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

Prosepe, I.

Chapter 8

Discussion

Clinical prediction models are a central part of medical research and can play an important role in medical practice. In this dissertation, we developed and applied estimation methods that advance the field of prediction under interventions for time-to-event outcomes.

Chapter 2 served as the motivation for the rest of the thesis, providing context for the estimation methods developed in the subsequent chapters. In our review of Covid-19 prediction models, we assessed whether the analysis strategies used in published studies were aligned with the intended purpose of the models. We found that, very often, there was a mismatch between the two. One important finding was that 64% of the reviewed papers recommended their model for decision-making purposes, which typically requires prediction under hypothetical interventions. However, these models were developed using methods that do not support that aim. Additionally, 21% of the prediction models did not have a clearly defined prediction estimand, making it unclear what their estimated risks actually represent. These findings highlight that it remains common practice to use estimation strategies that are not appropriate for the specific prediction estimand being targeted. One contributing factor is that appropriate estimation strategies for certain clinical settings are not well-established in literature, which motivated the methodological developments presented in this thesis.

In **Chapter 3** and **Chapter 4**, we explored the application setting of scarce medical resources — focusing on livers from deceased organ donors. In **Chapter 3**, we studied survival benefit in the context of liver transplantation for patients listed on the United Network for Organ Sharing (UNOS) waiting list [1]. This clinically focused work aimed to examine differences in access to transplantation between patients with and without hepatocellular carcinoma (HCC). The motivation stemmed from the known limitations of the MELD(-Na) score, which suboptimally captures mortality risk for HCC patients [2], whose outcomes are often driven by tumor progression rather than liver failure. To address this, an exception point system was introduced, granting HCC patients additional points [3–5], which however raised concerns over unintentionally adding inequities [6–8]. We investigated access to liver transplantation among patients with and without hepatocellular carcinoma (HCC), quantifying it in terms of survival benefit among those who actually received a liver transplant. We applied the methodology developed by Gong and Schaubel [9], which extends marginal structural models to handle the complexity of multiple baseline times and multiple time scales. The use of a marginal structural model was necessary to adjust for the time-varying confounding present in the data via inverse probability weighting. Multiple baseline times were needed to make predictions repeatedly over time. Two distinct time scales were relevant: calendar time, as treatment

decisions are based on the availability of donor organs, and patient's follow-up time, as survival predictions are made on this scale. Our findings indicated that patients without HCC experienced the greatest gain in life-years from transplantation, likely because they tended to receive transplants at higher MELD(-Na) scores — typically above 26 — where the survival benefit was most pronounced. In contrast, patients with HCC were more often transplanted at lower MELD(-Na) scores, frequently below 14, where the associated survival benefit was comparatively modest.

The results presented in **Chapter 3** highlighted potential disparities in liver transplant access that could potentially be addressed by shifting to a benefit-based prioritization rule — if the objective would be to maximize overall survival benefit. The methodology developed by Gong and Schaubel was designed to estimate survival benefit only among treated individuals. In their framework, benefit was defined as the difference between two quantities: a prediction of post-transplant survival (estimated among those who received a transplant) and a prediction under the intervention of no transplant (estimated using appropriate adjustment methods among those who did not receive a transplant). However, to inform a prioritization rule based on expected benefit, it is necessary to estimate survival benefit for *all* patients on the waiting list, regardless of whether they ultimately received a transplant. This requires extending the original methodology so that we can use prediction under interventions for *both* components of survival benefit — post-transplant survival and no-transplant survival. This motivated the work presented in **Chapter 4**.

