Pajouheshnia R, van Geloven N. The Disconnect Between Development and Intended Use of Clinical Prediction Models for Covid-19: A Systematic Review and Real-World Data Illustration. *Frontiers in Epidemiology.* 2022;2.
9. Hernán MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *American Journal of Public Health.* 2018;108(5):616–619.
10. Shi J, Norgeot B. Learning Causal Effects From Observational Data in Healthcare: A Review and Summary. *Frontiers in Medicine.* 2022;9.
11. Hernán M, Robins JM. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied ProbabilityBoca Raton: Chapman & Hall/CRC 2023.
12. Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, ter Borg PC. A Model to Predict Poor Survival in Patients Undergoing Transjugular Intrahepatic Portosystemic Shunts. *Hepatology.* 2000;31(4):864–871.
13. Kim WR, Biggins SW, Kremers WK, *et al.* Hyponatremia and Mortality among Patients on the Liver-Transplant Waiting List. *New England Journal of Medicine.* 2008;359(10):1018–1026.
14. Tschuor C, Ferrarese A, Kuemmerli C, *et al.* Allocation of Liver Grafts Worldwide – Is There a Best System?. *Journal of Hepatology.* 2019;71(4):707–718.
15. Gong Q, Schaubel DE. Estimating the Average Treatment Effect on Survival Based on Observational Data and Using Partly Conditional Modeling: Treatment Effect on Survival via Partly Conditional Regression. *Biometrics.* 2017;73(1):134–144.
16. OPTN/UNOS liver and intestinal transplantation committee . OPTN/UNOS Policy Notice Revisions to National Liver Review Board Policies. https://optn.transplant.hrsa.gov/media/2816/liver_nlrb-revised-policynotice-dsa_01252019.pdf.
17. Gong Q, Schaubel DE. Partly Conditional Estimation of the Effect of a Time-Dependent Factor in the Presence of Dependent Censoring. *Biometrics.* 2013;69(2):338–347.
18. VanderWeele TJ, Hernán MA. Causal Inference Under Multiple Versions of Treatment. *Journal of causal inference.* 2013;1(1):1–20.
19. Eurotransplant Registry. https://www.eurotransplant.org/about-eurotransplant/registry/.
20. Hernán MA. How to Estimate the Effect of Treatment Duration on Survival Outcomes Using Observational Data. *BMJ (Clinical research ed.).* 2018;360:k182.
21. van Houwelingen HC. Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics.* 2007;34(1):70–85.
22. Keogh RH, van Geloven N. Prediction Under Interventions: Evaluation of Counterfactual Performance Using Longitudinal Observational Data. *Epidemiology.* 2024;35(3):329.
23. Putter H, Fiocco M, Geskus RB. Tutorial in Biostatistics: Competing Risks and Multi-State Models. *Statistics in Medicine.* 2007;26(11):2389–2430.
24. Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen PK. Multi-State Models for the Analysis of Time-to-Event Data. *Statistical Methods in Medical Research.* 2009;18(2):195–222.
25. Andersen PK, Pohar Perme M. Inference for Outcome Probabilities in Multi-State Models. *Lifetime Data Analysis.* 2008;14(4):405–431.
26. Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic Models to Predict Health Outcomes: Current Status and Methodological Challenges. *Diagnostic and Prognostic Research.* 2018;2(1):23.

27. Schnellinger EM, Yang W, Kimmel SE. Comparison of Dynamic Updating Strategies for Clinical Prediction Models. *Diagnostic and Prognostic Research.* 2021;5(1):20.
28. Binuya MaE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological Guidance for the Evaluation and Updating of Clinical Prediction Models: A Systematic Review. *BMC medical research methodology.* 2022;22(1):316.
29. Tanner KT, Keogh RH, Coupland CAC, Hippisley-Cox J, Diaz-Ordaz K. Dynamic Updating of Clinical Survival Prediction Models in a Changing Environment. *Diagnostic and Prognostic Research.* 2023;7(1):24.
30. Van Calster B, Van Hoorde K, Vergouwe Y, *et al.* Validation and Updating of Risk Models Based on Multinomial Logistic Regression. *Diagnostic and Prognostic Research.* 2017;1(1):2.
31. Van Calster B, McLernon DJ, van Smeden M, *et al.* Calibration: The Achilles Heel of Predictive Analytics. *BMC Medicine.* 2019;17(1):230.