# UNIVERSITÀ DEGLI STUDI DI MILANO

Corso di Dottorato in Scienze per la Sanità Pubblica

Dipartimento di Scienze Cliniche e di Comunità

*DISCCO*

**Evaluation, development, and extension of flexible parametric models for the distributional analysis of censored and non-censored data: A framework for modeling and communicating the full distribution of health outcomes.**

S.S.D MED/01 SECS-S/02

**Giacomo Biganzoli**
R13736
ORCID n. 0000-0002-5228-8443

**TUTOR:** prof. Patrizia Boracchi

**CO-TUTOR:** prof. Federico Ambrogi

**COORDINATORE DEL DOTTORATO:** prof. Fabio Parazzini

**A.A. 2024-2025**

# Table of contents

# 1 Acknowledgments

Although for a PhD student they will never seem enough, three years of Academia are a lot. Research projects aside, they are a lot in terms of emotions: loneliness, happiness, anger, feeling incomplete, feeling at the top of the world, feeling not to deserve the position you have.

For someone used to finding patterns in data, figuring out my own growth pattern has been more difficult than expected. While in this thesis I highlight the need to embrace complexity in medical data, I caught myself thinking of my own life as a simple mountain to climb uphill—no switchbacks, downhills, or breaks. (Sorry for the cycling terms, I pedaled a lot during these times). This struggle taught me a crucial lesson: growing up as a researcher is not a predictable 'linear pattern', but something far more complex. The result is that I can't 'predict' where I'll be in the future, but I know I have gained invaluable professional and life experience. And I know exactly who to thank for helping me reach the final year of my PhD, alive.

First off, I want to thank Professor Patrizia Boracchi for her patience and guidance. Not only did she spend a lot of time teaching me how to be a good biostatistician and researcher, but above all she let me express and work on my ideas, encouraging me to pursue them and believing in them, although they might seem bizarre at first. I came to understand that this is one of the rarest qualities a Supervisor can have, and I am sure I'll pay it forward in my future career.

I also want to thank Dr. Giuseppe Marano for his constant assistance, for the insightful discussions and critical feedback, and for teaching me everything I know about Montecarlo simulations and R programming. I should also thank him for the wonderful movie suggestions that contributed a lot to my leisure time.

I would also like to thank Professor Joel Schwartz, who hosted me in his lab during my period as a visiting researcher at Harvard T.H. Chan School of Public Health in Boston. His lectures and insights were very important for my understanding of the issues in modeling epidemiological data and inspired a big portion of the work presented in this thesis.

Finally, I would also like to thank Professor Laura Antolini and Professor Clelia di Serio for dedicating some of their time to reviewing this thesis and giving me valuable insights and suggestions.

Beyond Academia, I had the fortune of being surrounded by wonderful people that I must acknowledge.

I would like to thank my Family for the love and support they provided. With Family I mean my beloved relatives: my Mamma and Papà, Caterina and Elia and brother Davide; my Nonni Marialuisa and Antonio; my Zii, Betta and Fabio and my cousins Sofia and Bea; but also my best Friends: Marco, Giorgia, Margherita, for the nights, the laughs and the getaways together, and Lorenzo (Il Falco) and Nicola (Il Picchio) for the numerous bike rides around the Alps.

Their support was especially vital as I discovered that living 6000 km from home is not that easy, but neither is it that difficult—especially when you find a house that feels like Home. I want to thank Andrea and Matt for having me in their wonderful place in Brookline, for showing me around New England and letting me live the real American experience in a warm,

welcoming atmosphere. I am also grateful to Mike, my spiritual guru, for the numerous chats we had in the kitchen about life, love, and the future. That home was made even better by Andrei and Farzaan, for their good cheer and for being the best housemates in the world. My American community also grew through my passions, and I want to thank my cycling buddy Ricky who, as soon as he met me, provided me with one of the most important things in life: a bike. Many thanks also to his friends, now also my friends, Charlotte, Tyler, and Max, for the miles spent riding together and the way-too-big ice creams we had after.

Saving the most important for last, I want to thank my girlfriend Benedetta, who cared about me and loved me every second we spent far from each other, for getting along with a mix of grumpiness and silliness broadcasted through FaceTime, and for waiting for me until late, not caring about sleep cycles. She inspires me a lot, and I wish I had the same passion and dedication for my research as she has for her archaeological studies.

Having said that, among all the things I learned, studied, and reasoned on during these years, one is now perfectly clear to me.

Time does not exist; it is just one of our several mental constructions. Love exists indeed, and it is everywhere

*Per quanto a un dottorando possano non bastare mai, tre anni di vita accademica sono un'enormità. Progetti di ricerca a parte, lo sono soprattutto sul piano delle emozioni: si passa dalla solitudine più profonda a momenti di pura felicità, dalla rabbia alla sensazione di essere incompleti, dal sentirsi in vetta al mondo al credere di non meritare il posto che si occupa.*

*Per uno come me, abituato a cercare pattern nei dati, decifrare il mio percorso di crescita è stato più difficile del previsto. Mentre in questa tesi spiego la necessità di accogliere la complessità dei dati medici, mi sono ritrovato a pensare alla mia vita come a una montagna da scalare tutta d'un fiato, senza tornanti, discese o soste. (Scusate i termini ciclistici, ma di chilometri in sella ne ho macinati parecchi in questi anni). Questa fatica, però, mi ha insegnato una lezione fondamentale: crescere come ricercatore non è un "percorso lineare", ma un cammino ben più complesso. Di conseguenza, non so "prevedere" cosa farò in futuro, ma so di aver messo in valigia un bagaglio inestimabile di esperienze, professionali e umane. E so con certezza chi devo ringraziare se sono arrivato alla fine di questo dottorato, e per di più vivo.*

*Il mio primo ringraziamento va alla Professoressa Patrizia Boracchi, per la sua pazienza e la sua guida sicura. Non solo ha speso innumerevoli ore per insegnarmi a essere un buon biostatistico e un buon ricercatore, ma, soprattutto, mi ha concesso la libertà di esprimere e coltivare le mie idee, spingendomi a seguirle e a credere in esse, anche quando potevano sembrare bizzarre. Ho capito che questa è una delle doti più rare in un supervisore, e farò di tutto per farne tesoro e ripagarla nella mia carriera futura.*

*Ringrazio anche il Dottor Giuseppe Marano per il suo supporto costante, per gli scambi illuminanti e i feedback critici, e per avermi insegnato tutto ciò che so sulle simulazioni Montecarlo e sulla programmazione in R. Devo ringraziarlo anche per le preziose "dritte" cinematografiche che hanno allietato il mio tempo libero.*

*Vorrei estendere la mia gratitudine al Professor Joel Schwartz, che mi ha accolto nel suo laboratorio durante il mio periodo come visiting researcher alla Harvard T.H. Chan School of*

*Public Health di Boston. Le sue lezioni e le sue preziose intuizioni sono state cruciali per comprendere le sfide della modellizzazione dei dati epidemiologici e hanno ispirato buona parte di questa tesi.*

*Un ultimo ringraziamento va alle Professoresse Laura Antolini e Clelia di Serio, per aver dedicato il loro tempo a questa tesi, offrendomi spunti e suggerimenti di grande valore.*

*Ma un dottorato non si fa solo tra le mura dell'università. Ho avuto la fortuna di avere accanto persone meravigliose a cui va il mio grazie più sincero.*

*Vorrei ringraziare la mia Famiglia per l'amore e il sostegno che non mi ha mai fatto mancare. E per Famiglia intendo non solo i miei affetti più stretti – Mamma, Papà, Caterina, Elia e mio fratello Davide; i nonni Marialuisa e Antonio; gli zii Betta e Fabio e le mie cugine Sofia e Bea – ma anche i miei migliori Amici: Marco, Giorgia, Margherita, per le serate, le risate e le fughe insieme, e Lorenzo (Il Falco) e Nicola (Il Picchio) per le infinite pedalate sulle Alpi.*

*Il loro affetto è stato un'ancora di salvezza, perché vivere a 6000 km da casa non è una passeggiata. Ma non è nemmeno così difficile, specialmente quando trovi una casa che puoi chiamare Casa. Ringrazio Andrea e Matt per avermi accolto nel loro splendido nido a Brookline, per avermi guidato alla scoperta del New England e per avermi fatto vivere la vera esperienza americana in un'atmosfera di calore e amicizia. Sono grato a Mike, il mio guru spirituale, per le innumerevoli chiacchierate in cucina sulla vita, l'amore e il futuro. Quella casa è stata resa ancora più speciale da Andrei e Farzaan, per la loro allegria e per essere stati i migliori coinquilini del mondo. La mia comunità americana si è allargata grazie alle passioni, e qui ringrazio il mio compagno di bici Ricky che, appena conosciuto, mi ha messo a disposizione una delle cose più importanti della vita: una bicicletta. Grazie di cuore anche ai suoi amici, ora anche miei, Charlotte, Tyler e Max, per i chilometri macinati insieme e per i gelati esageratamente grandi del dopo-corsa.*

*E infine, il ringraziamento più importante. Lo devo alla mia ragazza, Benedetta, per essersi presa cura di me e per avermi amato ogni singolo secondo che abbiamo passato distanti. Per aver sopportato quel misto di malumore e scemenza che le arrivava via FaceTime e per avermi aspettato fino a tardi, infischiandosene dei cicli del sonno. Lei è una grande fonte di ispirazione, e spero un giorno di avere per la mia ricerca la stessa passione e dedizione che lei ha per i suoi studi di archeologia.*

*E così, alla fine di questo percorso, tra tutte le cose che ho imparato, studiato e su cui ho ragionato, una mi è diventata chiarissima.*

*Il Tempo non esiste; è una nostra costruzione mentale. L'Amore, invece, esiste. Ed è ovunque.*

Giacomo (aka Il Condor)

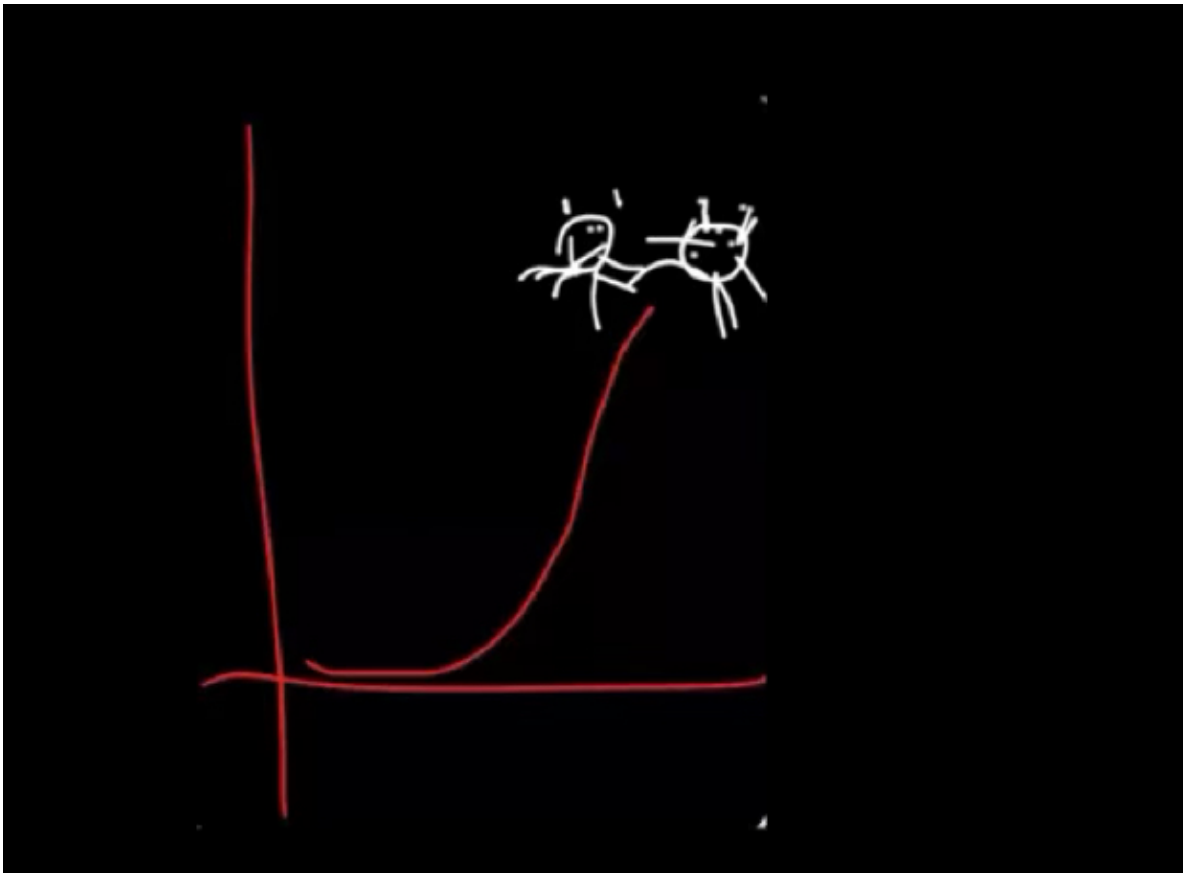


Figure 1.1: One of the best plots someone could ever made for me

# 2 Abstract

**Background:** Time-to-event analysis is a cornerstone of medical research, yet standard methods often rely on restrictive assumptions that limit their clinical utility and causal interpretability. The Hazard Ratio (HR), while popular, suffers from non-collapsibility and lacks a straightforward causal interpretation, particularly when proportional hazards (PH) assumptions are violated. Furthermore, single-number summaries often obscure the temporal dynamics of risk, failing to communicate *when* a treatment is most effective.

**Objective:** This thesis aims to evaluate, develop, and extend flexible parametric models to improve the estimation and communication of causal effects in survival analysis. It specifically targets the robust estimation of cumulative quantities (e.g., Risk Difference, Relative Risk) and the distributional analysis of health outcomes beyond simple summary statistics.

**Methods:** The work explores two primary modeling pathways: (1) indirect estimation via flexible hazard-based models (Discrete-Time Hazard and Piece-wise Exponential models) and (2) direct estimation using pseudo-observations.

- **Simulation Studies:** Extensive Monte Carlo simulations are conducted to determine the sample size requirements (Events Per Parameter) for these flexible models and to assess their performance under complex data-generating mechanisms (non-proportional hazards, time-varying effects).
- **Model Selection:** A novel statistical learning workflow is proposed, integrating bootstrap perturbation and time-dependent predictive metrics (e.g., Net Reclassification Improvement) to robustly select model complexity for the Cumulative Incidence Function (CIF) in competing risks settings.
- **New Estimators:** The thesis introduces an imputation-based Piece-wise Exponential model for the direct estimation of the sub-distribution hazard.
- **New Measures:** To improve risk communication, two novel estimands are defined: the Highest Risk Density Region (HRDR) and the Highest Net Risk Difference Region (HNRDR). These measures identify the specific time intervals where absolute risk is most concentrated or where the treatment effect is maximal.

**Results:** Simulation results indicate that flexible parametric models require approximately 20 events per parameter to achieve stable estimates of causal effects on the cumulative scale. The proposed imputation-based PWE model for competing risks demonstrates negligible bias and competitive performance against established methods. The HRDR and HNRDR frameworks successfully characterize the temporal distribution of risk, distinguishing between treatments

that delay event onset versus those that reduce overall event magnitude. Finally, the distributional regression framework is successfully generalized to non-censored continuous outcomes, offering a unified approach to dose-response analysis.

**Conclusion:** This thesis provides a comprehensive framework for the rigorous application of flexible survival models. By shifting the analytical focus from conditional rates (hazard ratios) to marginal cumulative effects and full probability distributions, it offers researchers tools to generate more causally sound, robust, and clinically interpretable evidence.

# 3 Background and Methodological Framework

## 3.1 Time-to-event analysis is about time: a historical perspective

The analysis of time-to-event (TTE) data is a cornerstone of medical and public health research, providing the fundamental tools to understand the temporal course of disease, the efficacy of interventions, and the prognosis of patients. The defining characteristic of TTE data is that the outcome encompasses not only *whether* an event occurred but, critically, *when*. This temporal dimension is fundamental, as the timing of an event—be it death, disease recurrence, or recovery—often dictates the perceived efficacy of a treatment and informs clinical decision-making. The history of the field reflects a continuous evolution of statistical methods designed to extract even more nuanced information from this temporal data, driven by the increasing complexity of the clinical questions being asked.

The origins of time-to-event thinking, in medical research, can be traced to early public health investigations that sought to characterize the natural duration of diseases. In a study in the British Medical Journal, Barlow and Leeming sought to describe the natural duration of common cancers at their primary sites (Lazarus-Barlow and Leeming 1924). Drawing on registry data, their objective was to quantify the interval between disease onset and death. The motivation was pragmatic and insightful: for chronic conditions like cancer, unlike acute illnesses, establishing a definitive "cure" is exceptionally challenging. The authors argued, therefore, that one of the most reliable indicators of therapeutic efficacy was the extension of survival time relative to a baseline or natural duration of the disease.

The statistical methods employed in this early work, while rudimentary by modern standards, reveal a profound interest in the full distribution of survival times. The results were presented in tables detailing the mean, minimum, and maximum durations for various cancers, but interestingly, the authors also reported the proportion of cases with duration below the mean. This reporting choice represents an early, intuitive attempt to capture the skewness of the survival distributions, demonstrating a recognition that, for researchers and stakeholders, the shape and characteristics of the distribution itself convey more meaningful information than a single summary statistic. This focus suggests an appreciation that a single average value could be misleading and that a more holistic description of event timing was necessary. This interpretive approach was reinforced by Greenwood's 1927 report ("A Report on the Natural Duration of Cancer." 1927), which also examined the natural duration of cancer, reporting expected survival times conditional on patient age and surgical procedures. In an era before the development of more advanced statistical tools, the expected survival duration was considered a direct and interpretable means of conveying results to a broad audience.

However, the application of standard statistical methods to TTE data is complicated by a defining methodological feature: censoring. Censoring arises when the time-to-event for some

individuals remains unknown, either because they are lost to follow-up or because the study concludes before they have experienced the event of interest. In such cases, the available information is incomplete, confirming only that the event time exceeds a certain threshold. While early work addressed censoring in primitive ways—for instance, by imputing time-to-death based on discharge status—the post-Second World War development of rigorous approaches to handle censored data marked a turning point for the field.

This methodological revolution was led by pioneering techniques such as the Kaplan-Meier estimator (1958) (Kaplan and Meier 1958) and the Nelson-Aalen estimator (W. Nelson 1969) (1960s), which allowed for the empirical reconstruction of the cumulative distribution function (CDF) while properly accounting for censored observations. Alongside the CDF and its corresponding probability density function (PDF), two new fundamental functions were introduced and rose to prominence: the survival function and the hazard function (Section 4.1). While the former tells the proportion of patients alive through-out the follow-up time, the latter describes the instantaneous probability of dying at a time point, given survival at that time. These became the principal tools for summarizing and interpreting TTE outcomes. An excellent example of the power of this new framework is the landmark clinical trial comparing 6-mercaptopurine (6-MP) to placebo for childhood leukemia. The results, presented as survival curves, clearly visualized the percentage of patients remaining in remission over time, powerfully demonstrating the benefit of 6-MP in a context where many children in the treatment arm were censored at the end of follow-up, having not yet relapsed (Gehan and Freireich 2011).

As therapeutic interventions advanced, the focus of clinical research expanded from the singular endpoint of death to more nuanced measures of disease progression. This led to the analysis of intermediate events, such as time-to-distant recurrence in breast cancer, and introduced the additional complexity of competing risks—events that preclude the occurrence of the primary event of interest. For example, a patient who dies from other causes cannot subsequently experience a distant recurrence. To address this, specific statistical frameworks were developed, centered on estimands like the cause-specific cumulative incidence function (CIF). The CIF quantifies the cumulative probability of experiencing a specific type of event first, in the presence of other event types. This estimand is associated with two distinct hazard functions: the cause-specific hazard and the sub-distribution hazard (Section 4.2).

## 3.2 Interpretational Issues of the Hazard Ratio

To evaluate the impact of covariates on time-to-event outcomes, a suite of regression models was developed, with the Cox proportional hazards model (Cox 1972) emerging as the dominant paradigm. Its extension to the competing risks setting, the Fine and Gray proportional sub-distribution hazards model (Fine and Gray 1999), similarly achieved widespread adoption for its ability to model on the CIF. The appeal of these semi-parametric models is undeniable. They allow for the estimation of prognostic effects—summarized by the hazard ratio (HR)—without requiring the specification of a rigid parametric distribution for the underlying event times. When the central assumption of proportional hazards (PH) holds—that is, when the ratio of hazards between two groups is constant over time—the HR provides an elegant,

single-number summary of a covariate's effect, making it exceptionally attractive to applied researchers.

However, the HR suffers from severe interpretational deficiencies (Weir et al. 2019). Although statistically robust, the HR is a ratio of *instantaneous event rates*, a subtle concept that is frequently misinterpreted in the medical literature (Sutradhar and Austin 2018) as a more intuitive relative risk (a ratio of probabilities). This has led to widespread confusion in communicating research findings to clinicians and patients. The common "fix" for NPH, modeling time-varying HRs, only exacerbates this problem. A time-varying HR that declines over time and falls below unity might be erroneously interpreted as a protective effect, when it may simply be a statistical artifact arising from a dwindling number of subjects at risk in the tail of the distribution.

Moreover, the widespread adoption and seductive simplicity of the HR mask profound limitations that can lead to misleading conclusions and flawed interpretations. The very feature that makes the HR so popular—its ability to condense a complex, time-dependent process into a single number—is also its greatest weakness. This is because the PH assumption is frequently violated in clinical reality, particularly in studies with long follow-up. Treatment effects are rarely constant; an intervention might have a delayed effect, a benefit that wanes over time, or a risk profile that leads to crossing survival curves. In these common scenarios of non-proportional hazards (NPH), a single, time-averaged HR becomes dangerously misleading, as it obscures critical time-dependent variations in efficacy or risk. For instance, in a breast cancer study, a large tumor size may be strongly prognostic for death early after surgery, but its effect may attenuate over longer follow-up. A constant HR would average away this crucial dynamic, mischaracterizing the biological process and reducing predictive accuracy.

## 3.3 The issues of the Hazard Ratio in causality

In epidemiological observational studies, understanding if a specific exposure (like a medication, an environmental toxin, or a lifestyle choice) causes a particular health outcome is often the main interest. The problem is that these studies are not randomized controlled trials. Because individuals are not randomly assigned to exposure groups, the core assumption of exchangeability (Section 4.3), guaranteed by randomization, is violated. This is particularly critical nowadays when, for instance, administrative databases are utilized for post marketing surveillance of drugs.

From a causal inference perspective, the HR is a fundamentally flawed measure of treatment or exposure effect (Hernán 2010). This is due to a phenomenon known as 'collider stratification bias'. The hazard function is, per its definition, conditional on survival up to a given time $t$. However, survival itself is a consequence of both the treatment and other prognostic factors (confounders). In a directed acyclic graph, survival is a "collider"—a variable influenced by two or more other variables. Conditioning on a collider, which is implicitly done when calculating a hazard, opens a spurious, non-causal statistical association between the treatment and the confounders within the stratum of survivors. Consequently, the HR is contaminated by this bias and does not represent a pure causal effect. This can be understood from another

perspective: as a study progresses, the risk sets at later time points are composed of selected subgroups of survivors who are no longer the randomized equivalents they were at baseline, undermining causal comparisons (Section 4.3). Imagine a study of a new medication for a severe illness. As time progresses, the group of patients still alive and being followed is no longer representative of the original study population. They are 'super-survivors' who have survived due to a combination of the treatment they received (or didn't) *and* their own underlying prognostic factors. Comparing the treatment and control groups at later time points is therefore no longer a fair, apples-to-apples comparison, because the groups have been selectively depleted. Calculating a hazard ratio implicitly makes these unfair comparisons at every moment in time, introducing a subtle but powerful bias known as collider stratification.

Furthermore, the HR is a non-collapsible measure, a property it shares with the odds ratio. This means that a marginal (unadjusted) HR can differ from a conditional (adjusted) HR even when adjusting for a variable that is not a confounder. This mathematical property arises from the non-linearity of the link function (complementary log-log) used in the Cox model and makes it difficult to interpret the HR as a stable, population-level effect measure, as its value can change simply by including non-prognostic variables in the model.

Nevertheless, it is important to note that approaches to causality focused on the hazard itself were proposed. Despite the limitations inherent to the potential outcomes view of risk, an alternative methodological lineage retains the hazard rate as the fundamental building block of causality. Rather than discarding the hazard due to the complexities of survivor selection, Dynamic Path Analysis explicitly models the dynamic evolution of the event process (Strohmaier et al. 2015).

## 3.4 Causal estimands of the average treatment effects (ATE)

The profound statistical, interpretational, and causal deficiencies of the hazard ratio, as detailed above, necessitate a fundamental shift in focus away from conditional rates and toward marginal effects. The field of causal inference provides a robust alternative through a class of estimands designed to capture the Average Treatment Effect (ATE) on clinically meaningful scales, such as the Risk Difference ( Equation 3.1 ), the Relative Risk ( Equation 3.2 ), usually evaluated at specific and pre-defined time-points of the follow-up, and the Difference in Restricted Mean Survival Time ( Equation 3.3 ) that offer more direct, robust, and interpretable causal effects.

The causal Risk Difference (RD), also known as the absolute risk reduction, quantifies the absolute change in the probability of experiencing an event by a specific time point $\tau$ due to the treatment. It is defined as the difference between the cumulative incidence functions under the two potential treatment scenarios:

$$\theta_{RD}(\tau) = \mathbb{P}(T^{(1)} \leq \tau) - \mathbb{P}(T^{(0)} \leq \tau) \tag{3.1}$$

This can also be expressed in terms of the potential survival functions, $S^{(a)}(t) = \mathbb{P}(T^{(a)} > t)$, as$\theta_{RD}(\tau) = (1 - S^{(1)}(\tau)) - (1 - S^{(0)}(\tau)) = S^{(0)}(\tau) - S^{(1)}(\tau)$.

The RD provides a direct and clinically meaningful measure of impact. For example, an RD of -0.05 at 5 years implies that the treatment causally reduces the absolute risk of the event occurring within 5 years by 5 percentage points. This measure is particularly valuable for clinical decision-making and public health policy, as it directly translates to the number of events prevented in a population.

The causal Relative Risk (RR), or risk ratio, measures the multiplicative effect of a treatment on the cumulative risk of an event by time $\tau$. It is defined as the ratio of the potential Cs:

$$\theta_{RR}(\tau) = \frac{\mathbb{P}(T^{(1)} \leq \tau)}{\mathbb{P}(T^{(0)} \leq \tau)} \tag{3.2}$$

An RR of 0.8 at 5 years would be interpreted as the treatment reducing the risk of experiencing the event by that time by 20% relative to the control group. The RR is often favored in epidemiological research because it can be more stable across populations with different baseline risks than the RD. However, its primary limitation is that it can be misleading about the absolute clinical impact. A large RR (e.g., 2.0, indicating a doubling of risk) may correspond to a negligible RD if the baseline risk in the control group, $\mathbb{P}(T^{(0)} \leq \tau)$, is very small (e.g., an increase from 0.001 to 0.002).

Finally, the Difference in Restricted Mean Survival Time ($\theta_{RMST}$) provides a summary of the treatment effect over a specified interval $[0, \tau]$, rather than at a single point in time. It is defined as the difference in the expected survival time, truncated at a pre-specified time horizon $\tau$:

$$\theta_{RMST}(\tau) = \mathbb{E}(T^{(1)}) \wedge \tau - \mathbb{E}(T^{(0)}) \wedge \tau \tag{3.3}$$

The $\theta_{RMST}$ has two complementary and intuitive interpretations. First, it represents the average event-free time gained (or lost) due to the treatment over the interval $[0, \tau]$. For instance, a $\theta_{RMST}(5 \text{ years}) = 0.5$ years means that, on average, the treatment extends event-free survival by half a year over the first five years of follow-up. Second, it is mathematically equivalent to the area between the two potential survival curves from time 0 to $\tau$:

$$\theta_{RMST}(\tau) = \int_0^\tau S^{(1)}(t)dt - \int_0^\tau S^{(0)}(t)dt = \int_0^\tau (S^{(1)}(t) - S^{(0)}(t))dt \tag{3.4}$$

This property makes it a robust summary measure of the cumulative separation between the survival curves over a clinically relevant period.

## 3.5 Flexibly modeling cumulative quantities

The reliability and robustness of the causal estimands of the ATE in survival analysis depends on the quality of the counter factual functions used to describe the potential outcomes such as the cumulative survival and cumulative incidence functions ( Equation 4.17 and Equation 4.18). This thesis explores the performance of, and develops upon, two distinct but

complementary methodological pathways for modeling these cumulative quantities and for obtaining the causal measure cited above, moving beyond the limitations of the traditional Cox and Fine & Gray frameworks.