Chapter 4 focused on developing an estimation strategy to dynamically estimate the conditional expected survival benefit for all patients on the transplant waiting list. We built on the work of Gong and Schaubel [9, 10], combining marginal structural models with the use of multiple baselines and multiple time scales. We proposed a different reweighting strategy, as we adapted it to target a broader population — the entire waiting list — rather than restricting estimation to only those who received a transplant. This methodological extension introduced several additional complexities. First, it required a careful formulation of the estimands and of the assumptions necessary for identification. Second, we needed to account for multiple versions of treatment [11], a challenge that was not present in the setting considered by Gong and Schaubel. In predicting outcomes under the intervention of not receiving a liver, there is only one version of non-treatment. However, in predicting outcomes under the intervention of receiving a liver, there are multiple treatment versions, as donor livers vary in quality depending on donor characteristics. With a simulation study, we showed that the proposed method outperforms simpler methods which either do not account for (time-dependent) confounding or ignore the need to combine two different time scales.

We demonstrated the practical value of our methodology by estimating and predicting the survival benefit of liver transplantation for patients with End-Stage Liver Disease in the Eurotransplant region. Using individual-level data from the Eurotransplant registry [12], we applied our model to patients with chronic liver cirrhosis. Our predictions enabled a comparison between the current MELD-based allocation system and a hypothetical benefit-based prioritization rule. A particularly striking finding was that a benefit-based allocation system could, on average, yield an additional 0.46 life-years per transplanted patient over a three-year horizon — compared to the current MELD-based approach — assuming the existing structure of the waiting list remains unchanged. This suggests there could be substantial potential for

improving outcomes through benefit-based allocation. However, because our analysis does not account for how the waiting list might change under a new prioritization system, the findings should be interpreted with caution.

Our results relied on the identifiability conditions of consistency, positivity, conditional exchangeability, and conditionally independent censoring — as well as the correct specification of the outcome and weight models. These assumptions needed to hold for both treatment strategies under investigation: (1) immediately assigning a specific liver and (2) never assigning any liver.

In the data application in **Chapter 4**, consistency appeared reasonable, as both interventions were well-defined. Conditional exchangeability was less straightforward, as it is common in observational data. This assumption is especially challenging for treatment strategies that involve ongoing decision-making, such as the "never assign a liver" strategy. In such strategies, a decision to continue withholding treatment must be made every time a liver becomes available, which means exchangeability must hold at all of these time-points — that is, all confounders influencing each decision must be fully observed at all of these time-points. By contrast, a "treat now" strategy typically involves a single treatment decision, which makes the exchangeability requirement less stringent. However, in our proposed estimation method, where predictions were made at multiple baseline times across different calendar dates, conditional exchangeability was required to hold at each of those calendar dates.

While conditional exchangeability is unlikely to fully hold in geneal, small enough violations — i.e. a small enough degree of unmeasured confounding — can still yield reasonably reliable estimates. In the Eurotransplant data used in **Chapter 4**, we considered conditional exchangeability to hold reasonably well, though not perfectly. Eurotransplant allocates livers based on a prioritization rule, primarily MELD score, which we were able to adjust for as it was available in the data. However, clinicians may decide to decline liver offers based on additional clinical information not captured in the Eurotransplant registry. For instance, a clinician might decline a low-quality liver for a relatively stable patient in the hope of receiving a better organ later. Some of these clinical considerations were reflected in observed covariates, but not all. One notable limitation was the absence of serum sodium in the Eurotransplant data, despite its importance as a predictor of mortality and its likely availability to some clinicians at the time of liver offer. By comparison, the US dataset used in **Chapter 3** was more detailed and included a broader range of clinical variables, making the conditional exchangeability assumption more plausible in that setting.

Positivity was also nontrivial in **Chapter 4**. Theoretically, positivity could be reasonably defended for both treatment strategies based on clinical context. Patients on the transplant waiting list were, by definition, considered eligible to receive treatment, implying a nonzero probability of receiving a transplant. This probability was also never equal to one, due to the scarcity of donor organs. The prioritization system in place does not account for donor liver quality, suggesting that any patient theoretically has some chance of receiving any treatment type. Moreover, variability in donor availability introduces further randomness: organs may be more or less readily available at a particular transplant center, and logistical constraints may lead to reallocating organs to a geographically closer, lower-priority patients. In practice, however, positivity was more problematic. For instance, patients with relatively stable condi-