The first pathway is a direct approach that leverages the technique of pseudo-observations. This powerful method, derived from jackknife re-sampling, generates a complete, "uncensored" outcome value for each individual at a series of pre-specified time points based on a non-parametric estimator like the Kaplan-Meier and Aalen-Johansen for the CDF or the CIF respectively. For instance, one can compute a pseudo-observation for the 5-year cumulative incidence for every subject in the study.

The great advantage of this transformation is that pseudo-observations can then be used as the outcome variable in a standard Generalized Linear Model (GLM). This allows for the direct modeling of the cumulative quantity of interest and offers immense flexibility in the choice of link function, enabling the estimation of effects on various scales, such as proportional odds or proportional risks. Furthermore, by including interaction terms between covariates and the function of time (smoothed with splines), this framework can readily accommodate time-varying effects on the cumulative scale, providing an appealing alternative when classical assumptions are violated.

The second pathway is indirect; it involves first building a flexible model for the hazard function and then deriving the cumulative quantities through their fundamental mathematical relationship (product limit estimator) (e.g., $F(t) = 1 - \prod_{s=1}^{t}(1 - \lambda(s))$). The two primary modeling frameworks for this purpose explored in this thesis are discrete-time hazard models and piece-wise exponential (PWE) models. While both partition the time axis into intervals, they differ in their formulation: discrete-time models treat time as ordinal and estimate conditional probabilities of failure in each interval, whereas PWE models assume a continuous-time process where the hazard is constant within each interval.

These frameworks are exceptionally flexible. The baseline hazard can be modeled non-parametrically by including dummy variables for each time period, and time-varying covariate effects are naturally handled via interactions between covariates and time or functions of time. A more convenient way is achieved by adopting splines functions. Splines allow the baseline hazard to adopt complex, non-monotonic shapes dictated by the data itself, using a limited set of parameters to yield smooth, continuous functions. This flexibility is not only crucial for accurately deriving cumulative risk estimates but also for accommodating non-proportional hazards by allowing the effect of a covariate to vary smoothly over time through interactions with the spline-based time terms.

The relative merits, specific applications, and comparative performance of these two complementary pathways will be a recurring theme in the chapters that follow. This is particularly relevant to our proposed workflow for model selection Section 3.7, where the choice of modeling strategy has significant practical implications, and in our development of novel estimators Chapter 8, which extends the indirect, hazard-based framework."

## 3.6 Simulation Study of EPP Requirements for Flexible Parametric Survival Models

While these flexible parametric models offer a powerful and versatile solution for estimating causal effects, their statistical performance is not guaranteed. Like any regression framework, their reliability is critically dependent on the amount of information available in the data—a dependency that, in survival analysis, is governed not by the total sample size, but by the number of observed events. This raises a crucial practical question regarding the sample size requirements for the stable application of these advanced models, a topic we address next.

A widely used heuristic for assessing sample size adequacy is the "Events Per Variable" (EPV) rule. However, this metric can be misleading as it does not fully account for the model's parametric complexity. A more precise and statistically rigorous metric is "Events Per Parameter" (EPP), which quantifies the ratio of events to the number of parameters being estimated.

The distinction is crucial because a single predictor variable can consume multiple parameters and, consequently, degrees of freedom. For example, a categorical predictor with k levels requires the estimation of k-1 parameters. Similarly, modeling a non-linear effect of a continuous predictor using a restricted cubic spline with k knots necessitates estimating k-1 parameters. Therefore, EPP provides a more accurate measure of the potential for data sparsity—a condition where the number of events is low relative to the model's complexity—which can lead to model instability, overfitting, and unreliable estimation of regression coefficients.

For the Cox proportional hazards model, a minimum of 10 EPP has been a long-standing, though debated, recommendation for achieving stable parameter estimates (Peduzzi et al. 1996, 1995; Riley et al. 2018). This flexibility, however, comes at the cost of increased parametric complexity. Unlike the semi-parametric Cox model where the baseline hazard is left unspecified, fully flexible parametric models require its explicit parametrization. The use of splines (e.g., B-splines) introduces several parameters to model the effect of time. Furthermore, if the proportional hazards assumption is violated, interactions between covariates and time must be specified, further increasing the parameter count.

For instance, a Cox model for the effect of a binary treatment, adjusting for a linear continuous predictor and a binary predictor, estimates only three parameters. In contrast, a PWE model of the same relationship, using a B-spline with one internal knot for the baseline hazard, would require additional four parameters. If the baseline hazard is stratified by a binary covariate, the parameter count would increase substantially again.

However, the theoretical flexibility of these models is accompanied by a critical, and largely unaddressed, practical question: what are the sample size requirements for their stable application? Existing guidelines, developed for the simpler semi-parametric Cox model, are insufficient for this new class of parametric models. Chapter 5 directly confronts this knowledge gap by providing the first systematic investigation of EPP requirements for flexible parametric models when the target of inference is the causal ATE.

We will conduct a Monte Carlo simulation study across five scenarios of increasing data-generating complexity, relaxing assumptions about prognostic relationships and hazard proportionality. The primary focus will be on evaluating the performance (i.e., bias, variance,

and mean squared error and absolute mean error) of clinically relevant ATEs—such as the risk difference (RD), risk ratio (RR), and RMST difference—estimated via standardization.

## 3.7 A workflow for modeling cumulative risk: model selection with the predictive power assessment

In addition to limited sample size, a primary challenge with flexible models is accounting adequately for the potential complexity of the prognostic relationship, avoiding inaccurate estimates from a misspecified model. As model complexity increases, so does the risk of overfitting or obtaining unreliable estimates; conversely, reducing model complexity may yield more stable results but could lead to oversimplified patterns Kantidakis et al. (2023) . The bias-variance trade off should be exploited to identify the "the best" model in an exploratory framework aimed at investigating prognostic relationship and/or predicting counter factual cumulative probability functions.

Several approaches can be considered to this end. Researchers might apply formal hypothesis tests to detect non-proportionality of hazards and interaction effects, while others, in an exploratory framework, might base their choice solely on minimizing information criteria Akaike (1992) computed from the model's maximum likelihood. Often the evaluation of the model complexity is based on the use of a single procedure on the original data and the simplest model is chosen when there is lack of statistical significance for the adopted test or at the same values of Akaike Information Criterion (AIC). No perturbation methods are typically applied to investigate the robustness of the model choice when a simple model structure is selected.

Alternatively, some researchers focus only on measures of predictive ability, such as discrimination and calibration, alongside internal and external validation procedures (Efron and Tibshirani 1994; HARRELL, LEE, and MARK 1996). The most popular measure of discrimination is the area under the Receiver Operating Characteristic (ROC) curve Harrell , (2015), which has been adapted for censored survival data, competing risks Wolbers et al. (2014) and as a time-dependent measure (Antolini, Boracchi, and Biganzoli 2005; Kamarudin, Cox, and Kolamunnage-Dona 2017) .

Other measures, such as net reclassification improvement, are proposed for quantifying the additional sensitivity and specificity a model gains from increased complexity (Alba et al. 2017; Pencina, D'Agostino, and Steyerberg 2010). In modeling complex effects, model discrimination and calibration may give additional useful information, as assessing the robustness of a model structure investigated. Indeed, they can be employed to evaluate how critical the consideration of complex time dependent and interaction effects in the model selection step is for the future predictions at the individual level. Although information criteria can be seen as indicators of model discrimination, they are somewhat incomparable, as they lack a minimum (absence of discrimination) and a maximum (perfect discrimination).

Given the numerous methods available for selecting model complexity, an important consideration is how to effectively combine these methodologies to identify a model that is both robust

in capturing the complexity of the phenomenon under study and remains interpretable. Additionally, it is beneficial to examine whether the performance of this integrated model complexity selection approach, which leverages information from the aforementioned methods, differs between the two methods of obtaining the cumulative quantities—on the instantaneous scale and directly on the cumulative scale for the pseudo-observation approach-.

Chapter 6 proposes a workflow to analyze data in a competing risk framework where CIF is the measure of interest and complex effects are expected. Particularly, a method to evaluate the performance of different model structures according to the issues above described is proposed and illustrated.

To this end, data from two historical breast cancer clinical trials (Umberto Veronesi et al. 1981; U. Veronesi et al. 2001), characterized by a long follow-up, will be analyzed. The analysis will be focused on a clinical interest concerning the risk of occurrence of distant recurrences (distant metastases) depending on treatment and axillary nodal status, which the main results were already published. Distant recurrences are recorded regardless of intra-breast tumor recurrences and contralateral tumors but are influenced by competing events such as other tumors or death (related or unrelated to the disease).

## 3.8 A Piece-wise exponential model for the sub-distribution hazard of an event

Among the models for directly and flexibly estimating the cause-specic cumulative incidence function of an event, an extension of the PWE model for this purpose has not yet been introduced.

PWE framework have been applied to model cause specific hazard (Equation 4.8 ) in competing risks (Bender et al. 2018). However, if the goal is to model the effects of the covariates on the CIF for a single primary outcome, constructing and combining multiple, potentially complex PWE-CSH models (each with interval-specific effects to capture non-proportionality for different event types) can lead to considerable model complexity and parameter burden, contrasting with the potential parsimony of an approach based on the sub-distribution hazard (Equation 4.11 ) for that single CIF.

Chapter 7 aims to fill this gap. We propose an extension of the PWE model specifically for estimating the SDH of a primary outcome. Rather than deriving and implementing a direct weighting scheme for the PWE framework in the SDH context, as proposed by Berger et al. (2018), we adopt an imputation-based strategy. This approach builds upon and extends the methodology proposed by Ruan and Gray (2008).

The core principle involves treating the potential censoring time for individuals who experience a competing event as missing data. This potential censoring time is then imputed from an estimated censoring distribution. A key advantage of this imputation strategy is that, once competing events are effectively transformed into imputed censorings, it allows for the application of existing, well-understood PWE fitting procedures.

To thoroughly assess this imputation-based PWE-SDH model, our simulation study evaluates its performance under correctly specified imputation models for the censoring distribution, reflecting different assumptions about the censoring mechanism and compares its performances with other flexible competing risks models.

## 3.9 Beyond cumulative risk for examining the full event-time distribution: Introducing the Highest Risk Density and Highest Net Risk Difference Regions

Having established and extended robust methods for modeling cumulative risk, the inquiry must be pushed a step further.

While it is technically feasible to estimate complex, time-varying Hazard Ratios (HRs), their communication to clinicians presents significant challenges due to the causal and interpretative issues previously discussed. Conversely, relying solely on ATEs derived from counterfactual cumulative risks is also problematic. The proportional hazards assumption is frequently violated in practice, and it follows that the proportionality of cumulative risks is equally unlikely to hold. Consequently, the difference between two time-to-event distributions cannot be captured by a single metric and must be analyzed across multiple follow-up times. Therefore, to effectively communicate treatment effects, it is often more intuitive and desirable to summarize the differences between survival distributions on the time scale (e.g., difference in median survival time) rather than on the probability scale (e.g., absolute risk reduction).

These limitations have motivated the adoption of alternative summary measures like the RMST as it relies on neither the proportional hazards nor the proportional risks assumption, and it summarizes the effect on a time scale and not a probability scale. The RMST integrates the entire survival experience within a specified interval into a single summary statistic. This could be considered both an advantage and disadvantage.

The difference between restricted mean survival times (RMST) can be misleading when the distributions of survival times are highly skewed, for the very same reason that a difference in simple means can be misleading for skewed non-censored data. The mean is sensitive to extreme values, and in survival analysis, a long "tail" of a few long-term survivors can pull the mean survival time upwards, potentially masking what is happening for the majority of the population.

While RMST mitigates this by cutting off the tail at a specific time point ($\tau$), it doesn't eliminate the problem entirely. The magnitude, and sometimes even the direction, of the difference in RMST can depend heavily on the chosen truncation time, $\tau$. If the survival curves cross, for example, choosing a $\tau$ before the crossing point will show one treatment is better, while choosing a $\tau$ after might show the opposite or no difference. A highly skewed distribution (a long tail) in one group can disproportionately inflate its RMST as $\tau$ gets larger.

Moreover, a new treatment might offer a huge survival benefit to a small subgroup of patients (creating a long, thin tail) but a negligible benefit to the majority. The difference in RMST

would be positive, driven by that small subgroup. This could be misinterpreted as a general benefit for all patients, when the difference in median survival time might be zero. Finally, the RMST summarizes the entire survival experience up to $\tau$ into a single number. It averages out early, dramatic differences and late, modest differences. If one drug works much faster but its effect wanes, while another works slowly but is durable, their RMSTs at a certain $\tau$ could be identical, completely hiding the different clinical dynamics.

Quantile regression for censored survival data offers another alternative to communicate the results on the time-scale, enabling the estimation of covariate effects on specific quantiles of the event-time distribution Peng (2021) . Its primary strength is the flexibility to model effect heterogeneity across the distribution. This flexibility, however, introduces its own analytical considerations. The selection of quantiles for analysis introduces a degree of subjectivity and potential for selective reporting. Moreover, because the effect is rarely constant across quantiles, summarizing the overall treatment effect becomes complex. A significant practical limitation is right censoring, which can preclude the estimation of higher quantiles, if the observed event rate is insufficient. It is worth noting that sometimes the tails of the distribution are often of main interest when quantile regression is applied.

The Accelerated Failure Time (AFT) model offers an appealing conceptualization of a treatment effect, formalizing it as a "stretching" of the time-to-event distribution by a constant factor. In clinical contexts where risk elimination is not assumed, the treatment acts to delay events and this temporal stretching is conceptually equivalent to a redistribution of the probability mass of events toward later follow-up times. However, AFT models require specifying a parametric distribution for the baseline failure time. An incorrect distributional assumption can lead to significant model misspecification and biased estimates of the treatment effect.

This thesis moves beyond conventional metrics based on the instantaneous (hazard) or cumulative (risk) probability scale. We also critique simple time-scale measures that condense the entire survival distribution into a single value, as such summaries are overly sensitive to the study's follow-up duration (truncation time). Instead, we argue that the investigation of the *whole* probability density functions (PDFs), $f(t)$ and CDF, $F(t)$ allows a more granular understanding of the redistribution of risk mass due to a treatment. A treatment benefit manifests not as a simple location shift of the density curve, but as a change in its shape -typically a reduction in height and increase in width- reflecting a temporal dispersion of risk. The analytical objective thus becomes the quantitative characterization of this risk density redistribution.

While adapted histogram and boxplot methods can visualize the PDF for censored data (Barnett and Cohen 2000), a more formal framework is needed to quantify and compare distributional shapes to maximize the communicability of a treatment's effect on a temporal scale, particularly for non-statistical stakeholders.

This raises the question of how to formally summarize such whole distributions. For right-skewed distributions characteristic of survival data, conventional intervals, such as those derived from quantiles or equal-tailed partitioning, are often sub optimal as they may not contain the mode of the distribution, that is the time point of highest risk.

In Chapter 8 we therefore propose a method based on the concept of high-density regions, first described by Hyndman (1995). We define a Highest Risk Density Region (HRDR) as the

narrowest interval over which a specified risk (probability of the event) mass, $1-\alpha$, integrates. This interval always contains the mode of the density of the time-to-event distribution. Within this framework, a treatment effect can be characterized by its influence on the location (timing) or scale (width) of these highest-risk intervals. A protective effect, for example, might manifest as a temporal delay of the interval or a widening of the interval, indicating that a longer time is required to accumulate the same risk mass. While the HRDRs provides an essential characterization of the temporal risk pattern within a specific treatment arm or exposure group, a formal measure is required to isolate and quantify the effect, that is a contrast between PDFs and CDFs. Following the logic of HRDRs, we propose a novel measure, that we call Highest Net Risk Difference Region (HNRDR), defined as the narrowest time window over which a specified net absolute risk difference is achieved. This measure exploits the instantaneous risk difference function to isolate the period of maximal therapeutic impact and it is advantageous in describing the improvement in survival time due to treatment.

This tension -between statistical convenience (e.g., the single-number summary of the HR) and clinical or causal interpretability-is a defining challenge across the entire field of TTE analysis. In the present work, we try to participate in a larger paradigm shift toward more descriptive, assumption-free, and communicable effect measures. The HRDR/HNRDR fits into this landscape by addressing a specific and important niche: characterizing the shape and temporal location of the risk distribution, something none of the other measures do as directly. In this framework HRDR/HNRDR not as a replacement for measures like RMST, but as a valuable complement that answers a different clinical question: "When is the treatment working the most?" and "When is the risk highest?".

## 3.10 A Unified Framework for distributional regression in Health Research

Up to now, this thesis adopts statistical frameworks that allow for modeling complex relationship between an exposure and a time-to-event censored outcomes and communicating complex model results with more interpretable and insightful statements about the outcomes. In this thesis, we argue that this shift of paradigm might be advisable also to the context of non-censored continuous, and possibly multivariate, health outcomes.

Understanding dose-response relationships of non-censored outcomes is fundamental in epidemiology and public health, as it provides crucial evidence for shaping policies that protect individual and community health (Doll and Hill 1956). Exposure is often analyzed on a continuous scale to capture the full complexity of its relationship with the outcome, including potential nonlinear effects (Redelmeier and Zipursky 2023; May and Bigelow 2005) and the disadvantages of using categorized continuous predictor variables have been extensively explained (Altman and Royston 2006; Royston, Altman, and Sauerbrei 2005). For analogous reasons, when dealing with a continuous outcome variable, it is preferable to use it on the original scale rather than categorizing it (Ragland 1992; Dawson and Weiss 2012). To illustrate the above issue, suppose the interest lies examining the relationship between physical activity -measured with Metabolic Equivalents (METs)- and metabolic health. A traditional strategy of analysis might involve assessing the presence or absence of Metabolic Syndrome

(MSy). To do that, one might model the (log) odds or prevalence of MSy as a function of PA using logistic or log-linear regression. However, this binary classification presents methodological challenges mostly concerning critical loss in the information one could retrieve from the continuous variables (Beckstead and Beckie 2010) and the naturally greater sample size required to fit a logistic model (Riley et al. 2018). Further motivation for not using the categorized MSy are, first, that it represents an advanced stage of metabolic dysfunction, whereas the beneficial effects of physical activity might be more evident in earlier stages of metabolic dysregulation. Additionally, MSy is a multifactorial condition, meaning its individual components could be affected differently by physical activity. Indeed, in the studies presented in the literature, secondary analyses often examine these components separately but still treat them as binary outcomes(Zeng et al. 2024).

In conclusion, a more rigorous approach would focus on evaluating the association of the exposure to the components of MSy, measured on their original scale, and to their joint distribution.

The most common approach for analyzing continuous outcomes involves modeling the expected value (mean) of the outcome variable as a function of exposure variables using ordinary least squares regression. The generalized linear model (GLM) extends this capability by accommodating non-Gaussian distributions through flexible distributional assumptions (e.g., binomial, poisson, gamma, inverse Gaussian).

Although GLMs allow a variety of distributions for the error term of the outcome variable, with the only requisite that they have to be part of the exponential family, the main focus is often on the expected value. Moreover this requires to specify the relationship between mean and variance.

Understanding the probability distribution of an outcome in relation to an exposure is crucial because exposure effects may alter not just the mean but the entire shape of the outcome distribution (Canale, Durante, and Dunson 2018). Even though one can obtain the conditional probability density function (PDF) of the response variable as a function of the covariates -considering the maximum likelihood estimates of central tendency and dispersion-, this is tightly linked to the distributional assumptions.

The linear model, which assumes normality and homoschedasticity of the (conditional) distributions, estimates an effect on the location of the distribution that is constant on all of its percentiles. However, an exposure could affect both the location of the distribution and the variability (scale) around that location. In other cases, an exposure might only impact the dispersion while leaving the location unchanged. A model focused solely on the location would erroneously conclude there is no association in the latter case. Furthermore, if an outcome's distribution is bimodal (or multimodal) for some non-accounted factor, its mean could fall in a low-density region between the peaks. Modeling this single, unrepresentative value would completely obscure the underlying nature of the relationship.

Several methods have been proposed in literature to address this limitation. Generalized Additive Model for Location, Scale and Shape (GAMLSS) is a flexible statistical modeling framework that extends GLMs and GAMs by allowing the modeling of more than just the mean (location) of a distribution (Stasinopoulos and Rigby 2007). However, the models can become complex and computationally intensive, often leading to convergence issues. There's

an increased risk of over fitting due to the use of multiple smoothers and distribution choices, which also require careful selection and justification. Additionally, interpreting the results—particularly for shape parameters—can be challenging (Bohl et al. 2012; Zavorsky 2025) .

Quantile regression involves modeling specific percentiles of the outcome distribution as a function of exposure variables (Khadka et al. 2023). By fitting several regression models for specific quantiles of the response distribution one can retrieve a 'discretized' version of the density function.

Non-parametric methods like conditional Kernel Density Estimation can estimate conditional PDFs (Fan 2004; Hall, Wolff, and Yao 1999; Hyndman and Yao 2002). Yet, cKDE struggles when multiple covariates are involved, especially with limited data. Other approaches, such as kernel regression, model covariate effects on expected values rather than full distributions (Nadaraya 1964).

The conceptual power of modeling the full probability distribution is not confined to the time-to-event setting. A central contribution of this thesis is the generalization of the flexible modeling machinery developed for survival analysis into a unified, flexible parametric framework for estimating the conditional PDF and CDF of *any* numerical health outcome, even in the absence of censoring. Chapter 9 reframes the tools of survival analysis, not as a niche methodology for a specific data type, but as a powerful and general approach to distributional regression.

## 3.11 Summary

This thesis assesses the performance of flexible statistical models for analyzing time-to-event data, particularly in scenarios with limited sample sizes, where the primary measure of interest is the causal Average Treatment Effect (ATE). A key contribution is the proposal of a statistical learning workflow designed to navigate model complexity, aiming for the most robust and generalizable estimates of cumulative probability functions used for risk communication and ATE computation.

Furthermore, this work introduces and evaluates a novel approach based on the piece-wise exponential model for estimating the cause-specific cumulative incidence function, (the probability of observing a focus event in the follow-up time, accounting for the other competing events) in the context of competing risks, comparing its performance against other flexible methodologies.

A central argument of this thesis is that the analysis of an exposure's impact should extend beyond single-point estimates to encompass the entire distribution of health outcomes, a principle applicable to both survival analysis and non-censored data. This is achieved by leveraging flexible statistical models to target and visualize complete probability density functions (PDF) and cumulative distribution functions (CDF), thereby summarizing shifts in probability mass. To this end, the concepts of Highest Risk Density Regions and High Net Risk Difference Regions are introduced for time-to-event analysis and their extension to non-censored health outcomes is demonstrated.

For clarity, a roadmap of the thesis is provided in Table 3.1 . Each subsequent chapter is structured like a scientific paper, with dedicated methods, results, and discussion sections. Methodologies specific to each chapter are presented therein, while a concluding "Technical Points" chapter details the foundational methods common to the entire body of work.

Table 3.1: A roadmap of the thesis

| Identified Methodological Challenge / Research Question | Limitations of Existing Methods | Relevant Thesis Chapter | Novel Contribution of this Thesis |
|---|---|---|---|
| How many events are needed for stable estimation of causal effects with flexible parametric models? | Existing "10 EPP" rules are for Cox models and focus on coefficients, not ATEs on the cumulative scale. | Chapter 5 | Provides the first EPP guidelines for DTH/PWE models targeting RD, RR, and RMST via standardization. |
| How can we robustly select model complexity (e.g., interactions, time-varying effects) for CIF estimation? | Standard methods (AIC on original data, hypothesis tests) can be unstable and lack robustness checks. | Chapter 6 | Proposes a statistical learning workflow integrating bootstrap perturbation and predictive accuracy metrics (NRI). |
| How can we flexibly model the sub-distribution hazard (SDH) to directly estimate the CIF? | No PWE model for the SDH exists, limiting flexibility for direct CIF modeling. The CSH approach is complex. | Chapter 7 | Introduces a novel imputation-based PWE model for the SDH. |
| How can we move beyond single-summary effects (HR, RMST) to communicate the full distributional impact of a treatment? | Existing measures average away or ignore the temporal dynamics of risk. PDF plots are hard to quantify. | Chapter 8 | Introduces the HRDR and HNRDR as novel causal estimands and communication tools for the time scale. |
| Can this distributional regression framework be generalized beyond survival analysis? | Tools for distributional regression are often complex (e.g., GAMLSS) or less integrated with causal methods. | Chapter 9 | Generalizes the flexible survival modeling machinery to estimate the conditional PDF of any continuous outcome. |

# 4 Technical Points

## 4.1 Functions that characterize any random variable, specifically time-to-event $T$

The probability distribution of the time-to-event random variable $T$ can be uniquely characterized by four fundamental, mathematically equivalent functions: the probability density function, the cumulative distribution function, the survival function, and the hazard function. A complete understanding of these functions and their interrelationships is essential for both standard and competing risks survival analysis. The precise mathematical definitions differ depending on whether time is treated as a continuous or a discrete variable. When the time-to-event variable $T$ is assumed to be a continuous, non-negative random variable, its distribution can be described by the following functions:

- Probability Density Function (PDF), $f(t)$: The PDF describes the probability of the event occurring within an infinitesimally small interval of time around $t$. It is formally defined as the derivative of the cumulative distribution function, $f(t) = F'(t)$, and can be conceptualized as the rate of event occurrence at a specific point in time. Mathematically,

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \tag{4.1}$$

  The PDF must satisfy $f(t) \geq 0$ for all $t \geq 0$ and $\int_0^\infty f(t)dt = 1$.

  The equivalent in the discrete setting is the Probability Mass Function (PMF), $f(t_j)$. The PMF gives the probability that the event occurs exactly at time $t_j$.

$$f(t_j) = P(T = t_j) \tag{4.2}$$

- Cumulative Distribution Function (CDF), $F(t)$: The CDF gives the cumulative probability that the event has occurred by or at time t. It is the integral of the PDF from the time origin to $t$.

$$F(t) = P(T \leq t) = \int_0^t f(u)du \tag{4.3}$$

  $F(t)$ is a non-decreasing function with $F(0) = 0$ and $\lim_{t\to\infty} F(t) = 1$. In the context of a single event type, this function is equivalent to the cumulative incidence of the event.

- Survival Function, $S(t)$: The survival function, a central quantity in survival analysis, is the complement of the CDF. It represents the probability that an individual has survived (i.e., has not experienced the event) beyond time t

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du \tag{4.4}$$

- Hazard Function, $h(t)$: The hazard function, or hazard rate, describes the instantaneous potential for the event to occur at time $t$, conditional on the individual having survived up to that time. It is a rate, not a probability, and its value can exceed 1. Its formal definition is

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{4.5}$$

In the discrete setting, the hazard is a conditional *probability*, not a rate. It is the probability that the event occurs during the interval beginning at $t_j$, given that the individual was event-free up to the start of that interval.

$$\lambda(t_j) = P(T = t_j \mid T \geq t_j) \tag{4.6}$$

## 4.2 Functions that characterize the random variable time-to-event $T$, in face of the competing events

When competing risks are present, the primary estimand of interest is often the absolute risk of a specific event type occurring over a given time frame. Standard methods are ill-suited for this task. This section introduces the appropriate quantity for this purpose, the cause-specific cumulative incidence function (CIF), and explores the underlying theoretical concept of the improper distribution of cause-specific failure times. A common but incorrect approach to estimating the probability of event type k is to treat all other event types as censored and calculate the complement of the Kaplan-Meier survival estimate for cause k, i.e., $1 - S^k(t)$. This method is fundamentally flawed because it operates in a hypothetical world where competing risks do not exist. By treating individuals who experience a competing event as censored, it violates the non-informative censoring assumption; these individuals are no longer at risk for the event of interest, a fact that is not true for those who are genuinely censored (e.g., lost to follow-up). This misspecification leads to a systematic overestimation of the true event probability. The correct quantity for estimating the absolute risk of a specific event in the presence of competing risks is the cause-specific cumulative incidence function (CIF), also known as the subdistribution function. The CIF for cause $k$, denoted $F_k(t)$ or $I_k(t)$, is formally defined as the joint probability that an event has occurred by time $t$ *and* that the event is of type $k$.

$$F_k(t) = P(T \leq t, J = k) \tag{4.7}$$

This function properly accounts for the fact that an individual must survive all other potential causes of failure up to a given time to be at risk for failure from cause $k$ at that time.