tions are extremely unlikely to be transplanted with a low quality liver, as their clinician would typically decline the offer. Moreover, at each calendar date, even after discretizing time, only a small number of patients received treatment. Even if they all received the same treatment type, such a small subset could not reliably represent the broader waiting list population at that calendar date, even after reweighting. Certain patient profiles may simply be unrepresented among treated individuals. To address this, we assumed that patient cross-sections at different calendar dates were comparable and pooled these cross-sections for estimation. Additionally, we imposed parametric assumptions on the effects of the covariates describing liver quality, allowing for borrowing of information, particularly in the case of continuous variables. This allowed us to relax our concerns over positivity but introduced concerns over correct model specification, both for the weights model and the outcome model.

While the estimation method we developed allows for the prediction of conditional survival benefit, there are different way to translate such predictions into an actionable prioritization rule in the setting of scarce medical resources. Should the liver simply be allocated to the patient with the highest predicted benefit at the time of availability? Or should the goal be to maximize survival benefit across the entire population over time? These approaches are not equivalent: allocating a liver to one patient now inherently precludes offering it to another who might not receive a second chance. Thus, maximizing survival benefit for an individual does not necessarily translate to maximizing overall survival across the population. Allocating livers in such a way that we maximize survival benefit across the entire population over time is likely more desirable, but pursuing that aim introduces additional challenges, making our current work a necessary but only preliminary step. To move toward a system that truly maximizes population-level survival, future work should focus on incorporating two key components: (i) reliable forecasts of the availability and types of donor livers over both short and long time horizons, which could potentially be achieved through predictions of organ supplies, assuming no major shifts in availability trends; (ii) the integration of these forecasts with the survival benefit predictions. This combination could inform dynamic prioritization strategies that optimize outcomes at the population level. Moreover, to assess how much additional survival benefit such an optimized system could deliver, we would also need to understand how the waiting list would evolve under the new prioritization rule. Future work on this would be essential for evaluating the real-world impact of a benefit-based system. One possible approach would be through large-scale simulation studies — such as those modeled after the Liver Simulated Allocation Model (LSAM) [13] — which can help assess the potential impact of different prioritization strategies in a realistic system.

Moving away from the context of scarce resources, contrasting expected survival with and without treatment can offer direct guidance for clinical decision-making. This is the setting considered in **Chapter 5**, where we developed and validated a decision-support algorithm for prophylactic platelet transfusion in preterm infants. The aim was to support clinicians in evaluating the question: "What is the 3-day risk of major bleeding or death for a preterm infant given their current characteristics if I administer a prophylactic platelet transfusion, and what is the risk if I do not?". In particular we focused on two intervention strategies: receiving a platelet transfusion within 6 hours and no transfusion for 3 days. This study was motivated by the fact that in neonatal intensive care units, the majority of platelet transfusions are given to non-bleeding infants with severe thrombocytopenia with the goal to prevent bleeding. How-

ever, it is uncertain if transfusion is beneficial for all infants with severe thrombocytopenia [14–19], and infants with similar platelet counts may have different bleeding risks due to different clinical conditions.

For model development, we combined again reweighting with landmarking. Even though platelet transfusions can vary in volume, duration, and infusion rate, we did not explicitly model these variations. Instead, we targeted the intervention "receiving a platelet transfusion within 6 hours", interpreted as receiving an average transfusion as typically administered in current clinical practice [11]. Even with this simplification, the modeling process surfaced practical and methodological challenges that we had not anticipated in our initial analysis plan. For instance, we assumed initially that the intervention "receiving a platelet transfusion within 6 hours" could be interpreted as receiving an average transfusion, as typically administered in current clinical practice. However, this interpretation turned out to be too simplistic. In the clinical setting we studied, transfusions could occur repeatedly—administering one now did not preclude more in the near future. While our model was designed to account for prior transfusions (by adjusting for the number of previous transfusions as a confounder and including it as a predictor), it did not explicitly account for future transfusions. As a result, the model predicted the outcome under the intervention of "receiving a platelet transfusion within 6 hours, followed by subsequent transfusions as typically occur in current practice".