The cause-specific hazard function for event type $k$, denoted $h_k(t)$, is the instantaneous rate of failure from cause $k$ at time $t$, given that the individual has not experienced *any* event of any type up to time $t$:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, J = k \mid T \geq t)}{\Delta t} \tag{4.8}$$

This definition highlights that the conditioning event is survival from all causes, $T \geq t$. In a discrete time setting with event times $t_j$, the cause-specific hazard is the conditional probability:

$$h_k(t_j) = P(T = t_j, J = k \mid T \geq t_j) \tag{4.9}$$

This is the probability of failing from cause $k$ in interval $j$, given survival to the start of interval $j$. The risk set for estimating $h_k(t)$ at a given time $t$ is composed of all individuals who are event-free and uncensored just prior to time $t$. Critically, if an individual experiences an event of type $j = k$ at some time $t^* < t$, they are removed from the risk set for all subsequent times $t > t^*$ for all causes.

The sub-distribution hazard function is defined as the hazard rate corresponding to the CIF, $F_k(t)$, by treating it as if it were a proper survival function, $1 - F_k(t)$. The SDH for cause $k$, denoted $\lambda_k(t)$, is defined as:

$$\lambda_k(t) = -\frac{d}{dt} \log(1 - F_k(t)) = \frac{f_k(t)}{1 - F_k(t)} \tag{4.10}$$

where $f_k(t) = F_k'(t)$ is the cause-specific density. The conditioning set in the probabilistic definition is what makes this approach unique:

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, J = k \mid T \geq t \text{ or } (T < t \text{ and } J \neq k))}{\Delta t} \tag{4.11}$$

In a discrete time setting with event times $t_j$, the sub-distribution hazard is the conditional probability:

$$\lambda_k(t_j) = P(T = t_j, J = k \mid T \geq t_j \text{ or } T < t_j \text{ and } J \neq k) \tag{4.12}$$

The defining feature of the SDH is its risk set. At time $t$, the risk set for the SDH of cause $k$ includes all individuals who have not yet experienced an event of cause $k$. This means it includes those who are event-free *and*, counter-intuitively, those who have already experienced a competing event ($J = k$) at a time prior to $t$. While these latter individuals are no longer biologically at risk for any event, their inclusion in the risk set is the mathematical device that allows the resulting model to directly target the CIF.

The CIF is intrinsically linked to the underlying cause-specific hazard functions of all event types. To derive this relationship, we first define the cause-specific density for event $k$, $f_k(t)$, as the product of the cause-specific hazard for event $k$, $h_k(t)$, and the *overall* survival function, $S(t)$, which is the probability of surviving from all causes. The overall survival function is determined by the sum of all cause-specific hazards:

$$S(t) = \exp\left[-\sum_{j=1}^{K}\int_0^t h_j(u)du\right] \tag{4.13}$$

The CIF is then the integral of this cause-specific density:

$$F_k(t) = \int_0^t f_k(u)du = \int_0^t h_k(u)S(u)du \tag{4.14}$$

This integral formulation makes explicit that the cumulative incidence of cause $k$ depends not only on its own hazard, $h_k(u)$, but on the hazards of all competing causes through the $S(u)$ term. This is the mathematical embodiment of the loss of one-to-one correspondence discussed previously. The CIF provides a clinically meaningful interpretation of risk. The sum of the CIFs for all $K$ event types at time t gives the overall probability that any event has occurred by that time, which is the complement of the overall survival function: $\sum_{k=1}^{K} F_k(t) = 1 - S(t)$.

## 4.3 The potential outcome framework in survival analysis

The potential outcomes framework, often referred to as the Rubin Causal Model (RCM), provides such a structure. First proposed by Neyman in the context of randomized experiments and later generalized by Rubin to encompass observational studies (Rubin 1974) , this framework shifts the focus from modeling associations to defining and estimating causal effects.

For each individual unit $i$ (e.g., a patient) and for each level of a treatment or exposure $a$ (e.g., $a = 1$ for an active drug, $a = 0$ for a placebo), there exists a potential outcome, denoted $T_i(a)$. This value represents the outcome that would have been observed for individual $i$ had they received treatment level $a$.

The individual causal effect ($\tau_i$) is then defined as a contrast between two or more potential outcomes for the same individual. In the survival context

$$\tau_i = T_i(1) - T_i(0) \tag{4.15}$$

For any given individual, we can only observe one of their potential outcomes—the one corresponding to the treatment they actually received. The potential outcome under the treatment not received is unobservable and is therefore counterfactual. This is known as the fundamental problem of causal inference.

Because individual causal effects are inherently unobservable, the focus of causal inference shifts from the individual to the population level. The goal becomes the estimation of average causal effects (ATE), such as the average difference in outcomes if the entire population were treated versus if the entire population were not treated:

$$ATE = E[T(1)] - E[T(0)] \tag{4.16}$$

Estimating average causal effects from observed data requires a set of crucial, untestable assumptions known as identifiability conditions. These assumptions form the logical bridge that allows us to use observable associations to make inferences about unobservable causal effects.

- *Consistency*: This assumption links the potential outcomes to the observed data. It states that the potential outcome for an individual under the treatment they actually received is the outcome we observe for that individual. If an individual $i$ receives treatment $A_i = a$, then their observed outcome is $T_i = T_i(a)$. This assumption requires that the specified intervention is well-defined and that there are no hidden or different versions of the treatment that could lead to different outcomes.
- *Exchangeability* (or *Ignorability*): This assumption requires that the treatment assignment is independent of the potential outcomes, formally $(T(1), T(0)) \perp A$. In a perfectly conducted randomized controlled trial, exchangeability is expected to hold by design. In observational studies, where treatment is not randomly assigned, this assumption is rarely plausible. Instead, we rely on conditional exchangeability, which states that treatment assignment is independent of potential outcomes *within strata of measured covariates* $L$: $(T(1), T(0)) \perp A|L$. This assumes that we have measured a sufficient set of covariates $L$ to control for all common causes of treatment and outcome (i.e., no unmeasured confounding).
- *Positivity* (or Overlap): This assumption requires that for every combination of covariates $L = l$ present in the population, there is a non-zero probability of receiving each level of the treatment: $Pr(A = a|L = l) > 0$ for all relevant values of $a$ and $l$. This ensures that we have data on both treated and untreated individuals across the confounder space, making comparisons possible.
- *Stable Unit Treatment Value Assumption (SUTVA)*: This is a broader assumption that encompasses two components: the consistency assumption described above, and the assumption of no interference. No interference means that one individual's treatment assignment does not affect the potential outcomes of any other individual.

With this foundation, the counterfactual survival function for a specific treatment $a$, denoted $S_a(t)$, is defined as the survival function of the corresponding potential event time:

$$S_a(t) = Pr(T(a) > t) \tag{4.17}$$

This estimand has a precise causal interpretation: it is the proportion of the population that would have survived beyond time $t$ if, possibly contrary to fact, every individual in that population had received treatment $a$. It describes the survival experience in a hypothetical world defined by a universal intervention.

The other descriptive functions follow directly from the definition of the counterfactual survival function, preserving the mathematical structure established in the classical setting. The counterfactual cumulative distribution function, $F_a(t)$, is the probability of experiencing the event by time $t$ in the world where everyone receives treatment $a$:

$$F_a(t) = Pr(T(a) \leq t) = 1 - S_a(t) \tag{4.18}$$

Similarly, the counterfactual probability density function, $f_a(t)$, represents the density of event times in this same hypothetical world:

$$f_a(t) = \frac{dF_a(t)}{dt} = -\frac{dS_a(t)}{dt} \tag{4.19}$$

These functions collectively describe the full distribution of potential event times under a specific intervention, maintaining the internal consistency of the classical framework within each counterfactual world.

The counterfactual hazard function, $h_a(t)$, is defined analogously to its classical counterpart, using the counterfactual density and survival functions:

$$h_a(t) = \frac{f_a(t)}{S_a(t)} \tag{4.20}$$

Its interpretation, however, is substantially more nuanced and reveals a core complexity in causal survival analysis. The counterfactual hazard $h_a(t)$ represents the instantaneous event rate at time $t$ among the sub-population of individuals who *would have survived to time $t$ if they had all been assigned to treatment $a$*. This definition is critical because the conditioning event, $T(a) \geq t$, is itself a post-intervention outcome. Consequently, a comparison between $h_1(t)$ and $h_0(t)$ is not a comparison of event rates in two fixed populations. Instead, it is a comparison of rates in two different populations of survivors at time $t$, whose compositions have been dynamically shaped by the very treatments being evaluated. For example, if an effective treatment prevents early deaths among frail individuals, the population of survivors at a later time point in the treated world will contain these frail individuals, whereas they would be absent from the survivor population in the untreated world. This treatment-induced selection into the risk sets over time is the reason that the hazard ratio, even from a randomized trial, does not have a straightforward causal interpretation as an individual-level effect and can be misleading if interpreted as a constant measure of biological efficacy.

## 4.4 Flexible models for the estimation of cumulative quantities

In the present thesis, we will use flexible models of the form

$$g(\cdot|X) = \beta X + \sum_{k=1}^{K} (\gamma_k + \delta_k X) B_k(\cdot) + \epsilon \tag{4.21}$$

The formulation stays completely within the generalized linear model framework (Aitkin, Francis, and Hinde 2005). Here $g$ is a link function that link the response variable represented by $\cdot$ to a linear combination of parameters. Here the response variable is either the discrete hazard function (Equation 4.6), the hazard function (Equation 4.5), or the cumulative distribution function (Equation 4.3). In the easiest example with just one covariate, $\beta$ represents the effect associated to the exposure $X$, $\gamma_k$ represent the parameters of the splines that specify the baseline function of the response variable, whereas $\delta_k$ are the parameters associated to the interaction terms of the baseline function with the covariate term, used to relax the proportionality assumption (see the following sections).

When the focus is on the CDF in case of censored data, pseudo-observations, as proposed by (Klein and Andersen 2005), are a great solution since they somehow remove the problem of censoring.

Pseudo-observations are derived using a jackknife (leave-one-out) statistical procedure. For a specific quantity of interest, $\theta(t)$, such as the survival probability at a fixed time $\tau$, $\theta(\tau) = S(\tau)$, the process is as follows:

1. First, compute the non-parametric estimate $\hat{\theta}(\tau)$ using the full dataset of $n$ subjects. For the survival function, this would be the Kaplan-Meier estimate, $\hat{S}_{KM}(\tau)$.
2. Next, for each subject $i = 1, \dots, n$, temporarily remove that subject from the dataset and re-calculate the estimate on the remaining $n-1$ subjects, yielding $\hat{\theta}_{-i}(\tau)$.
3. The pseudo-observation for subject $i$ at time $\tau$ is then defined as:

$$\hat{\theta}_i(\tau) = n \cdot \hat{\theta}(\tau) - (n-1) \cdot \hat{\theta}_{-i}(\tau) \tag{4.22}$$

In the absence of censoring, $\hat{\theta}_i(\tau)$ would simply be the observed outcome for subject $i$ (e.g., the indicator variable $I(T_i > \tau)$. The pseudo-observations thus serve as a complete-data substitute for the potentially unobserved outcome in the presence of censoring. These generated values can then be used as the response variable in a Generalized Linear Model (GLM) or, more robustly, a Generalized Estimating Equation (GEE) framework to estimate the effect of covariates. This allows for direct modeling of the survival probability at one or more time points without any proportional hazards assumption.

When the response variable is the discrete hazard Equation 4.6, the model is usually called discrete-time hazard model, specifically designed for situations where time is measured in discrete intervals (e.g., years, months, treatment cycles). Classically, the effect of a covariate $X$ on the hazard of $T$ is modeled in a proportion continuation ratio model for an ordinal outcome $T$ with $K$ categories, where the effect of $X$ is expressed on the log-(conditional) odds scale. This model belongs to the GLM family and is expressed as below:

$$\begin{cases} Y \sim \text{Multinom}(t_1, \dots, t_K; p_1, \dots, p_K) \\ \log\left(\frac{h(t_k|\mathbf{X})}{1-h(t_k|\mathbf{X})}\right) = \log\left(\frac{P(T=t_k|Y \geq t_k, \mathbf{X})}{1-P(T=t_k|Y \geq t_k, \mathbf{X})}\right) = \alpha_j + \beta\mathbf{X} \end{cases} \tag{4.23}$$

relationship involves the unknown parameters $\beta$ and $\alpha_j, (k = 1, ..., K)$ that are interpreted as follows.

- $\beta$ represents the change in the log- (conditional) odds of $T$ occurring in a specific interval (e.g., $T = t_k$) given that it has not occurred it in the categories before ($T \geq t_k$), for a one-unit increase in $X$.
- $\alpha_j$ represents the baseline log-odds

Continuation ratio model can be considered as binomial model on each interval $t_k$ of the variable $T$, thus an extended model matrix $\mathbf{M}$ consists of the interval $t_k$ of the variable $T$, individual covariate values $\mathbf{X}$, and an indicator variable that defines whether the value of $T$ for subject $i$ is realized in the interval $t_k$. For each subject, the first three of the variables

above may assume distinct values in different intervals, while the covariates are repeated for each interval in which the subject has not realized its exposure covariate value.

While for modeling pseudo-observations the family of the error distribution is *Gaussian* (depending on the status of the subject they assume values greater or lower than 0 or 1), the family for the discrete-time hazard model is *Binomial*. Thus the canonical link function of the pseudo-observation is the *identity*, whereas for the discrete time hazard model the *logit* is the canonical one. Nevertheless different link functions within the generalized linear model can be exploited to represent the response variable or the linear predictor in different ways (Ambrogi, Biganzoli, and Boracchi 2008; Tutz and Schmid 2016).

When the focus is the continuous hazard, the piecewise exponential model is a direct extension of the exponential model that relaxes the assumption of a constant hazard over the entire follow-up period. It achieves this by partitioning the time axis into a series of $K$ prespecified intervals, $(t_{k-1}, t_k]$, and assuming a different, but constant, baseline hazard within each interval. The model for an individual in interval $k$ is:

$$h(t|\mathbf{X}) = h_k \exp(\mathbf{X}\beta) \quad \text{for } t \in (t_{k-1}, t_k] \tag{4.24}$$

specification of the baseline hazard is semi-parametric, offering more flexibility than a single parametric form (like Weibull) but less than the completely non-parametric baseline of the Cox model. The primary advantage of the PEM is its natural ability to accommodate time-varying effects (non-proportional hazards). This is achieved by allowing the regression coefficients $\beta$ to differ across time intervals, yielding a model of the form

$$h(t|\mathbf{X}) = h_k \exp(\mathbf{X}\beta_k) \tag{4.25}$$

This explicitly models how a covariate's effect changes over discrete periods of follow-up.The PEM assumes the hazard is constant within each interval and can be fitted within a Poisson log-linear model. In an augmented model matrix, the failure indicators $\delta_{kj}$ are treated as independent Poisson observations with means $\mu_{kj} = \tau_{kj}\lambda_{kj}$, where $\lambda_{kj}$ is the hazard rate in interval $k$ for individual $j$. Assuming proportional hazards, this relationship is modeled as $\lambda_{kj} = \lambda_k \exp(X_j'\beta)$, where $\lambda_k$ is the baseline hazard in interval $k$. This leads to the log-linear model:

$$\log(\mu_{kj}) = \log(\tau_{kj}) + \log(\lambda_k) + X_j'\beta \tag{4.26}$$

formulation is highly flexible, as the baseline hazard component, $\log(\lambda_k)$, can be modeled smoothly, and time-dependent covariate effects can be included as interactions between covariates and the terms for the time indicators.

## 4.5 Flexible Baseline Specification

Instead of leaving the baseline hazard $h_0(t)$ completely unspecified (as in the Cox model) or assuming a parametric form (as in the Weibull model), flexible parametric models, use splines to flexibly model a transformation of the baseline hazard or survival or cumulative distribution function. A common and robust approach is to model the baseline hazard or baseline cumulative distribution function, as a a function of (log, logit, cloglog) time:

$\log(H_0(t)) = s(\log(t)|\gamma, \mathbf{k})$ where $s(\cdot)$ is a spline function with coefficients $\gamma$ and a set of knots $\mathbf{k}$ placed at quantiles of the observed event times. This formulation creates a fully parametric model that is highly flexible and can capture a wide variety of hazard shapes, including monotonic, unimodal, or more complex patterns. The resulting model provides smooth estimates of the baseline hazard and survival functions, which facilitates clinically useful outputs such as direct prediction of survival probabilities and robust extrapolation beyond the observed follow-up period.

The primary mechanism for modeling non-proportional hazards (i.e., time-varying covariate effects) with splines is through the use of interaction terms between the covariate and a spline function of time. The term $B_k(t)$ in Equation 4.21 are the basis functions of a spline of time (e.g., B-splines or restricted cubic splines).

A relevant aspect of spline fitting is the selection of the basis function, and the selection of the position and the number of knots. Usually, the number and the position of the knots is specified considering previous knowledge in the field. When this is not possible, the number of the knots is directly inserted into the model selection algorithm, while the position is pre-specified at quantiles of the distribution of the event-times of the primary outcome. In the present thesis we use B-splines or natural splines (B-splines with linearity constraints at the tail of the distribution) for modeling instantaneous quantities, whereas restricted cubic splines for modeling cumulative quantities.

B-splines are defined by a set of basis functions. The degree of the B-spline (often denoted as $p$) determines the order of these basis functions. The most used B-splines are cubic (degree $p = 3$). ). A B-spline curve is determined through a unique vector of time-points $t$**,** called knots, which identify the function domains of the polynomial segments. Essentially, the dimension of the knot vector determines the flexibility of the spline. A higher number of knots lead to higher flexibility of the curve.

A B-spline function $B(t)$ of degree $p$ with $N_{kn}$ interior knots is a linear combination of the $K + p$ basis functions:

$$B(t) = \sum_{j=-p}^{N_{kn}} \alpha_j B_{j,p}(t), \;\; t \in (t_0, t_{K-1}) \tag{4.27}$$

B-splines have the property of local support. Considering a knot $t_k$ each basis spans on $t_{k-2}$ to $t_{k+2}$. For each datapoint the sum of the basis functions is equal to 1.

Restricted cubic splines consist in truncated power basis function. Basis functions are defined as follows:

$$B_1(x) = 1, B_2(x) = x, ..., B_{d+1}(x) = x^d B_{d+2}(x) = (x - t_1)_+^d, ..., B_{K+d+1} = (x - t_k)_+^d \tag{4.28}$$

The restricted cubic splines have linearity constraints before the first and after the last knot of the knot vector $K$. The constraints are appealing in the specification of the baseline hazard/risk function, as they mitigate possible non-monotonic estimation of the functions

towards the end of the follow-up. Due to the constraints applied the number of the bases are $K-1$.

## 4.6 Flexibly modeling cause specific cumulative incidence function

Flexible model for the cause specific cumulative incidence function in the competing risk context build upon the model described before. The first approach build upon (Berger et al. 2018) and it is based on estimating the discrete time SDH Equation 4.12 of an event considering a binomial estimation scheme like in Equation 4.23 and a weighted maximum likelihood function.

The individual form of the weighted maximum likelihood function is:

$$L_i = \prod_{t=1}^{K-1} \{\lambda_r(t|x_i)^{y_{it}}(1-\lambda_r(t|x_i))^{1-y_{it}}\}^{w_{it}} \tag{4.29}$$

where $y_{it}$ values correspond to:

$$(y_{i1},...,y_{y,\tilde{T}_i},...,y_{i,k-1}) = \begin{cases} (0,...,0,1,0,...,0),\ if\ \Delta_i = 1 \\ (0,...,0,0,0,...,0),\ if \Delta_i = 0 \end{cases}, \tag{4.30}$$

and $w_{it}$ is the probability that within the interval t the i-th subject is still at risk of manifesting the event of interest. More specifically, by denoting $\tilde{t_i}$ the generic value of the variable $\tilde{T}_i$, the weights$(w_{i1}, w_{i2}, ..., w_{i\tilde{t_i}}, w_{i,(\tilde{t_i}+1)}, ..., w_{i,(K-1)}$ are defined as $(1,1,...,1,0,...,0)$ for subjects who experience the focus event (formally: if $\Delta_i = 1$ and $\epsilon_i = 1$). A similar formula holds for censored subjects, but the term $\tilde{t_i}$ should be substitute by the observed value of $C_i$. For subjects who experience a competing event, the value of $w_{it}$ in the intervals following $\tilde{t_i}$ is not 0 but depends on the probability of censoring:

$$w_{it} := \frac{\hat{S}_c(t-1)}{\hat{S}_c(\tilde{t_i}-1)}, \tilde{t_i} < t \leq K-1, \tag{4.31}$$

where $S_c()$ represent the survival function of $C$. Since these values are not observable, a solution is to substitute them with estimates obtained using the inverse probability of censoring weighting (IPCW) method, as done by Fine and Gray (1999) in the continuous-time framework.

Equivalently to the direct approach with pseudo-observations for the cumulative incidence, this second approach is based on the original transformation model using pseudo-observations computed for each subject at $P$ selected time points $(\tau_1,...,\tau_P)$ within the follow-up period considering however the non-parametric Aalen-Johansen estimates of the CIF.

# 5 Number of events per parameter needed to recover measure of average treatment effect in flexible regression survival analysis

As introduced in Section 3.6, the reliability of survival models depends on the amount of information in a dataset, which is best measured by the number of observed events. A robust metric for this is Events Per Parameter (EPP), which more accurately reflects a model's complexity than the simpler "Events Per Variable" rule. While a minimum of 10 EPP is a common guideline for standard Cox models, the highly flexible parametric models used in modern causal survival analysis are far more complex. These models require additional parameters to model the baseline hazard and time-dependent effects (Section 4.5), yet the EPP needed for their stable application is largely unknown. This chapter addresses this critical gap through a simulation study. We investigate the EPP requirements for estimating Average Treatment Effects (ATEs; see Section 3.4) from flexible models, evaluating their performance across various scenarios of increasing complexity. The focus of this study will be the indirect approaches for estimating cumulative quantities, which are commonly used for this purpose.

## 5.1 Methods

### 5.1.1 Simulation Study Design

The overall objective of the simulation study is to evaluate the performance of the discrete-time hazard and the piece-wise exponential models ( Section 4.4 ) in estimating standardized cumulative probability functions from which measures of ATE are computed. The performance is evaluated as a function of the number of events per parameter available, in addition to the complexity of the baseline hazard function underlying time-to-event distribution of a randomized clinical study.

To such end, we adopt a large scale Monte Carlo simulation study involving five survival scenarios with increasing level of complexity, either in terms of the baseline hazard function to smooth, or of the relationship between the covariates (interaction effects, time-dependent effects). For each scenario, we evaluated the performances of the models when 5, 10, 15, 20, 30, 40, 50 events per parameter were available.

### 5.1.2 Data-Generating Mechanisms (DGMs)

We will now describe the data-generating mechanism underlying each scenario adopted in the simulation. These ranged from standard proportional hazards (PH) models with simple or complex baseline hazards to scenarios with non-proportional hazards (NPH).

This being an initial investigation, we decided to limit our considerations where exposure ($C$) is completely independent from the covariates exchangeability ( Section 4.3 ) is automatically guaranteed by simulation.

#### 5.1.2.1 Scenarios 1-2-3: Proportional Hazards (PH) with Varying Baseline Hazards

The first three scenarios establish a baseline for model performance under the standard proportional hazards assumption. The general form of the log-hazard ($log(h(t))$) for a subject $i$ at time $t$ is:

$$log(h_i(t)) = \log(h_0(t)) + \beta_N N_i + \beta_C C_i + \beta_{TD}(TD_i - 0.1) \tag{5.1}$$

where the covariate effects are fixed at $\beta_N = 0.53$, $\beta_C = -0.4$, and $\beta_{TD} = 0.4$. The scenarios differ only in the baseline hazard, $h_0(t)$. For each subject, we simulate the binary covariate, $N$, from a Bernoulli distribution with $p = 0.65$; the binary exposure covariate, $C$, from a Bernoulli distribution with $p = 0.50$; and the continuous covariate, $TD$ (tumor dimension), from a Normal distribution $N(\mu = 1.5, \sigma = 0.25)$, truncated to the interval $[0, 2.5]$.

Namely, for Scenario 1, the baseline hazard is from a Weibull distribution with shape $\rho = 1$ and scale $\lambda = 160$. This is equivalent to an Exponential distribution with a constant hazard rate of $1/160$. The motivation behind the adoption of this distribution is that it represents the simplest case of a constant baseline hazard function.

For Scenario 2, the baseline hazard is from a Weibull distribution with shape $\rho = 1.5$ and scale $\lambda = 150$, corresponding to a monotonically increasing hazard function. This scenario evaluates the modeling approaches against a common parametric form with a simple, increasing hazard trend.

For Scenario 3, the baseline hazard, $h_0(t)$, is a complex, non-monotonic mixture of two spline-defined hazard profiles, as described previously. This distribution was selected to test the flexible spline-based model's ability to accurately capture a highly irregular baseline hazard shape under an otherwise standard PH structure.

To generate a multimodal hazard function similar to the disease-free survival functions encountered in clinical studies of breast cancer (Retsky and Demicheli 2014), we defined two distinct, baseline hazard profiles, $h_{0,1}(t)$ and $h_{0,2}(t)$. Each profile was constructed using monotone Hermite cubic splines, which were fit to a series of pre-specified knots defining the hazard's values and its derivative at key time points. Then, through the relationship between the hazard, the survival and the PDF, we obtained $S_1$ and $S_2$ and $f_1$ and $f_2$. From these two

profiles, a primary baseline hazard, $h_0(t)$, was created as a two-component mixture model, where the mixture proportion was set to $p = 0.65$:

$$h_0(t) = \frac{p \cdot f_1(t) + (1-p) \cdot f_2(t)}{p \cdot S_1(t) + (1-p) \cdot S_2(t)}$$

#### 5.1.2.2 Scenario 4: Proportional Hazards with a Covariate Interaction

This scenario builds upon the complex PH model of Scenario 3 by introducing a product interaction term between the binary covariates $N$ and $C$. The log-hazard function is:

$$\log(h_i(t)) = \log(h_{mix}(t)) + \beta_N N_i + \beta_C C_i + \beta_{TD}(TD_i - 0.1) + \beta_{NC} N_i C_i \tag{5.2}$$

The interaction effect is set to $\beta_{NC} = -0.288$.

This scenario verifies the model's ability to estimate a standard interaction effect, where the hazard ratio for one covariate depends on the level of another.

#### 5.1.2.3 Scenarios 5: Non-Proportional Hazards (NPH)

This scenario was designed to present a significant challenge by incorporating multiple, simultaneous violations of the proportional hazards assumption. It departs from the semi-parametric hazard modeling framework and instead generated event times from a fully parametric log-logistic distribution. In this case, covariates act multiplicatively on the time scale, which translates to a non-proportional effect on the hazard scale. The event time $T$ follows a log-logistic distribution with a constant shape parameter $\rho = 2$ and a subject-specific scale parameter $a_i$. The scale parameter is determined by the covariates with the following equation:

$$\log(\alpha_i) = \log(70) + \gamma_N N_i + \gamma_C C_i + \gamma_{TD} TD_i \tag{5.3}$$

### 5.1.3 Simulation of Cohort Data

For the first two and the fifth scenarios, the inverse of the CDF are parametrically defined, thus event times are generated by applying the inverse cumulative distribution function for the subject-specific model. Instead, for scenarios 3-4 event times are generated using numerical integration and root finding techniques (Crowther and Lambert 2013; Marano et al. 2025) on a discrete time grid from $t = 1$ to $t = 120$, based on the subject-specific cumulative distribution function.

All generated event times are subject to administrative censoring at 120 months and random censoring from loss to follow-up, which is simulated from an exponential distribution with a rate of $\lambda = 0.01$. The final observed time is the minimum of the true event time, administrative censoring time, and random loss to follow-up time.

As is standard in simulation studies, it is necessary to specify all the parameters pertinent to the simulation. However, since the hazard function is specified with parametric distributions or interpolating splines, the specific number of parameters of the baseline are unknown and have to be empirically defined.

To ensure each simulation cell had the desired EPP, we followed a two-step process. First, for each of the five scenarios, the 'true' number of parameters ($P$) required to flexibly model the baseline hazard with B-splines was determined empirically. This was achieved by fitting a series of models with an increasing number of interior knots to a very large dataset ($N = 100,000$) and selecting the spline complexity that minimized the Akaike Information Criterion (AIC). This number of parameters ($P$) then served as the denominator for the EPP calculation. Second, for each target EPP level (e.g., EPP=20), the required number of events was calculated as $E = EPP \times P$. Finally, the total sample size ($N$) for each simulation replicate was determined by generating subjects until the target number of events ($E$) was observed, accounting for the known event and censoring rates specified in the DGM.

### 5.1.4 Model Specification

For both the DTH and PWE models, the baseline log-hazard function was modeled flexibly using cubic B-splines. A data-dependent procedure was employed to determine the complexity of the baseline spline function for each analysis. Specifically, the number of internal knots was selected from a range of 0 to 6, maintaining the maximum number of baseline parameters at 10, by fitting a series of models and choosing the one that minimized the Akaike Information Criterion (AIC). This ensures that the model complexity is adapted to the information available in the data, balancing goodness-of-fit with parsimony.

Continuous covariate effect of the prognostic covariate TD was modeled assuming a linear relationship with the log-hazard (or the relevant model parameter). For scenarios designed to have time-varying coefficients (scenarios 3, 4, and 5), the interaction between the relevant covariate(s) and time was also modeled flexibly using a cubic B-spline with one internal knot, allowing the effect of the covariate to change over time.

### 5.1.5 Target estimands of the average treatment effect

The primary estimands of interest were the marginal (population-averaged) measures already described in Section 3.4, including Risk Difference ( Equation 3.1 ), Relative Risk ( Equation 3.2 ) and Difference in Restricted Mean Survival Time ( Equation 3.3 ) . Figure 5.1 shows the true estimand of interest for each scenario considered.

### 5.1.6 Estimation Procedure

The marginal ATEs were estimated using the principle of standardization, also known as g-computation. After fitting either the DTH or PWE model to a simulated dataset, the following procedure was performed:

1. For each subject in the original dataset, two counterfactual survival trajectories were predicted from the fitted model. First, by setting their exposure status to "exposed" (`C=1`) while keeping their observed covariates fixed. Second, by setting their exposure status to "unexposed" (`C=0`), again holding other covariates at their observed values.
2. These individual counterfactual survival predictions were then averaged across all subjects in the dataset. This step yields the marginal survival and cumulative incidence curves for the exposed and unexposed populations.
3. Finally, the marginal estimands (RD, RR, RRD, and $\Delta$RMST) were calculated directly from these averaged counter factual curves. This procedure effectively estimates the treatment effect that would be observed had the entire population been exposed versus had the entire population been unexposed, averaged over the empirical distribution of the covariates.

### 5.1.7 Performance Metrics

The performance of each estimation strategy was evaluated across the $M = 1000$ Monte Carlo simulations using the following statistical metrics. Let $\theta$ be the true value of an estimand (e.g., RD at a specific time point) and $\hat{\theta}_m$ be its estimate from the $m$-th simulation at time $t$.

1. **Bias**: The average difference between the estimated ATE and the true ATE, indicating systematic error.
$$\text{Bias} = \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta) \tag{5.4}$$
2. **Empirical Standard Error (ESE)**: The standard deviation of the ATE estimates across all simulations, representing the sampling variability of the estimator.
$$\text{ESE} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\hat{\theta}})^2} \quad \text{where} \quad \bar{\hat{\theta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{5.5}$$
3. **Mean Absolute Error (MAE)**: The average absolute difference between the estimate and the true value, providing a measure of the average magnitude of the errors.
$$\text{MAE} = \frac{1}{M}\sum_{m=1}^{M}|\hat{\theta}_m - \theta| \tag{5.6}$$
4. **Mean Squared Error (MSE)**: A measure of the overall accuracy of the estimator, which decomposes into the sum of the squared bias and the variance ($\text{ESE}^2$).
$$\text{MSE} = \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta)^2 = (\text{Bias})^2 + (\text{ESE})^2 \tag{5.7}$$

5. **Root Mean Squared Error (MSE)**: The squared root of the MSE.

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta)^2} \tag{5.8}$$

For sake of simplicity, we will show the results of the performance for the measures evaluated at specific time-points of the follow-up (60 and 120 months).

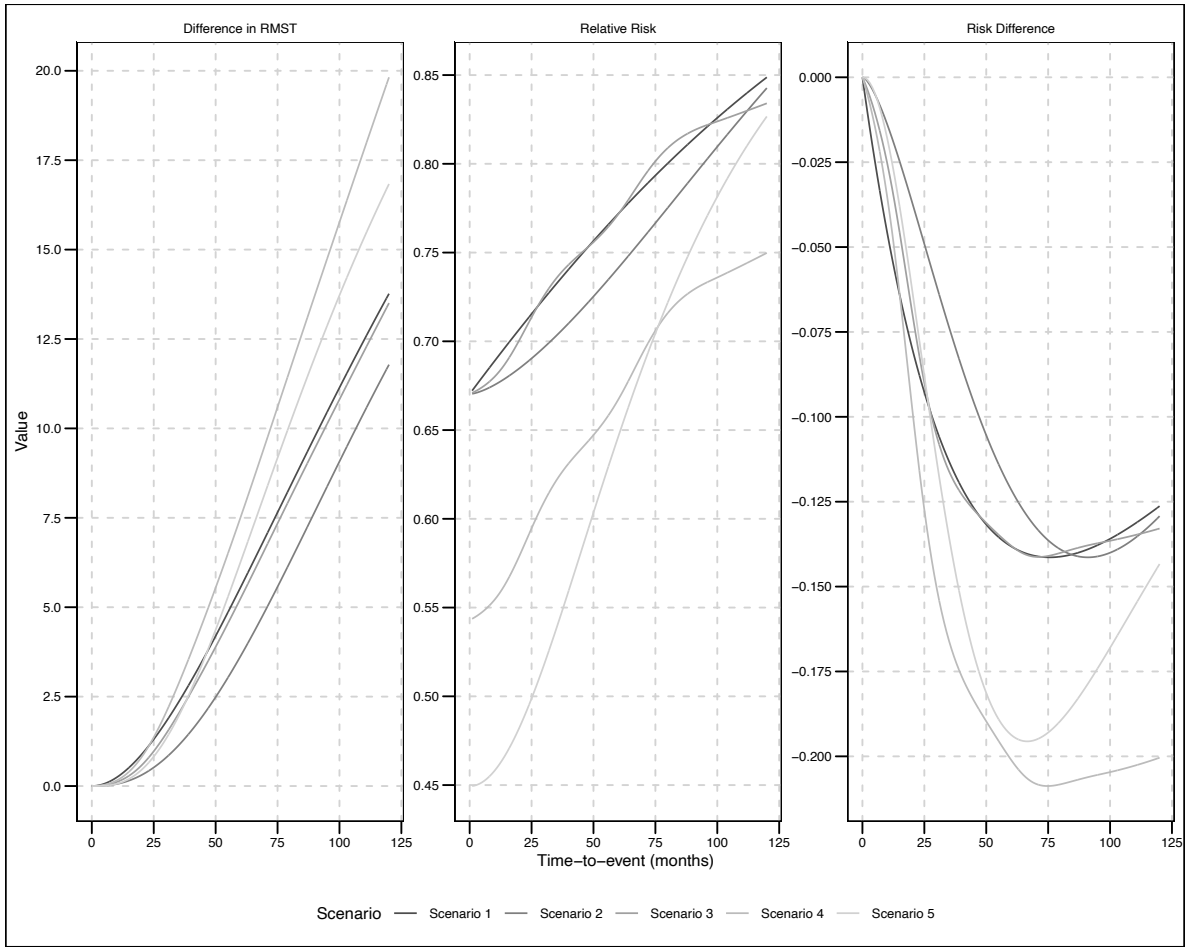



Figure 5.1: True measures of the average treatment effect for each scenario. These will be the target estimands in the simulation. From left to right panels, Difference in RMST, Relative Risk, Risk Difference.

## 5.2 Results

### 5.2.1 *Negligible Bias of ATE Estimators Across All Scenarios*

The first finding, consistent across all five simulation scenarios, both modeling approaches, and all estimands of interest, is that the estimators were found to be practically unbiased,

regardless of the number of event per parameter available. The analysis of the decomposition of the MSE reveals that the MSE of the ATE estimates was overwhelmingly driven by the variance component, as measured by the squared Empirical Standard Error.

Figure 5.2, Figure 5.3 and Figure 5.4 show that the bias of the estimators for the Risk Difference (RD), Risk Ratio (RR), and the difference in Restricted Mean Survival Time ($\Delta RMST$), respectively, is consistently negligible (values about 4 to 5 decimal places lower than the variance), approaching zero even at the lowest Events Per Parameter (EPP) level of 5. This fundamental result serves as a validation of the core modeling strategy. The use of flexible B-splines to parameterize the baseline hazard, coupled with a data-dependent selection of spline complexity via the Akaike Information Criterion (AIC), proves highly effective at capturing the true underlying data-generating mechanisms without introducing systematic error. This held true for simple, constant hazards (Scenario 1) as well as for complex, non-monotonic baseline hazards (Scenario 3) and severe violations of the proportional hazards assumption (Scenario 5).

The near-zero bias is a critical prerequisite for causal inference based on standardization (g-computation), as it confirms that the models can generate accurate counterfactual predictions upon which the ATE estimates are based. The absence of systematic error affirms the theoretical appropriateness of these flexible parametric models for the target estimands. Consequently, the central challenge identified by this study is not one of model misspecification or inherent bias, but rather one of statistical precision and estimator stability. The subsequent sections provide a detailed examination of how estimator variance, and by extension overall accuracy as measured by MSE and Mean Absolute Error (MAE), is critically dependent on the interplay between the available EPP and the complexity of the data-generating process.

### 5.2.2 *Impact of EPP on Estimator Variance and Precision*

Regardless of the complexity of the scenarios modeled and the estimand considered, the relationship between estimation error and EPP followed a distinct "L-shaped" curve, characterized by a steep reduction in both MSE and MAE as EPP increased from 5 to 15. Beyond an EPP of approximately 20, the performance gains from additional events per parameter became progressively smaller, indicating a point of diminishing returns in precision. This is visualized for Risk Difference, Relative Risk and Difference in RMST in Figure 5.5, Figure 5.6 and Figure 5.7, respectively.

In Table 5.1, Table 5.2 and Table 5.3 a comprehensive assessment of the performances of the flexible models is provided. An EPP of 5 or 10 proved insufficient for achieving reliable and stable ATE estimates. The high variance observed at these low EPP levels challenges the notion that the conventional 10 Events Per Variable (EPV) rule of thumb can be relaxed in simple settings. The inherent structure of spline-based models necessitates the estimation of multiple parameters to define the basis functions, even when the AIC selects a simple final form (e.g., zero internal knots). When the number of events is low, these parameters are estimated with high uncertainty. This uncertainty is then amplified through the non-linear standardization procedure used to calculate the marginal ATEs, resulting in high variance in the final estimands. This suggests that the very flexibility that makes these models powerful imposes a

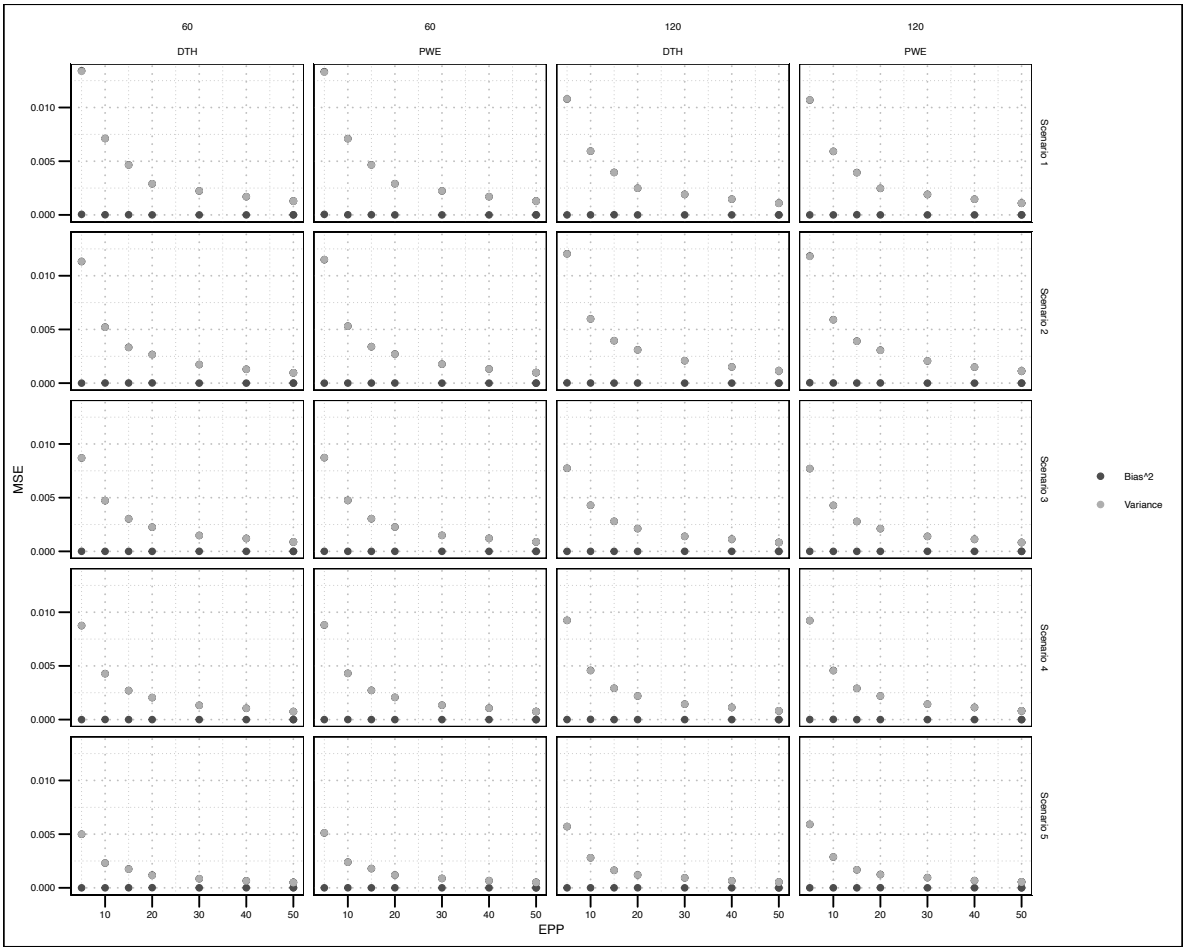



Figure 5.2: Bias-Variance decomposition of the Mean Squared Error (MSE) for the Risk Difference estimand. The total height represents the MSE, which is decomposed into its squared bias (darker shade) and variance (lighter shade) components.

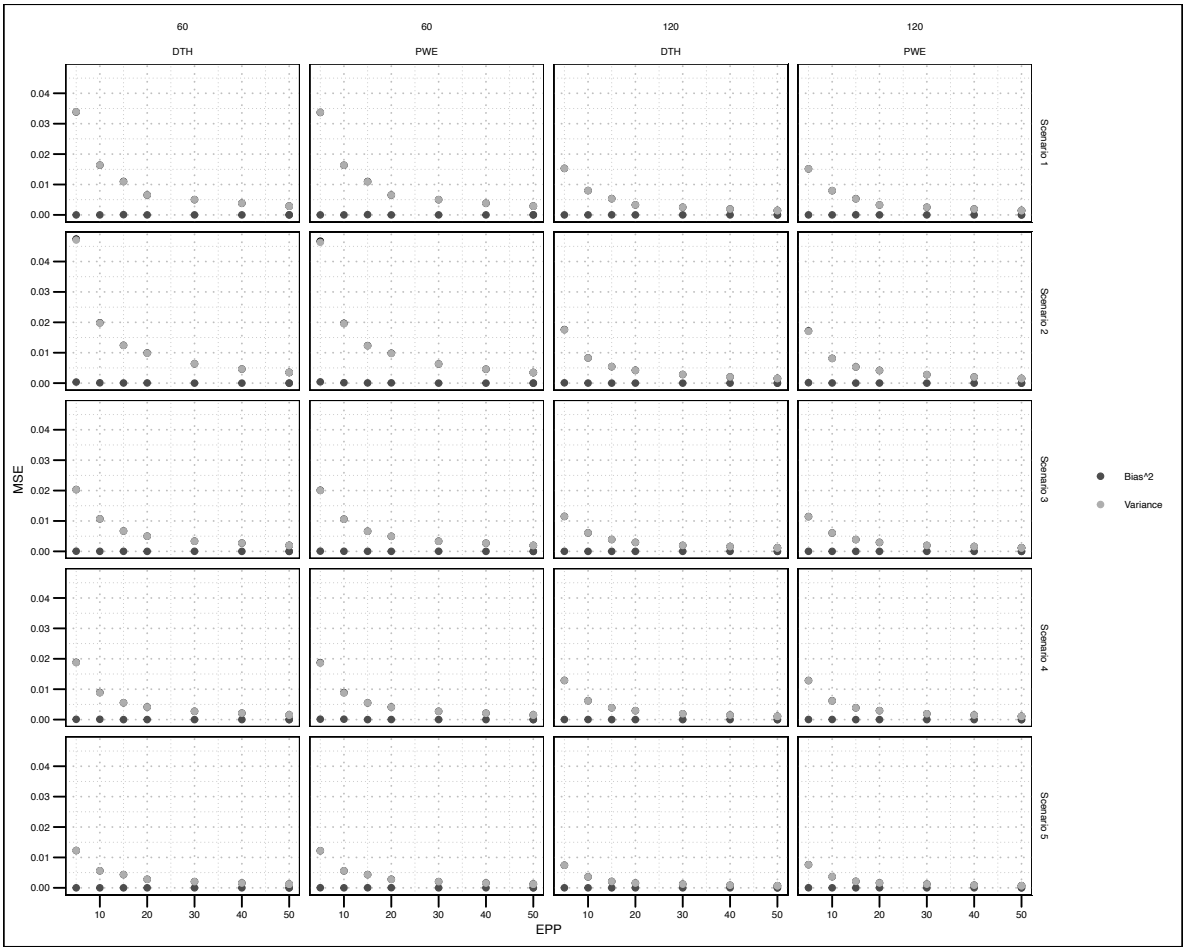



Figure 5.3: Bias-Variance decomposition of the Mean Squared Error (MSE) for the Relative Risk estimand. The total height represents the MSE, which is decomposed into its squared bias (darker shade) and variance (lighter shade) components.

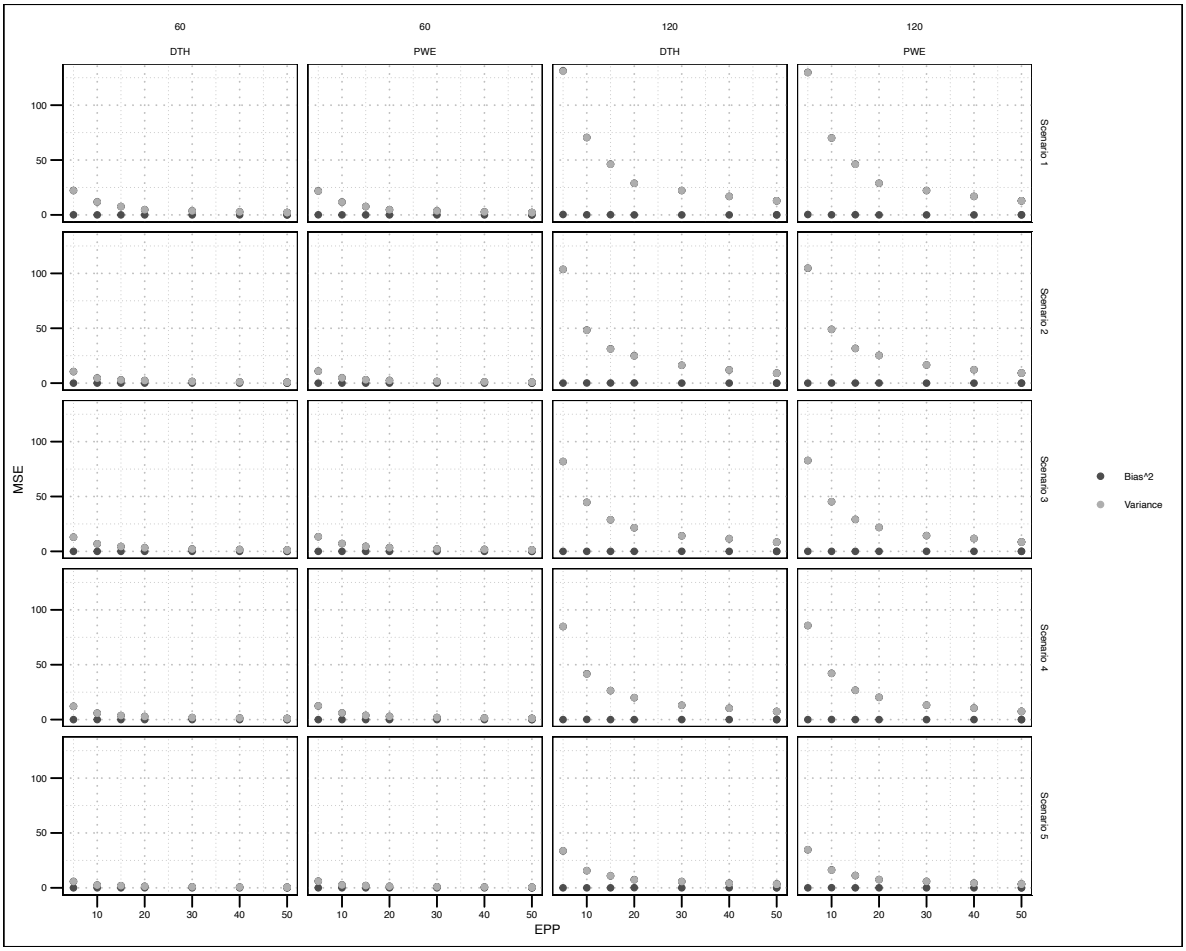



Figure 5.4: Bias-Variance decomposition of the Mean Squared Error (MSE) for the Difference in RMST. The total height represents the MSE, which is decomposed into its squared bias (darker shade) and variance (lighter shade) components.

higher minimum EPP requirement for stability compared to traditional semi-parametric Cox models, even when the underlying hazard structure is simple.

Although a higher EPP was expected to be required to achieve the same level of precision observed in the simpler Scenarios 1 and 2 was expected, for any given EPP level, the MSE and MAE in Scenarios 3 and 4 were consistently comparable to those of the preceding scenarios. For example, the MSE for the RD estimand at an EPP of 20 in Scenario 3 was comparable to the MSE observed at an EPP of 20 in Scenario 2.

In Scenario 3, the models successfully captured the irregular, multi-modal shape of the baseline hazard, as evidenced by the continued negligible bias in the ATE estimates. This highlights the strength of the AIC-driven B-spline approach in adapting to unknown and complex hazard functions. In Scenario 4, the inclusion of the interaction term added one additional parameter to the model's linear predictor. The results did not demonstrate a further increase in estimator variance compared to Scenario 3 at equivalent EPP levels.

The results of scenario 5 demonstrate the remarkable capacity of this flexible modeling approach to approximate the true NPH structure. As in the PH scenarios, the bias of the ATE estimators remained low and practically negligible across all EPP levels. This confirms that the interaction between covariates and a B-spline basis for time is an effective strategy for capturing complex time-dependent effects without introducing systematic error. Interestingly the RMSE and the MAE of the estimates associated to this scenario were even lower than those from simpler scenarios.

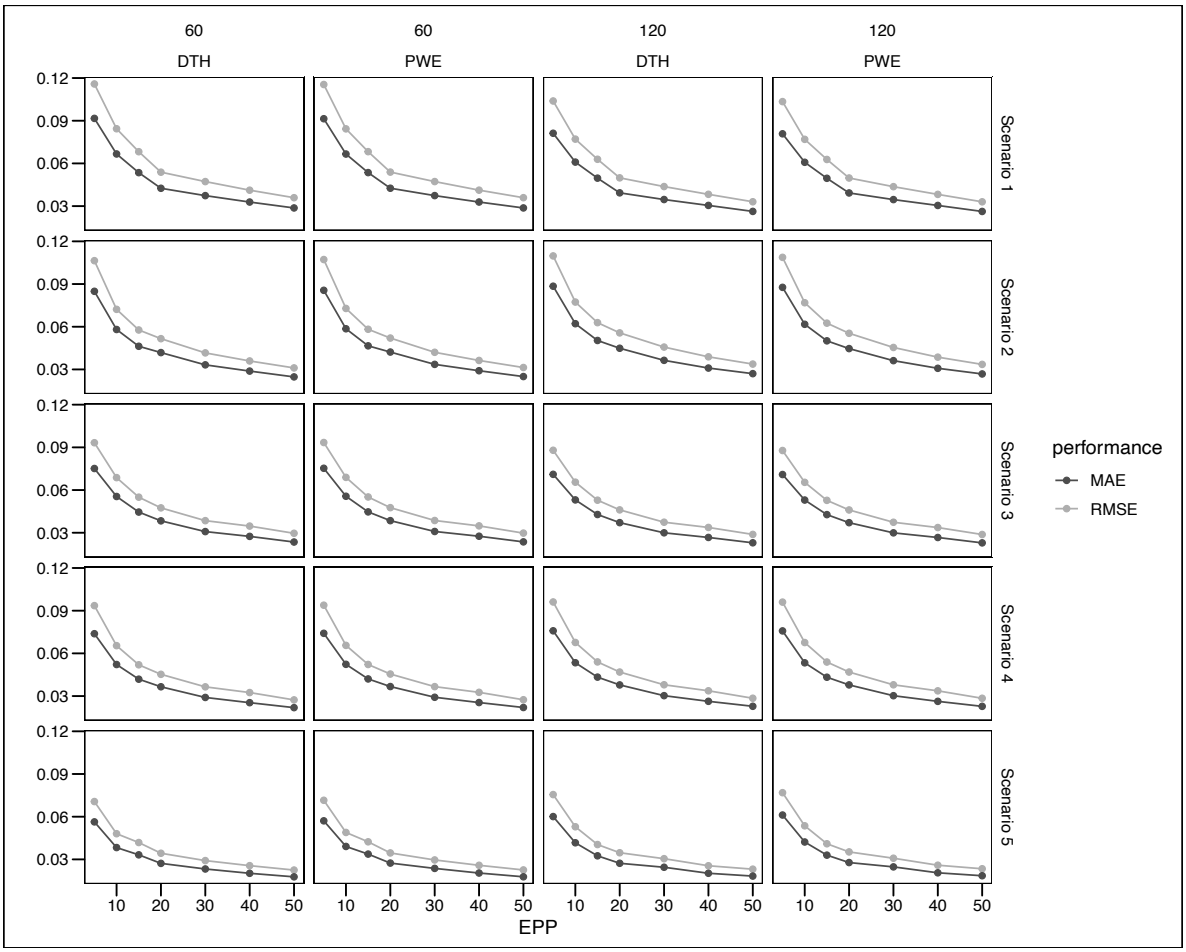



Figure 5.5: Results for the performance of the estimation of the Risk Difference ATE Measure as a function of the Event per Parameters at 60 and 120 months

Table 5.1: Model Performance by EPP for the Risk difference. Values shown are MAE / RMSE

| Scenario | Time | Truth | EPP5 | EPP10 | EPP15 | EPP20 | EPP30 | EPP40 | EPP50 |
|---|---|---|---|---|---|---|---|---|---|
| **DTH** | | | | | | | | | |
| Scenario 1 | 60 | -0.138 | 0.092/0.116 | 0.067/0.084 | 0.054/0.068 | 0.043/0.054 | 0.037/0.047 | 0.033/0.041 | 0.029/0.036 |
| Scenario 1 | 120 | -0.126 | 0.081/0.104 | 0.061/0.077 | 0.05/0.063 | 0.039/0.05 | 0.035/0.044 | 0.03/0.038 | 0.026/0.033 |
| Scenario 2 | 60 | -0.121 | 0.085/0.106 | 0.058/0.072 | 0.046/0.058 | 0.042/0.052 | 0.033/0.042 | 0.029/0.036 | 0.025/0.031 |
| Scenario 2 | 120 | -0.129 | 0.088/0.11 | 0.062/0.077 | 0.05/0.063 | 0.045/0.056 | 0.036/0.046 | 0.031/0.039 | 0.027/0.034 |
| Scenario 3 | 60 | -0.138 | 0.075/0.093 | 0.055/0.069 | 0.045/0.055 | 0.038/0.047 | 0.031/0.038 | 0.027/0.035 | 0.023/0.03 |
| Scenario 3 | 120 | -0.133 | 0.071/0.088 | 0.053/0.066 | 0.043/0.053 | 0.037/0.046 | 0.03/0.037 | 0.027/0.034 | 0.023/0.029 |
| Scenario 4 | 60 | -0.201 | 0.074/0.094 | 0.052/0.065 | 0.042/0.052 | 0.037/0.045 | 0.029/0.036 | 0.025/0.032 | 0.022/0.027 |
| Scenario 4 | 120 | -0.200 | 0.076/0.096 | 0.053/0.068 | 0.043/0.054 | 0.038/0.047 | 0.03/0.038 | 0.026/0.034 | 0.023/0.028 |
| Scenario 5 | 60 | -0.194 | 0.056/0.071 | 0.038/0.048 | 0.033/0.042 | 0.027/0.034 | 0.023/0.029 | 0.02/0.026 | 0.018/0.023 |
| Scenario 5 | 120 | -0.143 | 0.06/0.076 | 0.042/0.053 | 0.032/0.04 | 0.027/0.035 | 0.024/0.03 | 0.02/0.026 | 0.018/0.023 |
| **PWE** | | | | | | | | | |
| Scenario 1 | 60 | -0.138 | 0.091/0.115 | 0.067/0.084 | 0.054/0.068 | 0.043/0.054 | 0.037/0.047 | 0.033/0.041 | 0.029/0.036 |
| Scenario 1 | 120 | -0.126 | 0.081/0.103 | 0.061/0.077 | 0.05/0.063 | 0.039/0.05 | 0.035/0.044 | 0.03/0.038 | 0.026/0.033 |
| Scenario 2 | 60 | -0.121 | 0.086/0.107 | 0.059/0.073 | 0.047/0.058 | 0.042/0.052 | 0.034/0.042 | 0.029/0.036 | 0.025/0.031 |
| Scenario 2 | 120 | -0.129 | 0.088/0.109 | 0.062/0.077 | 0.05/0.063 | 0.045/0.055 | 0.036/0.045 | 0.031/0.039 | 0.027/0.034 |
| Scenario 3 | 60 | -0.138 | 0.075/0.093 | 0.056/0.069 | 0.045/0.055 | 0.039/0.048 | 0.031/0.039 | 0.028/0.035 | 0.024/0.03 |
| Scenario 3 | 120 | -0.133 | 0.071/0.088 | 0.053/0.065 | 0.043/0.053 | 0.037/0.046 | 0.03/0.037 | 0.027/0.034 | 0.023/0.029 |
| Scenario 4 | 60 | -0.201 | 0.074/0.094 | 0.052/0.066 | 0.042/0.052 | 0.037/0.045 | 0.029/0.037 | 0.025/0.033 | 0.022/0.027 |
| Scenario 4 | 120 | -0.200 | 0.076/0.096 | 0.053/0.068 | 0.043/0.054 | 0.038/0.047 | 0.03/0.038 | 0.026/0.034 | 0.023/0.028 |
| Scenario 5 | 60 | -0.194 | 0.057/0.072 | 0.039/0.049 | 0.034/0.042 | 0.027/0.035 | 0.024/0.03 | 0.02/0.026 | 0.018/0.023 |
| Scenario 5 | 120 | -0.143 | 0.061/0.077 | 0.042/0.054 | 0.033/0.041 | 0.028/0.035 | 0.025/0.031 | 0.021/0.026 | 0.019/0.023 |