A second complexity arose from the combination of landmarking and grace windows. Landmarking was used to accommodate the need for repeated predictions over time, while grace windows were introduced to align with the intervention strategy "initiate transfusion within six hours". An immediate "treat now" strategy was deemed unrealistic, as there is often a delay before a transfusion can be administered from the decision moment. Model development revealed that this combination added complexity to the interpretation of the target estimand. Specifically, the estimand corresponded to: "receiving a platelet transfusion within six hours, followed by subsequent transfusions as they typically occur in current clinical practice, under the assumption that in the first six hours, patients are transfused with timing as observed in the landmarked dataset". This assumption was not fully realistic given our data. In the landmarked dataset, transfusions often occurred at time zero—immediately upon patients meeting eligibility criteria. However, such immediate administration is uncommon in practice, where delays are frequent. One way to address this mismatch would be to estimate the typical delay between the decision to transfuse and the actual administration, and adjust the estimation strategy accordingly. Unfortunately, our data lacked detailed information on this timing, making it impossible to implement such an estimation strategy. As a result, we proceeded with an estimation strategy that does not account for this mismatch, only keeping this discrepancy in mind. While there is existing work on the identification and estimation of treatment strategies that include grace periods [20], the combination of grace periods and landmarking is a promising topic for future work.

Considerations regarding identifiability assumptions were similar to those discussed for **Chapter 4**, particularly for conditional exchangeability. The main simplification, compared to **Chapter 4**, was the absence of the calendar time scale, meaning that conditional exchangeability only needed to hold across follow-up time points rather than also across a set of calendar dates. As in **Chapter 4**, we had access to detailed clinical data, which supported the plausibility of conditional exchangeability holding reasonably well—though some minor unmeasured

confounding likely remained. Positivity assumptions in **Chapter 5** also paralleled those in **Chapter 4**, with the added simplification that we did not have to explicitly account for multiple versions of treatment. Nevertheless, a practical challenge remained: at each landmarking time point, only a small number of infants received transfusions. For this reason, we made similar assumptions as in **Chapter 4** — namely, that the populations at different landmarking times were comparable. We assumed a shared baseline hazard across these populations and introduced parametric assumptions to account for the effect of landmarking time. Under these added assumption, positivity seemed to hold reasonably well. Consistency also held reasonably well under the specific estimands discussed in the last two paragraphs: receiving a platelet transfusion within six hours (receiving a platelet transfusion of the type that typically occurs in current practice within 6 hours, followed by subsequent transfusions as typically occur in current practice), and not receiving a transfusion for three days.

In **Chapter 6**, we broadened the scope of interventions strategies. Until this point, our analyses had primarily focused on binary decisions of the form: "Should treatment be initiated now—yes or no?" However, in clinical practice, the question may be more nuanced. Rather than deciding whether to start treatment immediately, clinicians may ask: "When is the optimal time to initiate treatment, if at all?" This motivated a different type of prediction under intervention — namely, estimating outcomes under interventions that delay treatment initiation for a specified period.

It is important to emphasize that reweighting, on which we had so far relied, was not the best choice for evaluating the effect of treatment timing. To see this, consider the ideal randomized trial we would conduct to answer such a question. At baseline, individuals would be randomized to either initiate treatment immediately or not. After a short time interval, Δt, those who remain event-free in the "no treatment yet" group would be re-randomized to either initiate treatment or continue to delay. This process would repeat at each subsequent Δt until a pre-specified time horizon is reached. As $\Delta t \to 0$, this corresponds to infinite treatment initiation strategies — effectively an infinite number of treatment arms — which would require an infinitely large study population. In real-world observational data, even after discretizing time, we typically observe only a very small number of individuals eligible for treatment at each time point. Consequently, using inverse probability weighting would lead to these few individuals being assigned very large weights to represent the entire risk set, potentially leading to extreme variability in estimates. For this reason, a reweighting-based approach was not well-suited to this setting.