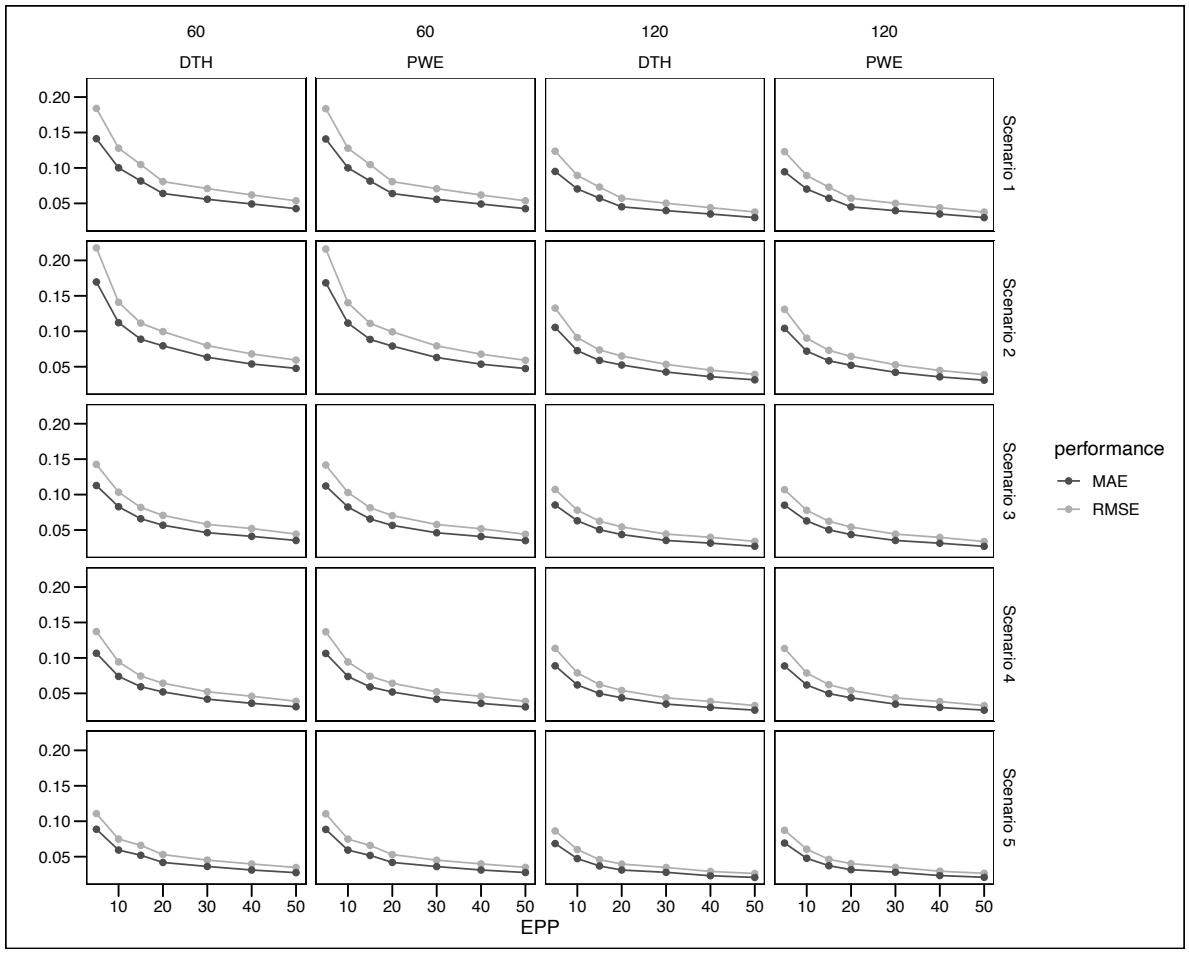



Figure 5.6: Results for the performance of the estimation of the Relative Risk ATE Measure as a function of the Event per Parameters at 60 and 120 months

Table 5.2: Model Performance by EPP for the Relative Risk. Values shown are MAE / RMSE

| Scenario | Time | Truth | EPP5 | EPP10 | EPP15 | EPP20 | EPP30 | EPP40 | EPP50 |
|---|---|---|---|---|---|---|---|---|---|
| **DTH** | | | | | | | | | |
| Scenario 1 | 60 | 0.772 | 0.141/0.184 | 0.1/0.128 | 0.082/0.105 | 0.064/0.081 | 0.056/0.071 | 0.049/0.062 | 0.043/0.054 |
| Scenario 1 | 120 | 0.849 | 0.095/0.124 | 0.07/0.089 | 0.057/0.073 | 0.045/0.057 | 0.04/0.05 | 0.035/0.044 | 0.03/0.038 |
| Scenario 2 | 60 | 0.741 | 0.17/0.218 | 0.112/0.141 | 0.089/0.112 | 0.079/0.1 | 0.063/0.08 | 0.054/0.068 | 0.048/0.059 |
| Scenario 2 | 120 | 0.843 | 0.105/0.133 | 0.073/0.091 | 0.059/0.074 | 0.052/0.065 | 0.043/0.053 | 0.036/0.045 | 0.031/0.039 |
| Scenario 3 | 60 | 0.772 | 0.113/0.143 | 0.083/0.103 | 0.066/0.082 | 0.057/0.071 | 0.046/0.058 | 0.041/0.052 | 0.035/0.044 |
| Scenario 3 | 120 | 0.834 | 0.085/0.107 | 0.063/0.078 | 0.05/0.062 | 0.044/0.054 | 0.035/0.044 | 0.031/0.04 | 0.027/0.034 |
| Scenario 4 | 60 | 0.667 | 0.107/0.137 | 0.074/0.094 | 0.059/0.074 | 0.052/0.064 | 0.042/0.052 | 0.036/0.046 | 0.031/0.039 |
| Scenario 4 | 120 | 0.750 | 0.089/0.114 | 0.062/0.079 | 0.05/0.062 | 0.044/0.054 | 0.035/0.044 | 0.03/0.039 | 0.026/0.033 |
| Scenario 5 | 60 | 0.646 | 0.089/0.111 | 0.059/0.075 | 0.052/0.066 | 0.042/0.053 | 0.036/0.045 | 0.031/0.04 | 0.027/0.035 |
| Scenario 5 | 120 | 0.827 | 0.068/0.086 | 0.047/0.06 | 0.037/0.046 | 0.031/0.04 | 0.028/0.035 | 0.023/0.029 | 0.021/0.026 |
| **PWE** | | | | | | | | | |
| Scenario 1 | 60 | 0.772 | 0.141/0.184 | 0.1/0.128 | 0.082/0.105 | 0.064/0.081 | 0.056/0.071 | 0.049/0.062 | 0.043/0.054 |
| Scenario 1 | 120 | 0.849 | 0.095/0.123 | 0.07/0.089 | 0.057/0.073 | 0.045/0.057 | 0.04/0.05 | 0.035/0.044 | 0.03/0.038 |
| Scenario 2 | 60 | 0.741 | 0.168/0.216 | 0.112/0.14 | 0.088/0.111 | 0.079/0.099 | 0.063/0.079 | 0.054/0.068 | 0.047/0.059 |
| Scenario 2 | 120 | 0.843 | 0.104/0.131 | 0.072/0.09 | 0.058/0.073 | 0.052/0.065 | 0.042/0.053 | 0.036/0.045 | 0.031/0.039 |
| Scenario 3 | 60 | 0.772 | 0.112/0.142 | 0.082/0.103 | 0.066/0.081 | 0.057/0.07 | 0.046/0.058 | 0.041/0.052 | 0.035/0.044 |
| Scenario 3 | 120 | 0.834 | 0.085/0.107 | 0.063/0.078 | 0.05/0.062 | 0.044/0.054 | 0.035/0.044 | 0.031/0.04 | 0.027/0.034 |
| Scenario 4 | 60 | 0.667 | 0.106/0.137 | 0.074/0.094 | 0.059/0.074 | 0.052/0.064 | 0.042/0.052 | 0.036/0.046 | 0.031/0.039 |
| Scenario 4 | 120 | 0.750 | 0.089/0.113 | 0.062/0.079 | 0.05/0.062 | 0.044/0.054 | 0.035/0.044 | 0.03/0.038 | 0.026/0.033 |
| Scenario 5 | 60 | 0.646 | 0.088/0.11 | 0.059/0.075 | 0.052/0.066 | 0.042/0.053 | 0.036/0.045 | 0.031/0.04 | 0.028/0.035 |
| Scenario 5 | 120 | 0.827 | 0.069/0.087 | 0.048/0.061 | 0.037/0.046 | 0.032/0.04 | 0.028/0.035 | 0.023/0.029 | 0.021/0.027 |

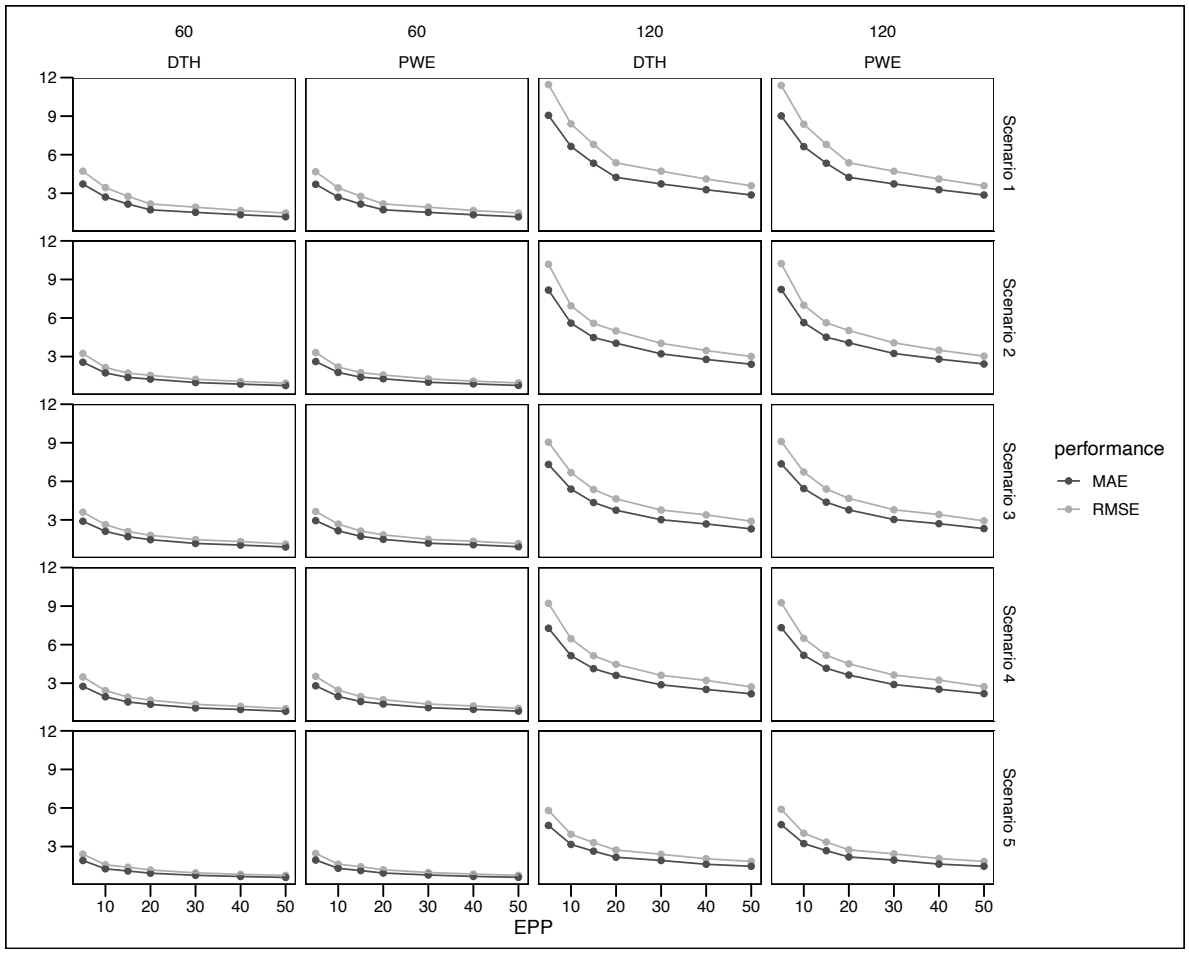



Figure 5.7: Results for the performance of the estimation of the Difference in RMST ATE Measure as a function of the Event per Parameters at 60 and 120 months

Table 5.3: Model Performance by EPP for the Difference in RMST. Values shown are MAE / True Value

| Scenario | Time | Truth | EPP5 | EPP10 | EPP15 | EPP20 | EPP30 | EPP40 | EPP50 |
|---|---|---|---|---|---|---|---|---|---|
| **DTH** | | | | | | | | | |
| Scenario 1 | 60 | 5.547 | 3.715/4.714 | 2.704/3.438 | 2.157/2.759 | 1.71/2.17 | 1.51/1.913 | 1.324/1.657 | 1.167/1.459 |
| Scenario 1 | 120 | 13.767 | 9.059/11.457 | 6.641/8.398 | 5.335/6.8 | 4.236/5.367 | 3.725/4.714 | 3.272/4.107 | 2.866/3.586 |
| Scenario 2 | 60 | 3.625 | 2.558/3.242 | 1.728/2.143 | 1.374/1.714 | 1.244/1.539 | 0.98/1.231 | 0.853/1.066 | 0.74/0.931 |
| Scenario 2 | 120 | 11.783 | 8.174/10.182 | 5.601/6.949 | 4.481/5.587 | 4.039/4.993 | 3.212/4.035 | 2.779/3.467 | 2.403/3.01 |
| Scenario 3 | 60 | 5.237 | 2.894/3.597 | 2.109/2.625 | 1.693/2.098 | 1.458/1.799 | 1.166/1.463 | 1.039/1.311 | 0.889/1.121 |
| Scenario 3 | 120 | 13.505 | 7.314/9.048 | 5.395/6.683 | 4.348/5.367 | 3.749/4.636 | 3.012/3.767 | 2.683/3.388 | 2.299/2.896 |
| Scenario 4 | 60 | 7.501 | 2.753/3.485 | 1.943/2.425 | 1.54/1.922 | 1.351/1.682 | 1.073/1.353 | 0.953/1.206 | 0.814/1.02 |
| Scenario 4 | 120 | 19.819 | 7.278/9.21 | 5.14/6.458 | 4.131/5.138 | 3.608/4.472 | 2.875/3.614 | 2.517/3.217 | 2.173/2.714 |
| Scenario 5 | 60 | 6.254 | 1.912/2.405 | 1.265/1.577 | 1.095/1.388 | 0.923/1.175 | 0.76/0.95 | 0.668/0.838 | 0.596/0.753 |
| Scenario 5 | 120 | 16.835 | 4.639/5.804 | 3.162/3.956 | 2.638/3.302 | 2.165/2.726 | 1.913/2.396 | 1.618/2.047 | 1.462/1.839 |
| **PWE** | | | | | | | | | |
| Scenario 1 | 60 | 5.547 | 3.694/4.671 | 2.695/3.418 | 2.158/2.76 | 1.712/2.172 | 1.51/1.914 | 1.324/1.658 | 1.166/1.458 |
| Scenario 1 | 120 | 13.767 | 9.019/11.394 | 6.627/8.373 | 5.333/6.799 | 4.238/5.369 | 3.725/4.712 | 3.271/4.107 | 2.864/3.585 |
| Scenario 2 | 60 | 3.625 | 2.61/3.308 | 1.764/2.19 | 1.394/1.749 | 1.269/1.568 | 1.001/1.259 | 0.872/1.088 | 0.751/0.947 |
| Scenario 2 | 120 | 11.783 | 8.222/10.234 | 5.641/7 | 4.507/5.625 | 4.069/5.025 | 3.24/4.069 | 2.8/3.496 | 2.423/3.033 |
| Scenario 3 | 60 | 5.237 | 2.94/3.653 | 2.149/2.671 | 1.72/2.132 | 1.484/1.833 | 1.182/1.483 | 1.063/1.339 | 0.91/1.148 |
| Scenario 3 | 120 | 13.505 | 7.358/9.1 | 5.434/6.727 | 4.374/5.397 | 3.775/4.668 | 3.026/3.786 | 2.704/3.415 | 2.32/2.922 |
| Scenario 4 | 60 | 7.501 | 2.794/3.534 | 1.97/2.459 | 1.573/1.96 | 1.379/1.715 | 1.094/1.375 | 0.965/1.229 | 0.83/1.04 |
| Scenario 4 | 120 | 19.819 | 7.32/9.258 | 5.168/6.493 | 4.162/5.174 | 3.635/4.504 | 2.896/3.635 | 2.529/3.239 | 2.186/2.73 |
| Scenario 5 | 60 | 6.254 | 1.947/2.463 | 1.305/1.628 | 1.127/1.427 | 0.936/1.194 | 0.783/0.978 | 0.674/0.855 | 0.601/0.758 |
| Scenario 5 | 120 | 16.835 | 4.701/5.894 | 3.229/4.036 | 2.676/3.346 | 2.187/2.742 | 1.942/2.427 | 1.632/2.064 | 1.463/1.843 |

## 5.3 Discussion & Conclusion

Several papers highlighted the need for more clinically interpretable measures in survival analysis (Stensrud et al. 2018; Uno et al. 2014; Verbeeck and Saad 2024). Measures like the Risk Difference, Relative Risk, and the difference in Restricted Mean Survival Time are more clinically meaningful and causally sound than the standard Hazard Ratio.

Flexible parametric models, such as the discrete-time hazard and piece-wise exponential models evaluated in this work, are well suited to obtaining such measures. They allow for time-varying effects of the covariates and they can directly estimate smoothed cumulative measures without relying on external estimators of the cumulative hazard, like the Breslow method used with Cox models. Moreover, these models allow for the implementation of inverse probability weights (IPW) or double robust methods to obtain measures of the ATE.

However, smoothing time-varying effects along with the baseline hazard, in addition to calculating average treatment effects increase the number of parameters in the models and it involves a series of computations (e.g., numerical integration, exponentiation, and ratios) that increase the error of estimation. This last aspect is in stark contrast with flexible parametric models on pseudo-ovnservations, from which, depending on the link function adopted, several clinically useful measures can be directly obtained (Ambrogi, Biganzoli, and Boracchi 2008). In this work, we did not compare the performances of the pseudo-observation model on a cumulative probability scale with the models on the instantaneous probability scale adopted here.

To our knowledge, no study has systematically assessed how well these flexible approaches estimate ATEs as a function of EPP. As our primary goal was to assess how closely flexible parametric models can approximate a known "ground truth" average treatment effect from our simulated data, we focused measure of prediction accuracy like the Mean Absolute Error and Root Mean Squared Error. Indeed they provide a clear, continuous measure of how performance degrades as the data becomes sparser.

However, we did not focus on calibration or discrimination performances, though these are interesting topics that have been partially addressed in machine learning literature (Infante, Miceli, and Ambrogi 2023).

Nevertheless, despite the adoption of splines, the bias was a limited component of the overall Mean Squared Error. The main issue was the variability of the estimates, which was clearly affected by the EPP. We found that an EPP greater than 20 is often needed to get precise estimates. Indeed, using a lower EPP can lead to imprecise estimates where a single study may substantially under- or over-estimate the true causal effect due to high sampling variability.

We also did not find major performance differences among the ATE measures, although the Relative Risk, being a ratio, appeared slightly more reliable than difference-based measures at smaller sample sizes. This may be because the ratio metric is scaled by the baseline risk, potentially stabilizing its variance relative to an absolute difference measure.

In this study, we guaranteed exchangeability by design (i.e., no confounding) to isolate the performance degradation induced by the splines and the final ATE numerical computations. Future studies will need to evaluate model performance when exchangeability is achieved

through either IP weighting or outcome modeling or double robust methods, also under scenarios of mispecification of the propensity model.

We also plan to extend this research into the context of competing risks, where flexible models are also adopted (G. Biganzoli, Marano, and Boracchi 2025) and where the CIF is often the causal quantity of interest (Rudolph, Lesko, and Naimi 2020).

In conclusion, this work is a first step in a research line focused on vetting flexible models for producing causally sound and interpretable results. The next critical question, which is addressed in the following chapter, relates to how complex the prognostic component of these models should be.

# 6 Modeling strategies for a flexible estimation of the cause-specific cumulative incidence function in the context of long follow-ups: model choice and predictive ability evaluation

As already introduced in Section 3.7, selecting an appropriate model complexity for the estimation of the cumulative quantities with flexible models for time-to-event data is crucial, as the reliability of the ATE depends on them. This chapter proposes a comprehensive workflow that integrates various methodologies—including information criteria, bootstrap validation, and measures of predictive ability—to evaluate and select a robust model.

We will illustrate this approach by analyzing data from two breast cancer clinical trials within a competing risk framework, focusing on the cause-specific cumulative incidence function of distant recurrences (DRs). However, these considerations are easily translatable to a context without competing events.

The analysis will specifically account for potential time-dependent and interaction effects while comparing an indirect modeling approach based on the sub-distribution hazard with a direct approach using pseudo-observations Section 4.6. The analysis must account for possible time-dependent effects of both treatment and axillary nodal status, as well as the interaction effect between treatment and nodal status.

For simplicity, standard regression splines will be used to specify a smooth function of time, and classic information criteria—AIC and the Quasi-Information Criterion (QIC)—will be considered for the model selection step (Akaike 1992; Cui 2007) .

In addition, to assess the robustness of the complex effects underlying the CIF of DRs, a non-parametric bootstrap procedure along with an AIC-based selection algorithm of the model complexity is implemented.

The discriminatory capacity of the model is evaluated using time-dependent measures of discrimination, such as the Concordance Index (C-Index) and Net Reclassification Improvement (NRI), with optimism correction (HARRELL, LEE, and MARK 1996). A time-dependent version of the category less NRI measure is proposed to evaluate the usefulness of increasing the model structure complexity at different follow-up times.

## 6.1 Methods

### 6.1.1 Model Complexity Selection

The proportionality of the SDH and the additivity of the effects are conditions that may not be satisfied in real applications, particularly in the presence of extended follow-up times and intricate biological phenomena. However, randomized clinical trials and longitudinal cohort studies are usually not designed to test complex interactions between covariates and/or time-dependent effects. In these cases, classical hypothesis tests can be underpowered or impossible to apply for selecting the proper specification of the baseline function. This motivates the adoption of alternative criteria for the model selection task.

Information criteria like AIC and QIC (for the pseudo-likelihood of the GEE) represent a relevant approach. AIC quantifies the amount of information lost when a specific model is used to describe the real underlying process generating the data and it is defined as:

$$AIC = 2K - 2\ln(\hat{\theta}; X), \tag{6.1}$$

where $K$ is the number of model parameters, and $\ln(\hat{\theta}; X)$ is the log-likelihood function evaluated at the maximum likelihood estimates. The best model among a set of candidates is the one with the minimum value of the criterion. Therefore, AIC favors the goodness of fit of a model while penalizing over-fitted models with numerous parameters. It is important to note that AIC and AIC-like statistics are relative measures for comparing two or more models and do not indicate how closely a fitted model reflects the "true" data-generating process (Burnham and Anderson 2004).

In the following section, the model selection strategy adopted in the current work is illustrated in detail. We considered a set of six regression models for distant recurrences, as shown below.

The least complex model is a proportional hazards model with no interactions among covariates. The subsequent models introduce interaction effects of increasing complexity, including interactions between covariates and the baseline SDH function. The most complex effect consists of a second-order interaction between two covariates (nodal status and treatment) and the baseline SDH/risk function (Model 1).

A second-order interaction refers to the combined effect of three variables on the response variable that cannot be explained by their individual effects or first-order interactions. These patterns are based on previous clinical knowledge and reflect actual research hypotheses about tumor dormancy.

In the formulas below:

- $Y$ represents the discrete SDH or the risk for the pseudo-values approach,

- $\Psi$ is the vector of spline basis functions for the baseline time function,

- $X_T$ is the dummy variable for treatment (0 = surgical resection of the primary tumor by mastectomy, 1 = surgical resection by quartectomy),

- $X_N$ is the dummy variable for axillary nodal status (0 = node-negative, 1 = node-positive),

- $\otimes$ indicates the tensor product between the dummy variables and the spline bases. The tensor product term generalizes interactions by modeling the effect of two or more predictors as a product of basis functions for each predictor**.**

As a complex multipeaked shape of the baseline SDH function was expected in the analysis of the datasets adopted (these datasets are illustrated in a subsequent paragraph), B-splines were considered in the indirect approach (Rosenberg 1995). In the direct approach, the target function is the CIF, so cubic splines with linearity constraints on the terminal knots were considered (Harrell , 2015) instead of B-splines, as less complexity in the functional shape was expected.

1. **Model 1**: A model with a second-order interaction effect between nodal status, treatment, and the baseline function. The proportionality of the SDH or the risk function is not assumed.

$$\log(Y) = \Psi\mathbf{B}_0 + X_T\beta_1 + X_N\beta_2 + X_T \times X_N\beta_3 + X_T \otimes \Psi\mathbf{B}_4 + X_N \otimes \Psi\mathbf{B}_5 + X_T \times X_N \otimes \Psi\mathbf{B}_6 \tag{6.2}$$

2. **Model 2**: A model with an interaction effect between nodal status and the baseline function, an interaction effect between treatment and the baseline function, and an interaction effect between nodal status and treatment. The proportionality of the SDH or the risk function is not assumed.

$$\log(Y) = \Psi\mathbf{B}_0 + X_T\beta_1 + X_N\beta_2 + X_T \times X_N\beta_3 + X_T \otimes \Psi\mathbf{B}_4 + X_N \otimes \Psi\mathbf{B}_5 \tag{6.3}$$

3. **Model 3**: A model with an interaction term between treatment and the baseline function, an interaction term between nodal status and treatment, and an additive effect of nodal status. The proportionality of the SDH or the risk function is not assumed for treatment but is assumed for nodal status.

$$\log(Y) = \Psi\mathbf{B}_0 + X_T\beta_1 + X_N\beta_2 + X_T \times X_N\beta_3 + X_T \otimes \Psi\mathbf{B}_5 \tag{6.4}$$

4. **Model 4**: A model with an interaction term between nodal status and the baseline function, an interaction term between nodal status and treatment, and an additive effect of treatment. The proportionality of the SDH or the risk function is not assumed for nodal status but is assumed for treatment.

$$\log(Y) = \Psi\mathbf{B}_0 + X_T\beta_1 + X_N\beta_2 + X_T \times X_N\beta_3 + X_N \otimes \Psi\mathbf{B}_4 \tag{6.5}$$

5. **Model 5**: A model with an interaction term between nodal status and treatment and an additive effect of treatment. The proportionality of the SDH or the risk function is assumed.

$$\log(Y) = \Psi\mathbf{B}_0 + X_T\beta_1 + X_N\beta_2 + X_T \times X_N\beta_3 \tag{6.6}$$

6. **Model 6**: An additive model with no interaction effects. The proportionality of the SDH or the risk function is assumed.

$$\log(Y) = \Psi\mathbf{B}_0 + X_T\beta_1 + X_N\beta_2 \tag{6.7}$$

The covariates used in the present work, other than the time variable, are all categorical. It is worth noting that, in cases where both categorical and numerical covariates are used, the following rules may be considered for specifying non-linear effects and interactions.

The specification of non-linear effects or their interactions with a categorical variable follows the same approach illustrated above. Furthermore, the specification of interaction effects among numerical variables—including, as a special case, a time-dependent effect of a numerical covariate—may be obtained by employing a tensor product between the spline basis matrix for the baseline sub-distribution hazard or risk function and the spline basis matrix for the numerical covariate's non-linear effect. A good example of the specification of a time-dependent effect of a numerical covariate (tumor size) on the cause-specific hazard of the event can be found in (Boracchi, Biganzoli, and Marubini 2003).

The model selection procedure was as follows: each of the six models above was fitted several times with an increasing number of spline knots. The optimal model was selected according to the minimum QIC (in the direct approach) or AIC (in the indirect approach). To assess the robustness of the AIC-based selection, a perturbation of the data was performed using sampling with replacement (i.e., non-parametric bootstrap). The frequencies of the best model selected over 1,000 bootstrap samples were then compared with the model selected using the original sample.

### 6.1.2 Measuring the discriminatory capacity of the models

While the Akaike Information Criterion (AIC) provides valuable insights into which model best describes the real process underlying the incidence of an event of interest, clinical decision-making requires additional metrics. Specifically, when using a model as a tool for predicting patient outcomes, measures of model discrimination should complement information criteria.

In most models, well-established prognostic factors are always included. The primary evaluation concerns whether incorporating complex terms—such as time-dependent effects or interactions between covariates—is necessary to predict the absolute cumulative risk of an event. Both overall and time-dependent measures of accuracy are relevant. In the literature, several accuracy measures have been adapted from classical survival analysis to the context

of competing risks. These include net reclassification methods, the area under the receiver operating characteristic curve (AUC), and prediction error curves.

### 6.1.3 5.2.5.1 Time dependent C-index for competing risks in the discrete-time SDH and model on pseudovalues

The C-statistic proposed by Harrell is widely recognized as a valid measure of discrimination and has been adapted for competing risks (Wolbers et al. 2014). A key advantage of this measure is that it depends solely on the cumulative incidence function of an event.

A model with optimal discrimination ability, as measured by the time-dependent C-index, should maximize the concordance probability:

$$P(M(X_i) > M(X_j)|D_i = 1 \; and \; T_i < T_j \; or \; D_j = 2) \tag{6.8}$$

This measure can be evaluated at different follow-up time points. The inverse probability of censoring weighting (IPCW) methodology is applied to address the issue of censored observations when evaluating concordance, ensuring that $T_i < T_j$ even when $j$ corresponds to a censored observation.

Given that the present study focuses on estimating the C-index for cumulative incidence (CIF) using two different modeling frameworks, we employ the time-dependent C-index measure proposed by Wolbers. This measure is applied in the context of the discrete-time sub-distribution hazard model and the transformation model on pseudo-observations at 5 and 10 years of follow-up.

$$C_{DR}(t) = \Big(F_{DR}(t, X_i) > F_{DR}(t, X_j)|D_i = DR\Big) \; and \Big( \; T_i \leq t \; \Big) and \; \Big(T_i < T_j \; or \; D_j = OT \; or \; D\Big) \tag{6.9}$$

### 6.1.4 Time-dependent, category-less, Net Reclassification Improvement (NRI) for competing risks

For models containing standard risk factors and possessing reasonably good discrimination, a substantial "independent" association of the new marker (i.e., an added predictor increasing model complexity) with the outcome is necessary to achieve a significantly higher time-dependent C-index. Moreover, the C-statistic does not provide insights into the relative merits of alternative models for risk prediction.

The category-less Net Reclassification Improvement (NRI) quantifies the improvement in discrimination achieved by incorporating an additional predictor into the model. Generally, an increase in discriminative ability arises from either an increase in the predicted probability of an event for subjects who experience it or a decrease in the predicted probability among those who remain event-free. The absolute NRI, adopted in this work, measures the proportion of

subjects correctly reclassified—non-event subjects with decreased event probability and event subjects with increased event probability—among all subjects.

To complement this measure, we extend the concept of category-less NRI as a time-dependent metric for improved discrimination. The original formulation of the category-less NRI is:

$$NRI = \frac{1}{N}\sum_{i}^{N} 1[P_{i,new}(event) - P_{i,old}(event) > 0,\ \Delta_i = 1\ \vee\ P_{i,new}(event) - P_{i,old}(event) < 0,\ \Delta_i = 0] \tag{6.10}$$

where $P_{i,new}(event)$ and $P_{i,old}(event)$ represent the predicted probabilities of an event from the new and old models, respectively, and $\Delta_i$ is the event indicator (1 for an event, 0 otherwise).

In our setting, the absolute time-dependent category-less NRI is defined as:

$$NRI(t) = \frac{1}{N}\sum_{i}^{N} 1\ [\hat{F}^{new}_{DR}(t|x_i) - \hat{F}^{old}_{DR}(t|x_i) > 0,\ \Delta_i = 1\ \ or\ \hat{F}^{new}_{DR}(t|x_i) - \hat{F}^{old}_{DR}(t|x_i) < 0,\ \Delta_i = 0] \tag{6.11}$$

where $F^{new}_{DR}(t|X_i)$ represents the CIF function of DR up to time $t$ for subject $i$ with covariate values $x$, predicted with a model of increased complexity, and $F^{old}_{DR}(t|X_i)$ represents the CIF function of DR up to time $t$ predicted with the basic additive model. The term $\Delta_i = 1(T_i < t\ \cap \epsilon_i = 1)$, where $\epsilon_i = 1$ defines the event of interest.

Thus, the category-less NRI in its time-dependent form assesses how effectively a new model, at time $t$, reclassifies subjects into higher or lower predicted risk categories relative to an old model, without relying on predefined risk categories. Instead, it focuses solely on whether an individual's predicted risk increases or decreases correctly relative to the actual outcome.

### 6.1.5 Illustration of the data

The "Milan 1" trial, held between June 1973 and May 1980, included 701 patients with unilateral breast cancer and no palpable axillary lymph nodes. Participants were randomly assigned to undergo either a radical mastectomy (MAST) or breast-conserving surgery followed by radiotherapy (BCS+RT). In the BCS+RT group, complete axillary lymph node dissection was performed alongside BCS, followed by radiotherapy, with adjuvant chemotherapy introduced for node-positive patients starting in 1976. The trial results indicated no significant differences in disease-free or overall survival between the two groups, suggesting that BCS+RT was comparable to, if not better than, MAST.

In the "Milan 3" trial, conducted from December 1987 to December 1989, 567 eligible patients with breast tumors up to 2.5 cm in size were randomized to either BCS with or without immediate radiotherapy (RT). All patients underwent complete axillary lymph node dissection, with adjuvant therapy provided to those with node-positive disease, tailored according to menopausal status and tumor estrogen receptor (ER) status. The follow-up procedures

mirrored those of the "Milan 1" trial. Initial results showed a reduced risk of local recurrence with RT following BCS, with no significant differences in overall survival between groups, except for patients with node-positive disease.

## 6.2 Results

### 6.2.1 Results from original data

At a first sight, the Aalen-Johansen estimates of the CIF of DRs (Figure 6.1) suggest that conditioned to the nodal status, a differential effect of the treatment arm could be present. Indeed, in Milan 1 trial, the detrimental effect of MAST seems a lot higher in N+ patients as respect to N- patients. Also, in Milan 3 trial the detrimental effect of omitting RT is visible in N+ patients, as respect to N- patients which on the contrary might be associated to a lower risk of DMs.

The visualization of the non-parametric estimates of CIF can give useful insights of the possible presence of the time-dependent effects. Indeed, a different shape it is observed between N+ and N-. N+ curves show two steep increases (one between 0 and 5 years, another between 5 and 10 years) separated by a flex, whereas N- curves show a more stable increase. This could reasonably motivate to specify different baseline risk functions in the pseudo-values models and different baseline of the sub-distribution hazard functions, at least for N+ and N- subjects, indicating the non-proportional hazard/risk for nodal status.

```
pdf
  2
```

```
TableGrob (1 x 2) "arrange": 2 grobs
  z     cells    name           grob
1 1 (1-1,1-1) arrange gtable[layout]
2 2 (1-1,2-2) arrange gtable[layout]
```

Figure 6.2 displays the results obtained with the indirect approach based on modeling discrete-time SDH. The model selected with AIC for the Milan 1 trial cohort shows a modulation effect of nodal status on the effect of treatment adopted. No interaction terms between the baseline SDH function with N nor with T was selected not supporting the drift from proportionality. Thus, the SDH curves display the same functional form with only a vertical shift in the cumulated risks determined by the T-N combination (**a**). Comparing the functions computed with the product-limit formula from SDH estimates, and those obtained with the AJ estimates, a relevant bump on of MAST N+ women is not fitted at all (**b**).

For the Milan 3 trial cohort the AIC selected model does not assume proportionality of the SDH for the effect of nodal status. However, conditioned on nodal status, proportionality is maintained leading to constant effects of treatment over time (**d**). This seems reasonable as also a different functional form of the CIF was observed from the AJ estimates. In

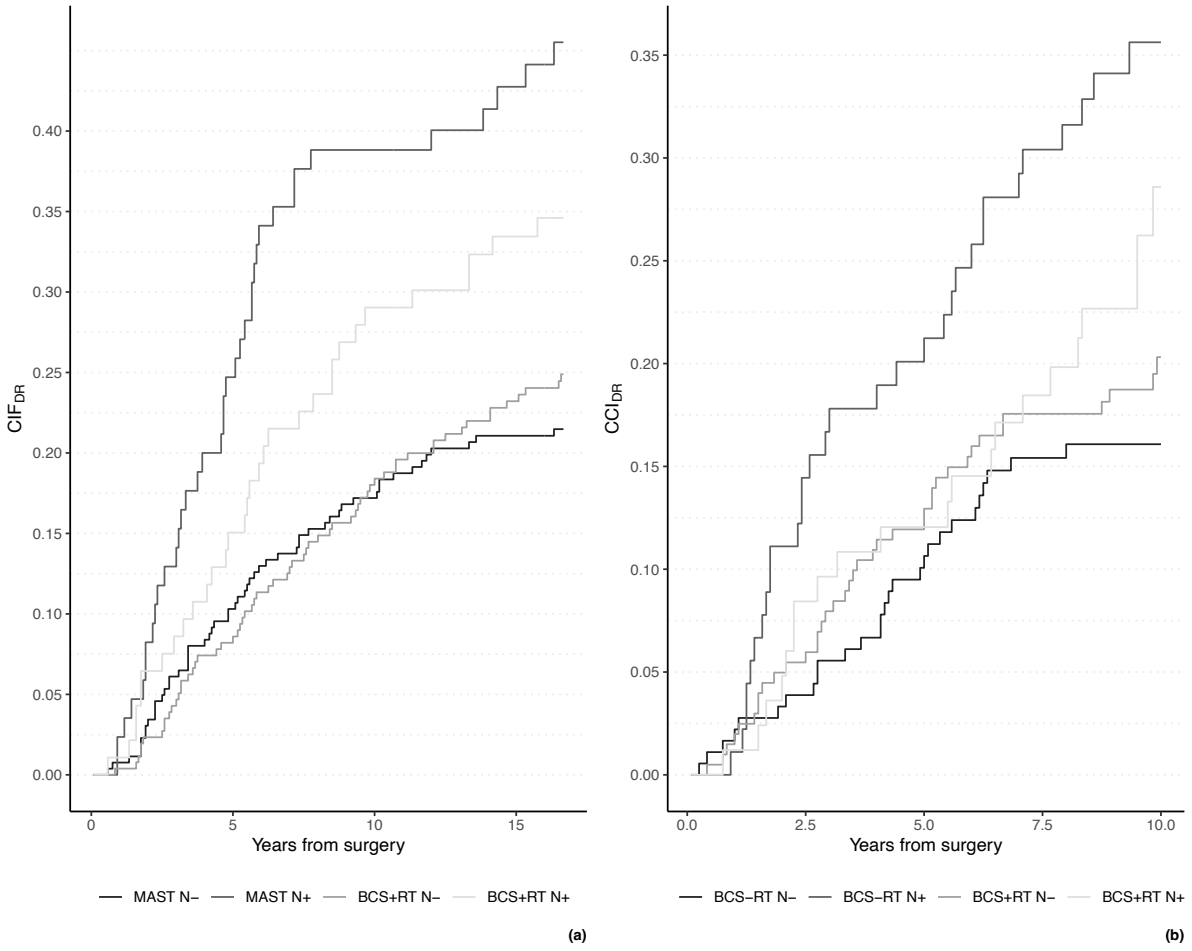



Figure 6.1: **Figure 1.** Non-parametric estimates of CIF of DRs conditioned on surgical procedure and lymph-nodal status in Milan 1 (a) and Milan 3 (b)

(**e**) the functions computed with the product-limit formula compared to the non-parametric estimates.

```
pdf
  2
```

```
TableGrob (2 x 2) "arrange": 4 grobs
  z     cells    name           grob
1 1 (1-1,1-1) arrange gtable[layout]
2 2 (1-1,2-2) arrange gtable[layout]
3 3 (2-2,1-1) arrange gtable[layout]
4 4 (2-2,2-2) arrange gtable[layout]
```

According to the exploratory aim, 30 time-points in the follow-up in which to compute the pseudo-observations were preferred to the suggested number of 8-10, with a slight increase of the computational cost. The QIC based model selection procedure chose a model structure with an interaction effect between N and TRT. No time-dependent effects were chosen suggesting essentially a proportional risks condition. In (**a, b**) of Figure 6.3 estimates of of distant recurrences of women underwent to a BCS+RT in MI1 and MI3 compared to the other treatment are displayed.

```
pdf
  2
```

```
TableGrob (1 x 2) "arrange": 2 grobs
  z     cells    name           grob
1 1 (1-1,1-1) arrange gtable[layout]
2 2 (1-1,2-2) arrange gtable[layout]
```

### 6.2.2 Perturbation of the data: the non-parametric bootstrap approach

The procedure of model selection based on AIC or QIC has been applied to 1000 bootstrap sample to evaluate the robustness of the model selected. Figure 6.4 reports the relative frequency of model structure selection separated for modeling framework and trial.

The non-parametric bootstrap procedure revealed substantial variability in selecting the level of complexity for the SDH modeling framework in the Milan 1 cohort. Notably, an interaction effect between treatment and nodal status appears reasonable, as the additive model structure is selected in less than 10% of 1000 independent samples. Interestingly, a distinct baseline SDH function for N+ and N- is supported (66%: 28% for Model structure 4, 24% for Model structure 1, and 14% for Model structure 2).

In the Milan 3 cohort, a robust interaction effect between nodal status and treatment is observed, with the additive model structure selected only 1.5% of the time. Furthermore,

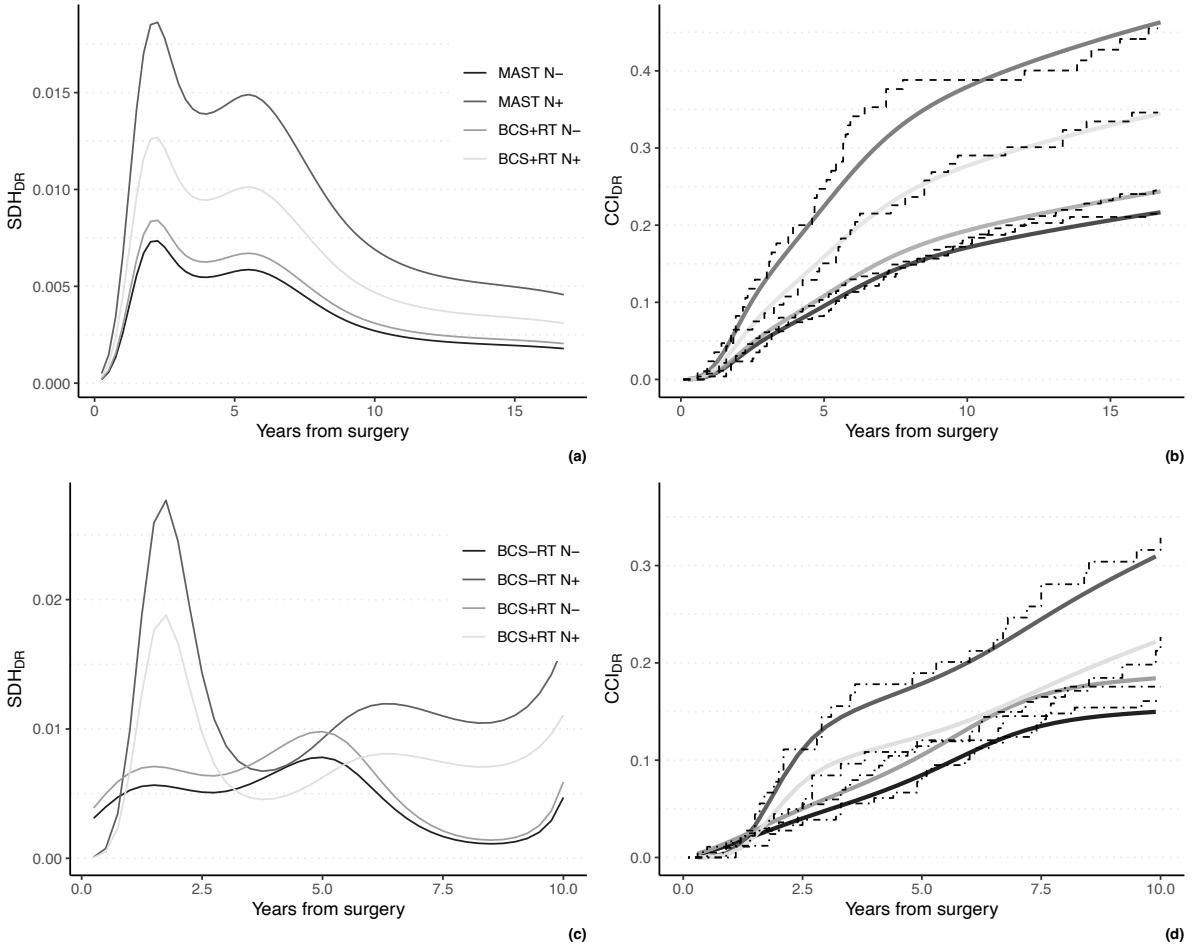



Figure 6.2: In **(a)** and **(c)** the point estimates of the discrete-time SDH function are illustrated conditioned on surgical procedure and lymph-nodal status for Milan 1 and Milan 3 trial respectively. In **(b)** and **(d)** the computed CIF values derived by the product-limit formula utilizing SDH estimates are reported (solid lines) and compared to the non-parametric Aalen-Johansen estimates (dashed lines).

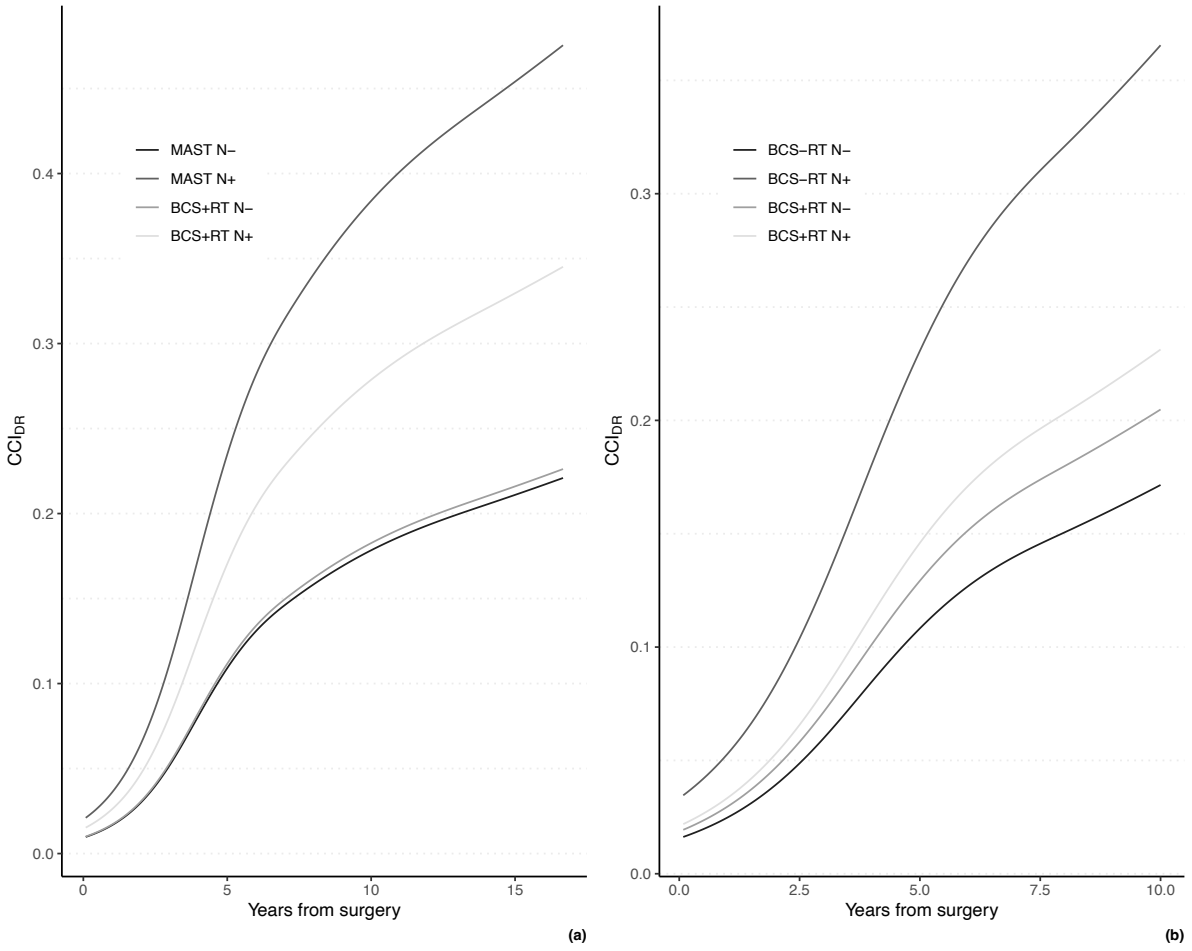



Figure 6.3: Results of the analysis with the method based on pseudo-observations to estimate the CIF of DRs for Milan 1 **(a)** and Milan 3 **(b)**

there is a pronounced indication of a differential functional form of the baseline SDH among node-positive and node-negative patients. Specifically, the interaction term between baseline SDH and N is maintained 85% of the time (47% for Model structure 2, 35% for Model structure 4, and 13% for Model structure 1).

Concerning the pseudo-values modeling framework, it generally favors lower levels of complexity compared to the SDH counterpart. In the Milan 1 cohort, the modulation effect of nodal status on treatment effect remains evident but is less robust at 68%, and the selection of time-dependent effects is infrequent. In the Milan 3 cohort, the interaction effect is robust at 75%, although 25% of the times the additive model is chosen.

It appears that the data perturbation procedure does not have a clear response on which is the most robust model structure, particularly in the sub-distribution hazard modeling framework. Additional insights on the practical utility of considering more complex model structures still selected by the algorithm based on the information criteria might be derived by the assessment of their predictive ability.

Regarding the selection of the number of the knots and as consequence the number of the spline basis, for the sub-distribution hazard baseline in Milan 1 regardless of the structure complexity, 3 interior knots (7 spline basis) were mostly selected. A greater variability of selection is shown for the sub-distribution hazard baseline of Milan 3.

On the contrary, for the baseline risk function of both Milan 1 and Milan 3, 3 knots (2 basis) were mostly preferred indicating that more complexity was not needed.

```
pdf
  2
```

### 6.2.3 Time dependent measures of discrimination

No relevant differences in the discriminatory capacity in relation to model complexity are present between the pseudo-value and SDH modeling frameworks. Figure 6.5 displays the time-dependent C-index evaluated for the six model structure at 5, and 10 years of the follow-up time. In the Milan 1 trial cohort, the additive model, despite being less frequently selected in the non-parametric bootstrap procedure, exhibits the same discriminatory capacity at 5 and10 years as respect the models that specify modulation effect of nodal status on treatment and the interaction effect between baseline risk function and nodal status. This was expected since the as the index is merely considering ranks of the CIF. Interestingly, these outperforms models that only specify a modulation effect of nodal status on treatment effect and the model that do not specify time-dependent effect of treatment, despite these models were those most selected in the bootstrap procedure.

```
pdf
  2
```

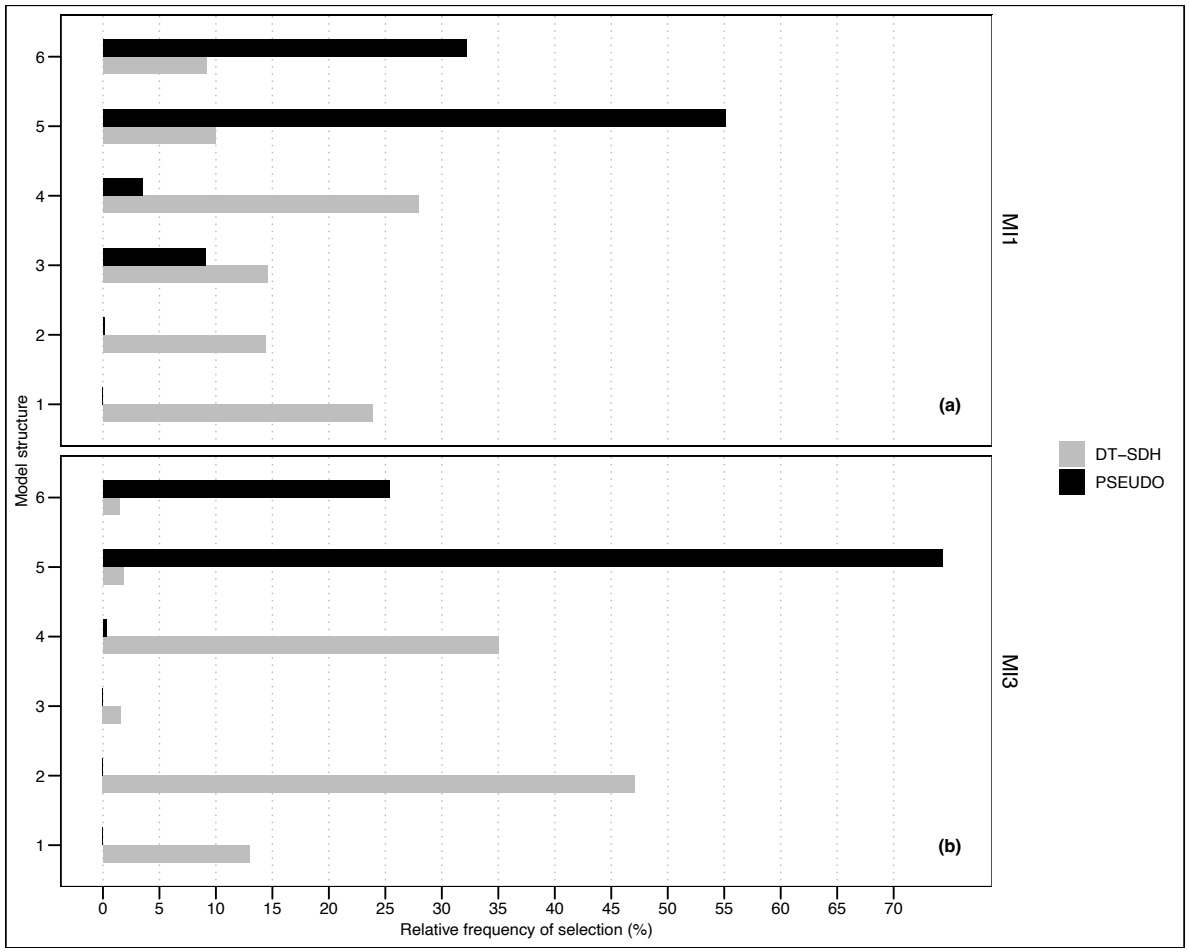



Figure 6.4: Relative frequency of model selection with the pertubation procedure through bootstrap resampling for Milan 1 **(a)** and Milan 3 **(b)**.