We proposed an approach that combines an illness-death model—a specific type of multistate model—with g-computation to estimate the causal effect of treatment delay using observational data subject to baseline confounding. We formally stated the causal assumptions required for identification and the modeling assumptions needed for estimation. Through a simulation study, we demonstrated that the proposed method makes more efficient use of the data compared to the cloning–censoring–reweighting approach. We applied the proposed methodology to estimate the effect of treatment delay on a cohort of 1896 couples with unexplained subfertility seeking intrauterine insemination.

A limitation of the estimation method we proposed in **Chapter 6** is that it does not account for time-varying confounding, which can arise when treatment decisions are influenced by

a patient's evolving health status. Extending the methodology to incorporate time-varying confounders could be a valuable direction for future research. A second limitation is that the method does not handle competing events, which are relevant when patients may become ineligible for treatment over time. For example, a patient initially planned for treatment might later develop conditions preventing its administration. Addressing this would require redefining the target estimand and adapting the identifiability conditions, both of which present important and interesting opportunities for further methodological development.

Finally, in **Chapter 7**, we addressed the challenge of keeping clinical prediction models up to date as clinical practices evolve and predictions become outdated. For instance, outcomes following liver transplantation may improve as surgical techniques advance; platelet transfusions may be administered differently as new evidence emerges on optimal volume, duration, or infusion rate; and hospital mortality rates may shift with the introduction of novel therapies. While dynamic model updating methods exist [21–24], they typically rely on the accumulation of sufficient new data [25, 26], which can be slow. To address this, we proposed an interventional updating approach that incorporates external evidence — such as findings from clinical trials — on treatment efficacy to enable more timely updates of predictions under changing intervention strategies. We applied our methods using electronic health records from 3236 patients hospitalized with Covid-19 in four Dutch hospitals in 2020 and 2021. We trained an initial prediction model to estimate the 28-day risk of mortality from the time of hospital admission, conditional on a set of clinically relevant covariates. The model was developed using data collected between March and July 2020. We applied dynamic model updating between August 2020 and May 2021, using two approaches: standard updates and interventional updates. We compared the performance of both approaches using metrics of discrimination, calibration, and overall accuracy. In our case study, interventional updating did not improve discrimination over standard updating, but achieved better calibration.

We emphasize that although the work in **Chapter 7** focuses on updating factual prediction models, the same principles apply to models used for predictions under different intervention strategies. Even a model designed to inform decisions — such as whether a patient should receive a liver transplant — makes predictions based on the assumption that the surrounding clinical context does not change over time. This includes implicit assumptions that surgical techniques do not improve over time, or that no new treatments emerge which might reduce the urgency of transplantation. However, clinical care evolves. Thus, having a model capable of prediction under interventions does not imply it accounts for all future changes in practice. Model updating remains essential, and our proposed interventional updating framework offers a way to keep such predictions current, particularly in rapidly changing clinical environments.

In conclusion, this dissertation developed and assessed statistical methodology for prediction under intervention of time-to-event outcomes. We introduced novel methods, grounded in causal inference theory, and applied them across a range of clinical settings, including liver transplantation waiting lists, preterm infants in need of prophylactic platelet transfusion, and risk stratification of hospitalized patients with Covid-19. These applications illustrated both the promise and limitations of prediction under intervention, shedding light on the methodological complexities involved and emphasizing the practical value of such approaches in supporting clinical decision-making.