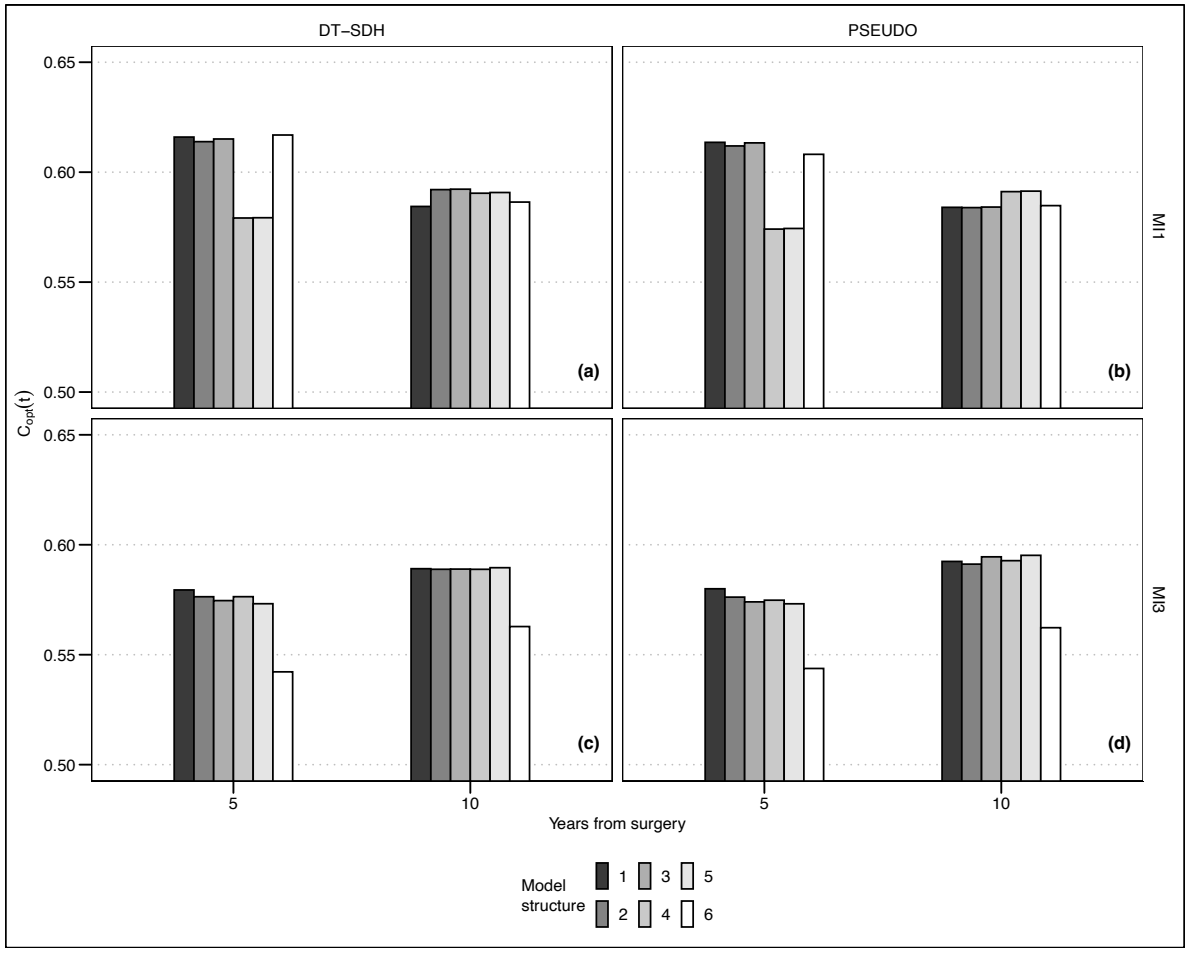



Figure 6.5: Time-dependent discriminatory capacity of the model structures assessed by means of time-dependent C-index. In **(a)** and **(b)** how the discriminatory capacity of the model structures in Milan 1 trial varies by increasing their complexity starting from the basic additive and proportional hazard model is shown for the sub-distribution hazard and the pseudo-observation modeling frameworks respectively. In **(c)** and **(d)** the same is shown for Milan 3 trial data.

Conversely, in the Milan 3 trial cohort, the additive model displays the lowest value, while multiplicative models, with or without a time-dependent effect of nodal status and treatment, demonstrate the highest discriminant ability. Notably, at a ten-year follow-up, relative differences in discriminatory capacity among the multiplicative and additive models diminish for Milan 1 but persist for Milan 3, where the additive model retains its lowest discriminant capacity.

```
pdf
  2
```

Figure 6.6 shows the results of the time-dependent NRI computed on the different model structures. In the SDH modeling framework, for Milan 1, Model structure 4, despite showing lower discriminatory ability, is the structure that reclassifies most efficiently patients, as respect to the additive model. The most complex model structures diminish the number of patients correctly reclassified. For Milan 3, considering the interaction effect between nodal status and treatment causes the highest increase in the number of patients correctly reclassified as having a distant recurrence. More complex interaction or time dependent effects do not increase further the reclassifying performance.

In the pseudo-values modeling framework, model structure 4 despite again the low discriminatory capacity show good performance as more complex model structures. However, considering the interaction effect of treatment and nodal status for Milan 1 data show a diminished increase in the number of reclassified patients but still a substantial increase (50%). For Milan 3, the interaction effect of nodal status and treatment is sufficient to substantially increase this number, and as what was seen in the SDH modeling framework, additional complexity does not add in the performance.

## 6.3 Discussion

In the context of competing risks, the CIF of an event is the preferred measure for clinicians and epidemiologists, as it directly represents the risk of a specific event occurring first. Consequently, models that enable direct inference on CIF are essential. The Fine & Gray (F&G) proportional sub-distribution hazard model remains the most widely used method in this context due to its similarity to the Cox model and the widespread availability of software. For long-term follow-up studies, flexible competing risks models that incorporate baseline hazard/risk offer significant advantages in specifying the functional form of time-dependent effects. While interaction terms and non-linear effects can be easily incorporated into the F&G model, specifying flexible functional forms for time-dependent effects or complex higher-order interactions may be limited to simple shapes. Additionally, when predicting CIF, a flexible parametric approach using regression splines eliminates the need for a Breslow-type estimator of the cumulative incidence function, which heavily depends on the pattern of observed events. Instead, this approach relies directly on smoothed model-predicted probabilities.

Although a stratification approach within the F&G framework (Zhou et al. 2010) addresses challenges related to non-proportional hazard effects, it does not provide estimated effects

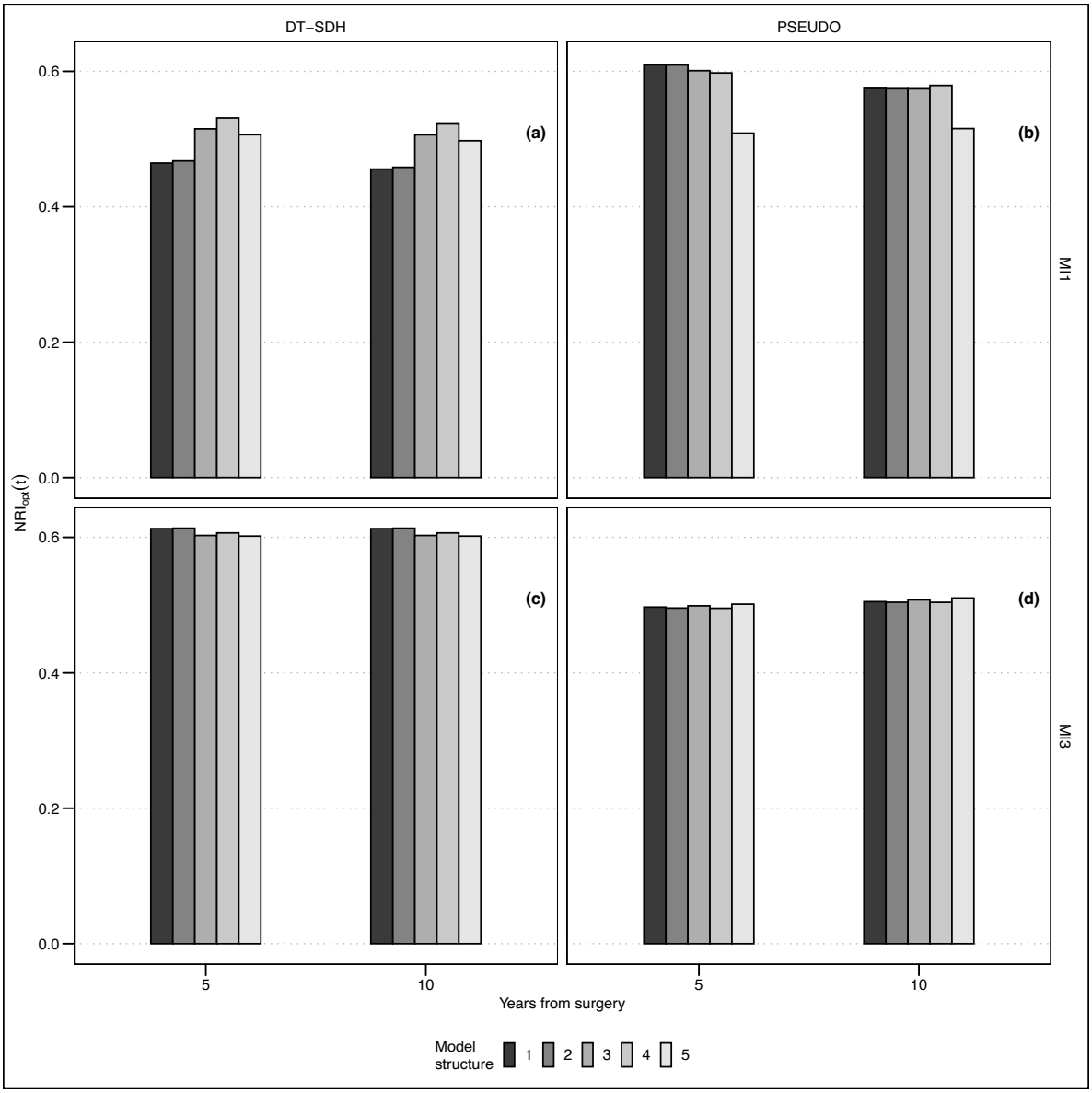



Figure 6.6: Time dependent measures of category-less absolute net reclassification improvement. In **(a)** and **(b)** the reclassification of the women of Milan 1 trial as a function of increasing complexity of the model structures is shown for the sub-distribution hazard and the pseudo-observation modeling frameworks respectively. In **(c)** and **(d)** the same is shown for the women of Milan 3 trial.

for stratification variables, which may limit its utility for clinicians interested in these effects. Furthermore, as noted by an anonymous reviewer, model selection between stratified and non-stratified models using information criteria is less straightforward with the F&G model compared to the approaches proposed in this work.

In this study, we employed two flexible competing risks modeling frameworks. The discrete-time sub-distribution hazard model demonstrated greater sensitivity in identifying time-dependent effects and ensured monotonic CIF estimates. However, a limitation of this method is the lack of a straightforward measure for the standard errors of CIF estimates. Variability can still be assessed using bootstrap procedures or the delta method.

The transformation model based on pseudo-observations provides a more direct and practical approach, although its sensitivity to time-dependent effects may depend on the choice of spline bases. It may also fail to guarantee monotonic estimates starting from zero at time zero for all covariate combinations, particularly when proportional risks are modeled. In such cases, monotonic splines, such as the integrated splines proposed by (Ramsay 1988), which enforce monotonicity under positivity constraints, could be beneficial. A key advantage of this framework is its ability to directly estimate variability around the estimates due to the asymptotic properties of pseudo-values.

Careful attention is required when specifying model complexity to avoid overfitting or underfitting. To address this, we introduced a statistical learning workflow that combines domain knowledge, graphical representation, and model complexity selection with perturbation methods and predictive metrics. This workflow enhances the robustness, accuracy, and interpretability of CIF estimation.

We also developed a modified version of the Net Reclassification Index (NRI) to account for the time dependency of prognostic relationships among covariates, enabling NRI evaluation at different follow-up points in the context of competing risks. When domain knowledge and graphical representations suggest covariate interactions or time-dependent effects not initially evident in the data, combining perturbation methods with tailored predictive performance measures can improve model robustness. Ideally, the model structure most frequently selected among bootstrap samples and exhibiting the highest predictive performance at different follow-up times should be chosen. For complex models, the added predictive capacity should be critically evaluated, balancing explainability with performance. If perturbation methods yield inconclusive results, prioritizing a model with significantly better predictive performance may be advisable, while carefully considering any loss in explainability. Additionally, comparing out-of-sample predictive performance using k-fold cross-validation with tailored metrics can further assess robustness. In this study, we used bootstrap model selection frequencies as an exploratory tool to assess variability and optimism correction, allowing full utilization of the available sample size without data partitioning.

Caution is advised when relying solely on discrimination indices such as the AUC, which reflect only rank-based measures. Discrepancies between the C-index and time-dependent NRI may arise, with NRI being more sensitive to the effects of increased model complexity. For example, in the analysis of the Milan 1 trial data within the sub-distribution hazard framework, the original data identified a model with an interaction effect between treatment and nodal status. Data perturbation revealed uncertainty in the most robust model structure,

with the most frequently selected model including a time-dependent effect of nodal status and an interaction effect between treatment and nodal status. Although this model showed lower discriminatory capability at 5 years based on the C-index, it exhibited the highest NRI at both 5 and 10 years.

It is important to note that neither the NRI nor the Integrated Discrimination Improvement (IDI)—whether continuous or categorical—are considered "proper" scoring rules. These metrics may suggest improvements even for biomarkers that add no real value, particularly when models are poorly calibrated, leading to inaccurate probability estimates (Hilden and Gerds 2013). While overfitting is less likely in models built with a small number of well-known prognostic markers, especially within the same study population, model calibration remains critical and should always be verified before further discrimination analyses (Leening et al. 2014). Summary statistics like NRI and IDI should not be interpreted in isolation but should be complemented with other metrics, such as changes in AUC, calibration measures, and decision-analytic tools, to assess whether model improvements translate into meaningful clinical utility. In this study, we combined NRI with AUC (or C-index), adapted for competing risks, applied to the training dataset with optimism correction. The choice of NRI in this work does not imply it is mandatory for all analyses; researchers should select metrics based on their strengths and limitations.

In both the pseudo-values and sub-distribution hazard frameworks for the Milan 1 data, the additive model demonstrated comparable discriminatory power to more complex models, suggesting that added complexity may be unnecessary. However, NRI indicated that incorporating more complex effects offered advantages over simpler models without time-dependent effects, consistent with the biological phenomenon's dynamics. These findings highlight that traditional metrics like the C-index may miss nuances captured by NRI, underscoring its value in model evaluation for CIF contexts. In the analysis of the Milan 3 trial data using the sub-distribution hazard framework, the original data identified a model with a time-dependent effect of nodal status and an interaction effect between treatment and nodal status. Data perturbation revealed significant uncertainty in identifying the most robust model structure, but the model with only an interaction effect between treatment and nodal status was favored by both the C-index and NRI at 5 and 10 years.

The methods used in this study are particularly straightforward when analyzing the CIF of a single specific event. In contrast, cause-specific hazard (CSH) approaches (Cheng, Fine, and Wei 1998; Ambrogi, Biganzoli, and Boracchi 2009) require additional modeling of the event-free survival function. A critical area for future research is comparing the performance of flexible methods in estimating CIF functions with CSH-based methods, especially as more flexible CSH estimation approaches have been proposed. For instance, (Belot et al. 2010) extended the Lunn and McNeil model to integrate time-dependent effects for event-specific covariate effects within a single model, while (E. M. Biganzoli et al. 2006) introduced a neural network with a multinomial activation function to derive smooth estimates of the cause-specific hazard function for each event.

This integrated approach is well-suited when there is a strong hypothesis about key prognostic factors but may be suboptimal in high-dimensional contexts requiring feature selection. In such cases, machine learning models (Ishwaran et al. 2014), combined with rule extraction

methods or the double penalization algorithm (Rodríguez-Girondo et al. 2013; Marra and Wood 2011) within the generalized additive model framework, may offer greater utility.

## 6.4 Conclusion

Accurate modeling of the cause specific cumulative incidence function is essential for both exploratory analyses, providing reliable prognostic information to patients and computing measures of average treatment effect.

With the availability of highly flexible models, researchers must carefully balance the level of complexity when modeling prognostic relationships among covariates. While relying on simpler models—assuming additivity and proportional hazards—may seem like a safe strategy to avoid over fitting, it often leads to under-fitted estimates, resulting in missed discoveries or inaccurate prognostic information.

Conversely, when substantial subject-specific knowledge is available, relying exclusively on machine learning algorithms for model selection can be problematic, as these methods may fail to adequately explain prognostic relationships or control for overfitting.

This paper proposes a workflow that integrates data perturbation and multiple measures of model discrimination to guide model selection. We believe this approach will be valuable for clinicians and epidemiologists working with long-term follow-up data and complex diseases, such as chronic conditions.
While the discrete-time hazard model has been extended for the sub-distribution hazard estimation in context where time is naturally grouped, a relevant question is if the piece-wise exponential model, that instead addresses a continuous time process, can be extended as well. This will be the task addressed in the next chapter.

# 7 A Piece-wise exponential model for the sub-distribution hazard of an event

In competing risks analysis, the cumulative incidence function (CIF) provides a clinically intuitive measure of event probability over time. While the Fine and Gray model directly targets the sub-distribution hazard (SDH), its proportional hazards assumption can be restrictive. Alternative flexible methods are often limited to discrete time intervals or rely on complex weighting schemes. This chapter introduces a novel, flexible, and continuous-time framework to address these limitations by extending the Piecewise Exponential (PWE) model to the competing risks setting.

Rather than developing a complex new weighting scheme, which can be computationally intensive and difficult to implement, our proposed method adopts a more elegant imputation-based strategy. This approach is advantageous as it reframes the competing risks problem into a pseudo-standard survival analysis, allowing for the direct application of well-established and computationally efficient PWE fitting procedures.

We conducted a comprehensive Monte Carlo simulation study to evaluate the performance of our proposed imputation-based PWE exponential model. The study was designed to assess the method's prediction accuracy and robustness across a wide range of clinically relevant scenarios, by varying the underlying data generating process, the censoring type (informative, non-informative) and the rate behind censoring mechanism, and the proportion of the primary event compared to the competing event(s).

## 7.1 Methods

### 7.1.1 Data-Generating Mechanisms

We simulate competing risks data under two primary data-generating models for the event of interest (event 1) and the competing event (event 2). Each subject has two binary covariates, $Z_1$ and $Z_2$, drawn from independent $Bernoulli(0.5)$ distributions.

**Proportional Sub-distribution Hazards (PSH) Model**

This first data-generating process (DGP) simulates right-censored time-to-event data under a competing risks framework, specifically adhering to the assumptions of the Fine and Gray proportional subdistribution hazards model (Fine and Gray 1999). The subdistribution hazard for the event of interest ($k = 1$) is specified as $\lambda_1(t|Z) = \lambda_{1,0}(t)\exp(\gamma' Z)$, which corresponds to a conditional Cumulative Incidence Function (CIF) of $F_1(t;Z) = P(T \leq t, E = 1|Z) = 1-[1-F_{1,0}(t)]^{\exp(\gamma' Z)}$. The baseline CIF, $F_{1,0}(t)$, is constructed from an improper log-logistic

distribution (i.e., a distribution whose density does not integrate to 1, allowing for a fraction of the population that never experiences the event) by scaling a proper log-logistic CDF by a baseline incidence probability, $p_{base}$. This choice facilitates the simulation of complex, non-monotonic baseline hazards.

The simulation process begins by generating a covariate matrix $Z$ of independent Bernoulli variables for a cohort of $n$ individuals. For each individual, the ultimate probability of experiencing event 1, $P(E_i = 1|Z_i) = 1 - (1 - p_{base})^{\exp(\gamma' Z_i)}$, is calculated. The event type $E_i$ is then assigned via a Bernoulli trial using this probability. Individuals not assigned to event 1 are designated to have the competing event ($E_i = 2$). Event times are subsequently generated conditional on the assigned event type using inverse transform sampling. For an individual destined for event 1, an event time $T_i$ is generated by inverting its conditional distribution, $P(T \leq t|E = 1, Z) = F_1(t; Z)/P(E = 1|Z)$, using a uniform random variate. For individuals assigned to the competing event, times are drawn independently from a separate Weibull distribution, creating a mixture model structure for the overall DGP.

**Non-Proportional Sub-distribution Hazards (NPSH) Model**

In this case we follow the method proposed by Beyersmann et al. (2009) . Data are generated by specifying the cause-specific hazards (CSH) for each event. The CSH for event 1, $h_1(t|Z)$, is modeled using a log-logistic distribution, while the CSH for event 2, $h_2(t|Z)$, is modeled using a Weibull distribution. The overall cumulative hazard, $H(t|Z) = H_1(t|Z) + H_2(t|Z)$, is computed, so that the overall event time $T$ for each subject is generated via inversion sampling by numerically solving $H(T|Z) + \log(U) = 0$ for a standard uniform random variable $U$, using the R `uniroot` function.

The cause of failure is then determined probabilistically based on the relative magnitude of the cause-specific hazards at the generated event time $T$, where $P(\text{Cause} = 1|T, Z) = h_1(T|Z)/(h_1(T|Z) + h_2(T|Z))$. For this model, the true log-hazard ratios for the CSH of event 1 are $\beta_{11} = -0.3$ for $Z_1$ and $\beta_{12} = 0.5$ for $Z_2$. For event 2, the coefficients are $\beta_{21} = -0.2$ for $Z_1$ and $\beta_{22} = 0$ for $Z_2$.

By generating proportional cause-specific hazards, we automatically simulate non-proportional sub-distribution hazards, this is due to the relationship between the CSH and the SDH well discussed in Putter, Schumacher, and Houwelingen (2020) .

### 7.1.2 Censoring mechanism

For each scenario, we target final censoring proportions of $\approx 30\%$, $\approx 45\%$, and $\approx 60\%$ by adjusting the rate parameter of the random censoring distribution and the administrative censoring time. In particular, we adjust the event-rate and the administrative censoring such that the final proportion of administratively censored observations is $\approx 15\%$. Then we adjust the exponential rate parameter of the censoring distribution to add random drop-outs to reach the desired levels of censored observations.

The mechanisms include:

- **Independent Censoring:** The censoring process is completely independent of subject covariates. Censoring times are drawn from a standard exponential distribution. These scenarios serve as a baseline to evaluate the method's performance under ideal censoring conditions.
- **Dependent Censoring:** To mimic more realistic and challenging settings, another scenario is designed to induce dependent censoring. Here, the probability of being censored was directly related to the same covariate, $Z_1$, that influenced the event risk. This creates a dependency between the event and censoring times, a known source of bias for standard survival estimators. This was achieved by making the censoring distribution's parameters a function of $Z_1$. For the exponential model, the rate parameter was defined as rate $\times \exp(0.3 \cdot Z_1)$.

### 7.1.3 Sample size

A key requisite of the design is to ensure an adequate number of event per parameter to obtain proper estimates of the cause specific cumulative incidence function. Although focused in a context without competing risks, the simulation study in Section 3.6 indicated that at least a number of 15-20 events per parameter are required. In this simulation, we aimed to reach a number of EPP between 15 and 20.

The total number of parameters to be estimated in the final piecewise exponential (PWE) model is determined by summing the degrees of freedom for the baseline hazard and the covariate effects. The model's baseline hazard was flexibly modeled using natural spline bases with 6 degrees of freedom (corresponding to 5 interior knots), whose locations are determined from a large preliminary simulated dataset of $N \geq 100000$. This level of flexibility was determined to be sufficient based on preliminary analyses of simulated data and represents a balance between capturing the potentially complex, non-monotonic shapes of the underlying hazard functions and avoiding overfitting, which could inflate the variance of the estimates. In addition to these 6 spline parameters, the model includes main effects for covariates $Z_1$ and $Z_2$, leading to 8 model parameters in total.

For the NPSH model, time-varying effects for the covariates are also included and modeled using a natural spline with a single interior knot at the median event time, to maintain a smooth functional form of the time-varying effect. Thus, in addition to 6 model parameters associated to the splines, the two parameters associated to the main covariate effect, additional four model parameters which specify the interaction of the covariate with the baseline hazard function add up to 12.

After determining the total number of parameters, and after identifying the rate parameters required to achieve the target censoring proportions, the initial sample size is scaled to ensure a minimum of primary events per estimated parameter. This adjustment accounts for the loss of events due to higher censoring and guarantees the stability of the model estimates across all conditions. For each scenario, 1000 replications are performed.

### 7.1.4 Proposed Estimation Method

The proposed imputation method is specifically designed to align with the risk set definition of a sub-distribution hazard. In a standard cause-specific analysis, a subject is removed from the risk set after experiencing a competing event. However, the sub-distribution hazard model requires such individuals to remain in the risk set to correctly model the crude probability of the event of interest. The imputation approach achieves this directly: by replacing the competing event time with an imputed, later censoring time, the subject contributes person-time to the risk set beyond their competing event, thereby correctly inflating the denominator of the hazard calculation in line with the sub-distribution hazard philosophy. The specific algorithm is as follows.

First, we fitted a model to the censoring distribution, considering only subjects who were censored for reasons other than the competing event. The choice of model depended on the censoring mechanism being simulated. In scenarios with independent censoring, we followed the original imputation methodology proposed by Ruan and Gray (2008), where the survival function for censoring was estimated non-parametrically using the Kaplan-Meier estimator. In scenarios with dependent censoring, a more flexible approach was utilized. We fitted a PWE model to the censoring distribution to account for covariate effects (Donoghoe and Gebski 2017), modeling the baseline hazard with a simple, smooth function, as the underlying censoring distributions were not expected to be complex.

We performed 10 imputations for each subject with a competing event. This number is commonly found to provide stable estimates for the parameters of interest while maintaining computational efficiency

Second, for each of the 10 imputation rounds, a potential censoring time was imputed for all subjects who experienced the competing event. This was done by drawing a random time from the conditional censoring distribution, given that the censoring event must occur after the observed competing event time ($T_C > T_{compete}$). This draw was performed via inversion sampling of the conditional survival function derived from the fitted model for censoring.

Third, using each of the 10 newly censoring–completed datasets, a final PWE model was fitted to the sub-distribution hazard of the primary event, using the pre-specified spline structure for the baseline hazard and time-varying effects as described in the design section. The final cumulative incidence function and model coefficients were obtained by averaging the predictions and estimates across the 10 imputations.

Algorithm 1 summarizes the estimation procedure.

Imputation and Modeling Procedure

1. **Model the Censoring Distribution:**
    - Fit a model to the time-to-censoring distribution using only subjects who were censored for reasons other than the competing event. The choice of model depends on the assumed censoring mechanism:
        - **(a) Independent Censoring:** Estimate the survival function for cen-

soring non-parametrically using the Kaplan-Meier estimator
  - **(b) Dependent Censoring:** Fit a Piecewise Exponential (PWE) model to the censoring distribution to account for covariate effects

2. **Impute Potential Censoring Times:**
   - For each of the **m = 10** imputation rounds:
     - For every subject who experienced the competing event, impute a potential censoring time by drawing a random value from the conditional censoring distribution.
     - This draw is conditioned such that the imputed censoring time ($T_C$) must occur after the observed competing event time ($T_{compete}$), i.e.,

$$T_C > T_{compete}$$

     - The random draw is performed using inversion sampling on the conditional survival function derived from the model fitted in Step 1.
   - This step yields 10 distinct "censoring-completed" datasets.
3. **Fit Final Models and Pool Results:**
   - Using each of the 10 completed datasets, fit a final PWE model to the sub-distribution hazard of the primary event. This model incorporates the pre-specified spline structure for the baseline hazard and any time-varying effects.
   - Combine the results across the 10 imputations to obtain the final estimates:
     - The final model **coefficients** are calculated by averaging the estimates from the 10 fitted models.
     - The final **cumulative incidence function (CIF)** is obtained by averaging the predictions from the 10 fitted models.

### 7.1.5 From model predicted rates to cause-specific cumulative incidence function

Details about the Piecewise Exponential Model framework are reported in Section 4.4. Briefly, the fitted Poisson GLM predicts the piecewise-constant (now sub-distribution) hazard rate, $\hat{\lambda}_j$, for each time interval $j$, allowing to obtain an overall smooth hazard function throughout the follow-up, if splinesare adopted. The cumulative sub-distribution hazard at time $t_k$ is then calculated as the sum of these hazards over the interval lengths, $\hat{H}^{sd}(t_k) = \sum_{j=1}^{k} \hat{\lambda}_j \Delta_j$. Finally, the cumulative incidence function is obtained via the transformation $\text{CIF}(t) = 1 - \exp(-\hat{H}(t))$.