References

1. 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.
2. Vitale A, Cucchetti A, Qiao G, *et al.* Is Resectable Hepatocellular Carcinoma a Contraindication to Liver Transplantation? A Novel Decision Model Based on "Number of Patients Needed to Transplant" as Measure of Transplant Benefit. *Journal of Hepatology.* 2014;60(6):1165–1171.
3. Alver SK, Lorenz DJ, Marvin MR, Brock GN. Projected Outcomes of 6-month Delay in Exception Points versus an Equivalent Model for End-Stage Liver Disease Score for Hepatocellular Carcinoma Liver Transplant Candidates. *Liver Transplantation.* 2016;22(10):1343–1355.
4. Goldberg D, Mantero A, Newcomb C, *et al.* Predicting Survival after Liver Transplantation in Patients with Hepatocellular Carcinoma Using the LiTES-HCC Score. *Journal of Hepatology.* 2021;74(6):1398–1406.
5. Freeman RB, Gish RG, Harper A, *et al.* Model for End-Stage Liver Disease (MELD) Exception Guidelines: Results and Recommendations from the MELD Exception Study Group and Conference (MESSAGE) for the Approval of Patients Who Need Liver Transplantation with Diseases Not Considered by the Standard MELD Formula. *Liver Transplantation.* 2006;12(Supplement 3):S128-S136.
6. Northup PG, Intagliata NM, Shah NL, Pelletier SJ, Berg CL, Argo CK. Excess Mortality on the Liver Transplant Waiting List: Unintended Policy Consequences and Model for End-Stage Liver Disease (MELD) Inflation. *Hepatology.* 2015;61(1):285–291.
7. Berry K, Ioannou GN. Comparison of Liver Transplant–Related Survival Benefit in Patients With Versus Without Hepatocellular Carcinoma in the United States. *Gastroenterology.* 2015;149(3):669–680.
8. Washburn K, Edwards E, Harper A, Freeman R. Hepatocellular Carcinoma Patients Are Advantaged in the Current Liver Transplant Allocation System. *American Journal of Transplantation.* 2010;10(7):1652–1657.
9. 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.
10. 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.
11. VanderWeele TJ, Hernán MA. Causal Inference Under Multiple Versions of Treatment. *Journal of Causal Inference.* 2013;1(1):1–20.
12. Eurotransplant Registry. https://www.eurotransplant.org/about-eurotransplant/registry/.
13. SRTR . Liver Simulated Allocation Model.
14. Stanworth SJ, Clarke P, Watts T, *et al.* Prospective, Observational Study of Outcomes in Neonates With Severe Thrombocytopenia. *Pediatrics.* 2009;124(5):e826-e834.
15. Muthukumar P, Venkatesh V, Curley A, *et al.* Severe Thrombocytopenia and Patterns of Bleeding in Neonates: Results from a Prospective Observational Study and Implications for Use of Platelet Transfusions. *Transfusion Medicine.* 2012;22(5):338–343.
16. Baer VL, Lambert DK, Henry E, Christensen RD. Severe Thrombocytopenia in the NICU. *Pediatrics.* 2009;124(6):e1095-e1100.
17. Deschmann E, Saxonhouse MA, Feldman HA, Norman M, Barbian M, Sola-Visner M. Association of Bleeding Scores and Platelet Transfusions With Platelet Counts and Closure Times in Response to Adenosine Diphosphate (CT-ADPs) Among Preterm Neonates With Thrombocytopenia. *JAMA Network Open.* 2020;3(4):e203394.
18. Resch E, Hinkas O, Urlesberger B, Resch B. Neonatal Thrombocytopenia—Causes and Outcomes Following Platelet Transfusions. *European Journal of Pediatrics.* 2018;177(7):1045–1052.
19. van der Staaij H, Hooiveld NMA, Caram-Deelder C, *et al.* Most Major Bleeds in Preterm Infants Occur in the Absence of Severe Thrombocytopenia: An Observational Cohort Study. *Archives of Disease in Childhood - Fetal and Neonatal Edition.* 2025;110(2):122–127.
20. Wanis KN, Sarvet AL, Wen L, *et al.* Grace Periods in Comparative Effectiveness Studies of Sustained Treatments. *Journal of the Royal Statistical Society Series A: Statistics in Society.* 2024;187(3):796–810.
21. 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.
22. Schnellinger EM, Yang W, Kimmel SE. Comparison of Dynamic Updating Strategies for Clinical Prediction Models. *Diagnostic and Prognostic Research.* 2021;5(1):20.

23. 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.
24. 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.
25. 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.
26. Van Calster B, McLernon DJ, van Smeden M, *et al.* Calibration: The Achilles Heel of Predictive Analytics. *BMC Medicine.* 2019;17(1):230.