### 7.1.6 Performance Evaluation

The performance of the method is evaluated by comparing the estimated CIF for event 1 against the true, analytically derived CIF at 121 time points from 0 to 120 months for the

PSH and 0 to 70 for the NPSH. The differences in the follow-up length are due to a difference intensity of the processes between PSH and NPSH.

The evaluation is conducted for all four covariate profiles ($Z_1, Z_2 \in \{0, 1\}$). The performance metrics used to calculate at each time point for each scenario are the Bias Equation 5.4, Variance Equation 5.5**,** Root Mean Squared Error (RMSE, Equation 5.8) and the Mean Absolute Error (MAE, Equation 5.6).

Moreover, as a benchmark, for the scenarios with non-informative censoring, we also consider a reanalysis of the same samples with the discrete-time sub-distribution hazard model and the transformation model on pseudo-observations computed on the Aalen-Johansen estimates of the CIF (Section 4.6) .

## 7.2 Results

### 7.2.1 *The Imputation-Based PWE Model Achieves Negligible Bias Across All Scenarios*

Figure 7.1 provides a comprehensive bias-variance decomposition of the Mean Squared Error (MSE) and clearly demonstrates the model's unbiasedness. Across all 12 scenarios, the squared bias (top row) remains close to zero and contributes minimally to the overall MSE, even under the challenging conditions of non-proportional hazards and 60% informative censoring. As expected for a cumulative estimand, the variance (middle row) is the primary driver of MSE, increasing with follow-up time. This pattern confirms that the imputation procedure correctly handles competing events without introducing systematic error into the CIF estimates.

### 7.2.2 *The PWE Model Demonstrates Robust and Competitive Performance in Estimating the CIF*

Overall, Table 7.1 shows that the performance of the PWE for the estimation of the CIF is reasonably good, regardless the complexity underlying the model for the outcome, and the presence of informative censoring and the proportion of censored observations. This means that the implementation of the imputation method within the piece-wise exponential model framework with splines is robust. This holds true also for the non-proportional hazards scenarios, where interaction terms with the baseline hazard specified with splines are specified. Considered the moderate sample size available (1000 N with >15-20 events per parameter), the RMSE and the MAE were always below the 0.02 for a baseline probability of 0.3, and around max 0.03 at baseline probability of 0.6.

The prediction accuracy of the PWE framework has revealed comparable with the ones of the other flexible methods, if not slightly better in specific cases especially in terms of the bias. The results of the formal comparison are shown in Table 7.2.

Collectively, these findings, presented in Figure Figure 7.1 and Table 7.1 and Table 7.2, support our central hypothesis: the proposed imputation-based PWE framework provides accurate and

robust estimates of the cause-specific CIF under a wide range of data-generating processes and censoring mechanisms.

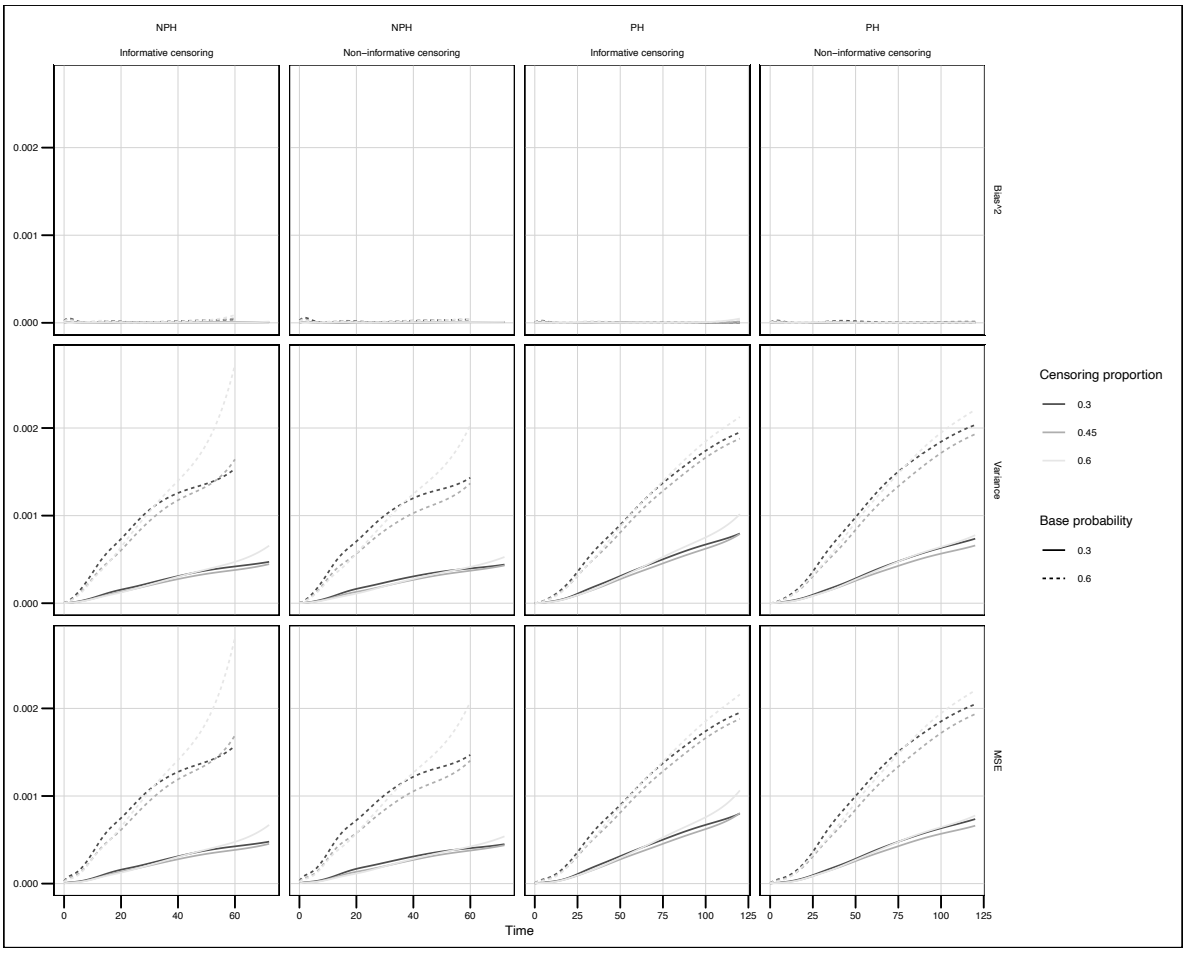



Figure 7.1: Bias-Variance Decomposition of the Mean Squared Error (MSE) for the Estimated Cumulative Incidence Function (CIF). The MSE (bottom row) and its components, squared bias (top row) and variance (middle row), are plotted against follow-up time in months. Results are aggregated across all four covariate profiles and are shown for scenarios under proportional (PH) and non-proportional (NPH) sub-distribution hazard assumptions, varying censoring proportions (30%, 45%, 60%), and baseline event probabilities (0.3, 0.6).

## 7.3 Discussion

The extensive simulation study demonstrates that this PWE-SDH framework provides accurate and robust estimates of the CIF. Across 12 diverse scenarios, encompassing both proportional and non-proportional sub-distribution hazards, varying degrees of censoring, and the challenging condition of covariate-dependent (informative) censoring, the proposed method

Table 7.1: RMSE and MAE of the CIF estimated with PWE model for the SDH. These are aggregated over covariates and time.

| Base Probability | Censoring % | RMSE | MAE |
|---|---|---|---|
| DGM 1: Proportional SDH - CM 1: Non-Informative Censoring | | | |
| 0.3 | 30% | 0.01890 | 0.01348 |
| 0.3 | 45% | 0.01790 | 0.01286 |
| 0.3 | 60% | 0.01886 | 0.01335 |
| 0.6 | 30% | 0.03340 | 0.02457 |
| 0.6 | 45% | 0.03162 | 0.02310 |
| 0.6 | 60% | 0.03333 | 0.02417 |
| DGM 1: Proportional SDH - CM 2: Informative Censoring | | | |
| 0.3 | 30% | 0.01956 | 0.01409 |
| 0.3 | 45% | 0.01874 | 0.01343 |
| 0.3 | 60% | 0.02035 | 0.01435 |
| 0.6 | 30% | 0.03220 | 0.02367 |
| 0.6 | 45% | 0.03110 | 0.02262 |
| 0.6 | 60% | 0.03257 | 0.02369 |
| DGM 2: Non-proportional SDH - CM 1: Non-Informative Censoring | | | |
| 0.3 | 30% | 0.01607 | 0.01186 |
| 0.3 | 45% | 0.01520 | 0.01109 |
| 0.3 | 60% | 0.01571 | 0.01133 |
| 0.6 | 30% | 0.02996 | 0.02254 |
| 0.6 | 45% | 0.02785 | 0.02075 |
| 0.6 | 60% | 0.03059 | 0.02247 |
| DGM 2: Non-proportional SDH - CM 2: Informative Censoring | | | |
| 0.3 | 30% | 0.01622 | 0.01190 |
| 0.3 | 45% | 0.01537 | 0.01130 |
| 0.3 | 60% | 0.01662 | 0.01198 |
| 0.6 | 30% | 0.03063 | 0.02294 |
| 0.6 | 45% | 0.02951 | 0.02189 |
| 0.6 | 60% | 0.03310 | 0.02402 |

Table 7.2: Comparison of the performance of the CIF estimation by other flexible method (the discrete-time sub-distribution hazard approach and the pseudo-observation approach) with PWE model for the SDH. The measures are aggregated over covariates and time.

| method | Censoring % | BIAS | VARIANCE | MSE | RMSE | MAE |
|---|---|---|---|---|---|---|
| DGM 1: Proportional SDH - Base probability: 0.3 | | | | | | |
| Discrete-time SDH | 30% | 0.00079 | 0.00035 | 0.00036 | 0.01894 | 0.01366 |
| PWE SDH | 30% | 0.00042 | 0.00036 | 0.00036 | 0.01890 | 0.01348 |
| Pseudo-observations | 30% | 0.00083 | 0.00041 | 0.00043 | 0.02062 | 0.01534 |
| Discrete-time SDH | 45% | 0.00141 | 0.00031 | 0.00032 | 0.01794 | 0.01304 |
| PWE SDH | 45% | 0.00110 | 0.00032 | 0.00032 | 0.01790 | 0.01286 |
| Pseudo-observations | 45% | 0.00147 | 0.00037 | 0.00038 | 0.01946 | 0.01436 |
| Discrete-time SDH | 60% | 0.00143 | 0.00035 | 0.00036 | 0.01890 | 0.01349 |
| PWE SDH | 60% | 0.00114 | 0.00035 | 0.00036 | 0.01886 | 0.01335 |
| Pseudo-observations | 60% | 0.00186 | 0.00042 | 0.00043 | 0.02074 | 0.01486 |
| DGM 1: Proportional SDH - Base probability: 0.6 | | | | | | |
| Discrete-time SDH | 30% | 0.00304 | 0.00108 | 0.00112 | 0.03341 | 0.02504 |
| PWE SDH | 30% | 0.00282 | 0.00111 | 0.00112 | 0.03340 | 0.02457 |
| Pseudo-observations | 30% | 0.00450 | 0.00125 | 0.00131 | 0.03619 | 0.02749 |
| Discrete-time SDH | 45% | 0.00202 | 0.00098 | 0.00101 | 0.03176 | 0.02355 |
| PWE SDH | 45% | 0.00184 | 0.00099 | 0.00100 | 0.03162 | 0.02310 |
| Pseudo-observations | 45% | 0.00316 | 0.00109 | 0.00112 | 0.03345 | 0.02510 |
| Discrete-time SDH | 60% | 0.00138 | 0.00110 | 0.00112 | 0.03344 | 0.02455 |
| PWE SDH | 60% | 0.00088 | 0.00111 | 0.00111 | 0.03333 | 0.02417 |
| Pseudo-observations | 60% | 0.00292 | 0.00127 | 0.00129 | 0.03588 | 0.02636 |
| DGM 2: Non-proportional SDH - Base probability: 0.3 | | | | | | |
| Discrete-time SDH | 30% | 0.00194 | 0.00022 | 0.00023 | 0.01512 | 0.01126 |
| PWE SDH | 30% | 0.00153 | 0.00025 | 0.00026 | 0.01607 | 0.01186 |
| Pseudo-observations | 30% | 0.00132 | 0.00023 | 0.00024 | 0.01556 | 0.01178 |
| Discrete-time SDH | 45% | 0.00185 | 0.00019 | 0.00020 | 0.01410 | 0.01039 |
| PWE SDH | 45% | 0.00153 | 0.00023 | 0.00023 | 0.01520 | 0.01109 |
| Pseudo-observations | 45% | 0.00120 | 0.00020 | 0.00021 | 0.01452 | 0.01082 |
| Discrete-time SDH | 60% | 0.00151 | 0.00019 | 0.00020 | 0.01420 | 0.01033 |
| PWE SDH | 60% | 0.00151 | 0.00024 | 0.00025 | 0.01571 | 0.01133 |
| Pseudo-observations | 60% | 0.00120 | 0.00022 | 0.00022 | 0.01488 | 0.01087 |
| DGM 2: Non-proportional SDH - Base probability: 0.6 | | | | | | |
| Discrete-time SDH | 30% | 0.00284 | 0.00085 | 0.00089 | 0.02989 | 0.02287 |
| PWE SDH | 30% | 0.00260 | 0.00088 | 0.00090 | 0.02996 | 0.02254 |
| Pseudo-observations | 30% | 0.00202 | 0.00100 | 0.00105 | 0.03235 | 0.02459 |
| Discrete-time SDH | 45% | 0.00234 | 0.00074 | 0.00077 | 0.02775 | 0.02100 |
| PWE SDH | 45% | 0.00272 | 0.00076 | 0.00078 | 0.02785 | 0.02075 |
| Pseudo-observations | 45% | 0.00294 | 0.00093 | 0.00098 | 0.03127 | 0.02320 |
| Discrete-time SDH | 60% | 0.00146 | 0.00089 | 0.00092 | 0.03029 | 0.02252 |
| PWE SDH | 60% | 0.00263 | 0.00092 | 0.00094 | 0.03059 | 0.02247 |
| Pseudo-observations | 60% | 0.00533 | 0.00139 | 0.00150 | 0.03867 | 0.02688 |

exhibited negligible bias. The variance of the CIF estimates was the main component of the MSE. It increased over the follow-up time, as it is expected for cumulative estimands, but overall remained well-controlled, resulting in low overall Mean Squared Error.

Furthermore, when benchmarked against established flexible methods—namely, discrete-time SDH models and regression on pseudo-observations—the PWE-SDH approach demonstrated comparable, and in some cases slightly superior, performance, solidifying its position as a powerful tool for competing risks analysis.

The transformation model on pseudo-observation uses jackknife-based pseudo-values of the non-parametric Aalen-Johansen CIF estimate as the outcome in a generalized estimating equation, allowing for direct regression on the CIF. The simulation results show the PWE-SDH is highly competitive with this approach. Conceptually, the pseudo-observation approach is a form of imputation or better a data transformation technique, as it transforms censored data into complete data. As the pseudo-observation method is a two-stage process (non-parametric estimation followed by regression), also the imputation-based PWE model is an integrated, likelihood-based approach (post-imputation).

The discrete time SDH model uses a weighted binary regression approach. This weights are computed considering an IPCW scheme, before the modeling step. Thus the approach also consists in a two stage modeling procedure. The PWE-SDH model can be viewed as a continuous-time analogue. The choice between these two frameworks may depend on the nature of the data collection process; discrete-time models are a natural fit for data collected at fixed intervals (e.g., scheduled follow-up visits), while the continuous-time PWE model is well-suited for data where event times are recorded with high precision. The comparable performance observed in the simulations suggests that both are valid and powerful tools in the analyst's arsenal.

For modeling the censoring distributions, we used different approaches depending on the censoring mechanism underlying the data: a non-parametric KM estimator in the presence of non-informative censoring and a PWE model specifying the baseline censoring hazard with splines in the presence of informative censoring.

The validity of the imputation-based approach hinges on the correct specification of the model for the censoring distribution. A critical and logical next step for this research is to formally evaluate the model's robustness to misspecification of the censoring distribution. Future simulation studies should be designed to quantify the magnitude of bias introduced when, for instance, a simple non-parametric Kaplan-Meier estimator is used for imputation when the true censoring process is strongly covariate-dependent. The results of such studies would be invaluable for establishing practical guidelines for applied researchers on how flexibly the censoring model must be specified to ensure valid inference.

A pooled model for the censoring distribution, in cases where the censoring is covariate-dependent, could lead to greater bias but have lower variances associated to the estimates of the CIF. This was already encountered in Donoghoe and Gebski (2017) for the estimates of the sub-distribution hazard ratio, when evaluating weighting schemes for the Fine & Gray proportional sub-distribution hazards model.

Moreover, while no particular differences in the performance were found between informative and non-informative censoring scenarios, indicating that the PWE model was efficient in imputing censoring times, an open question is whether the KM estimator or the Cox model could be also sufficient at this task. A PWE model with splines, might overfit the data, especially with limited censoring proportion and when considering non-proportional hazards scenarios ( arguably possible also for the censoring distributions). Equivalently to covariate information, considering multiple parameters for smoothing the censoring distribution might eventually increase the variance of the CIF estimates. In other words, a question that we should address in the future is how flexible the censoring model should be for a stable estimation of the CIF.

The performance of the PWE-SDH model with smaller sample sizes, a larger number of covariates (including continuous or multi-category predictors), or in high-dimensional settings remains also to be explored. The data expansion inherent in PWE models, which creates one row of data for each interval a subject is at risk, can lead to very large datasets. This may present computational challenges in scenarios with many covariates or long follow-up times partitioned into many intervals.

Future studies will assess the proposed method in the presence of different sample sizes and multidimensional data and they will try to address the issue of the flexibility of the censoring model behind the imputation of the censoring times.

In conclusion, the proposed imputation-based PWE-SDH model is a novel, flexible, and robust method for the analysis of competing risks data. By marrying the flexibility of PWE models with an accessible imputation framework, it provides a powerful tool for directly modeling the CIF, accommodating non-proportional hazards, and handling covariate-dependent censoring. We think tath the approach is also easily extendible to more advanced non-linear models (Fornili et al. 2018). Its strong performance in this preliminary simulation study and its implementation using standard statistical software position it as a valuable addition to the biostatistician's toolkit for modeling in the presence of competing risks.

With this study we concluded the part of thesis dedicated to the assessment of the performances and the introduction of new solutions to flexibly model cumulative quantities, critical to obtain clinically useful and causally sound measure of effect associated to an exposure. However, a critical question remain: can we exploit better the estimated cumulative quantities to communicate risk of an event associated to an exposure? This will be the aim of the following chapter.

# 8 Highest Risk Density Region for the communication of the impact of a treatment covariate on the time-to-event distribution

Building on the established need for more communicable measures of treatment effect beyond the Hazard Ratio, this chapter addresses the inherent limitations of existing time-scale alternatives like the Restricted Mean Survival Time (RMST) and quantile regression. We therefore shift the focus from single-summary statistics to a more granular analysis of the entire time-to-event distribution. To formally quantify how a treatment redistributes risk over time, we introduce two novel concepts: the Highest Risk Density Region (HRDR), which identifies the narrowest time interval of peak risk, and the Highest Net Risk Difference Region (HNRDR), which isolates the period of maximal therapeutic impact. These tools provide a complementary framework for answering clinically relevant questions about when and how a treatment exerts its effect.

## 8.1 Methods

### 8.1.1 Formalizing the Causal Restricted HRDR Estimand

Not all individuals will experience an event by the end of the follow-up time $\tau$, making some time-to-event data right censored. Unless the adoption of a parametric survival model (e.g., AFT), obtaining estimates of the $f_a(t|t > \tau)$ and $F_a(t|t > \tau)$ is impossible without extrapolation. For this reason, we restrict the support of $T(a)$ to $[0, \tau]$ where $\tau$ represents a relevant ending time-point of the follow-up.

Let $T(a)$ be the original, unconditional time-to-event random variable with PDF $f_a(t)$ and CDF $F_a(t)$. We define a new, truncated random variable, $T^*(a) = T(a)|T(a) \leq \tau$, whose distribution is that of $T(a)$ conditional on the event $T(a) \leq \tau$. The support for $T^*(a)$ is the interval $[0, \tau]$. For notational simplicity, the term $(a)$ and the subscript $_a$ will be omitted.

The properties of $T^*$ can be derived from first principles using the definition of conditional probability. The CDF of $T^*$, denoted $F^*(t)$, is the probability that $T^*$ is less than or equal to some time $t$ within its support. For any $t \in [0, \tau]$:

$$F^*(t) = P(T^* \leq t) = P(T \leq t|T \leq \tau) \tag{8.1}$$

By the definition of conditional probability, this is:

$$F^*(t) = \frac{P(T \leq t \cap T \leq \tau)}{P(T \leq \tau)}$$

Since $t \leq \tau$, the event $\{T \leq t\}$ is a subset of the event $\{T \leq \tau\}$, so their intersection is simply $\{T \leq t\}$. Therefore, the expression simplifies to:

$$F^*(t) = \frac{P(T \leq t)}{P(T \leq \tau)} = \frac{F(t)}{F(\tau)} \quad \text{for } 0 \leq t \leq \tau \tag{8.2}$$

For $t > \tau$, $F^*(t) = 1$. The restricted PDF, $f^*(t)$, is the derivative of this conditional CDF with respect to $t$:

$$f^*(t) = \frac{d}{dt}F^*(t) = \frac{d}{dt}\left(\frac{F(t)}{F(\tau)}\right) = \frac{1}{F(\tau)}\frac{dF(t)}{dt} = \frac{f(t)}{F(\tau)} \quad \text{for } 0 \leq t \leq \tau \tag{8.3}$$

For $t > \tau$, $f^*(t) = 0$.

The restricted PDF, $f^*(t)$, is the original unconditional PDF, $f(t)$, rescaled by the constant $1/F(\tau)$. This constant is the inverse of the probability of an individual belonging to the sub-population of interest, and it serves as the normalizing constant that ensures $\int_0^\tau f^*(u)du = 1$.

From these, we define the potential restricted PDF, $f_a^*(t)$, which describes the distribution of event times for the sub-population that *would have* an event by time $\tau$ if they received treatment $a$:

$$f_a^*(t) = \frac{f_a(t)}{F_a(\tau)} \quad \text{for } t \in [0, \tau] \tag{8.4}$$

This is the PDF for the truncated potential outcome variable $T^*(a) = T(a)|T(a) \leq \tau$.

### 8.1.2 Highest Risk Density Regions (HRDRs)

Building upon Hyndman 1995, we define the $(1-\alpha)100\%$ HRDR estimand, denoted $\Psi_{HRDR}(\tau, \alpha)$, as the pair of highest density regions for the potential outcome distributions under treatment and control:

$$\Psi_{RHDR}(\tau, \alpha) = (R_{1,\alpha}, R_{0,\alpha}) \tag{8.5}$$

where, for each treatment arm $a \in \{0, 1\}$, the region $R_{a,\alpha}$ is defined as:

$$R_{a,\alpha} = \{t \in [0, \tau] : f_a(t) \geq k'_{a,\alpha}\} \tag{8.6}$$

The threshold $k'_{a,\alpha}$ is the largest value such that the integral of the *unconditional* marginal PDF over this region equals a pre-defined proportion of the probability mass within the truncated population:

$$\int_{R_{a,\alpha}} f_a(t)dt = (1-\alpha)F_a(\tau) \tag{8.7}$$

This version of HRDR answers specifically the question of how the treatment changes the timing of the most likely $(1-\alpha)100\%$ of the failures of the sub-population of individuals who fail by time $\tau$. This method cleanly focuses on the timing of events for *those who fail*, captured by comparing the $(1-\alpha)100\%$ HRDRs of the two normalized conditional distributions. This is a question about the timing of risk, normalized for any difference in the overall event rate by time $\tau$.

However, another relevant question would be how treatment changes the specific time-window during which a fixed absolute probability mass (i.e, risk) of failures is most likely to accumulate for the sub-population of individuals *who are followed* by time $\tau$. This HRDR version compares the periods of highest risk of the two groups, accounting for the same absolute probability of events that we will call $P_{mass}$.

Suppose that the research question involves the characterization of a time-window in which 20% of the patients are mostly expected to experience the event. Panel h of Figure 1 indicates that for a 20% risk mass ($\alpha = 0.2$), the HRDR (solid light line) for the control group is the interval $[26, 45]$ months, whereas for the treatment group (solid dark line), it is $[43, 74]$ months. This approach translates the treatment effect into a quantifiable shift and extension of the interval of maximal event concentration. This is a question about the timing of absolute risk accumulation in the population as a whole.

Naturally, $P_{mass}$ must be less than the total probability of an event occurring by time $\tau$ in either group: $0 < P_{mass} < \min(F_0(\tau), F_1(\tau))$. For each treatment arm $a \in \{0, 1\}$, we would calculate the required coverage level, $(1-\alpha_a)$, that yields this fixed mass when multiplied by the group's total event probability, $F_a(\tau)$.

$$(1-\alpha_a) = \frac{P_{mass}}{F_a(\tau)} \tag{8.8}$$

if $F_0(\tau) \neq F_1(\tau)$, then the required 'coverage levels' $\alpha_0$ and $\alpha_1$ will be different.

For each group, we would then find the standard $(1-\alpha_a)\%$ HRDR of its restricted distribution, $f^*_a(t)$. This means finding the region $R^{FM}_{a,P_{mass}}$ such that:

$$R^{FM}_{a,P_{mass}} = \{t \in [0, \tau] : f_a(t) \geq k'_{a,\alpha_a}\} \tag{8.9}$$

where the threshold $k'_{a,\alpha_a}$ is chosen so that the integral of the unconditional marginal PDF over this region is exactly equal to the target mass, $P_{mass}$:

$$\int_{R^{FM}_{a,P_{mass}}} f_a(t)dt = (1-\alpha_a)F_a(\tau) = P_{mass} \tag{8.10}$$

While the first version of the HRDR is better suited for studying the pure temporal dynamics of the risk density, independent of differences in cumulative incidence, the AR-HRDR is arguably more clinically relevant for communicating absolute risk to clinicians, patients and stakeholders. For this reason, in the paper we will focus on this second version.

### 8.1.3 The Highest Net Risk Difference Region

Up to now we have considered intervals that describe how risk of an event is distributed over the follow-up time but we did not address methods to characterize the *difference* between treatments. A naive initial approach would be to compare the interval features, such as interval length, interval mid-point or interval boundaries by calculating ratios or differences.

By comparing the difference or the ratio between the mid-points or the lengths of the intervals one could isolate the effect of treatment on the location or the scale of the distribution. Notably, a ratio between HRDR features would be conceptually similar to the interpretation of the coefficients of the AFT models.

Assuming no cure fraction, a treatment or exposure does not eliminate risk but rather shifts its distribution over time. As said before, within a finite follow-up period, a beneficial intervention is one that effectively postpones a quantum of risk to a later time. This perspective gives rise to two fundamental questions.

One first question would be what the cumulative magnitude of the risk shift (i.e., the absolute risk reduction or increase) attributable to the treatment or exposure over the course of the follow-up is. This is usually addressed by analyzing the difference between the cumulative risk or cumulative survival functions of the groups, where the maximum difference quantifies the maximal net shift in risk mass.

Another relevant question would be what the temporal dynamics of this risk shift is. That is, how is the risk shift realized over time—as an acute, high-impact effect over a short duration, or as a sustained, moderate effect over a longer period.

Thus, a more direct characterization of the net treatment effect, than the comparison of the HRDR, can be obtained by analyzing the instantaneous risk difference function, i.e., the difference between the two PDFs. This difference function quantifies the instantaneous risk reduction at each point in time, identifying periods of maximal treatment efficacy.

The instantaneous risk difference function is defined as $d^*(t) = f_1^*(t) - f_0^*(t)$. The value of $d^*(t)$ at any time point represents the instantaneous net rate of change in event probability due to the treatment. Where $d^*(t) < 0$, the treatment is preventing events that would have occurred in the control group at that time. Where $d^*(t) > 0$, the treatment is causing excess events relative to the control group, by shifting events that would have occurred earlier, later.

Formally, we define the HNRDR as the set of time points where $d(t)$ is most intense, such that the integral of $d(t)$ over this set equals a pre-specified absolute net risk difference, $C$. For a net benefit (negative $C$), the region is:

$$R_C = \{t \in [0, \tau] : d(t) \leq k_C\} \quad \text{such that} \quad \int_{R_C} d(t)dt = C \tag{8.11}$$

where $k_C$ is the data-driven threshold constructed with the same methodology considered for the computation of the HRDRs.

While the cumulative risk difference, D(t), or the difference in RMST, $\Delta RMST(\tau)$, provide the overall magnitude of the net treatment effect, the HNRDR uniquely identifies the temporal location where that effect is most intensely realized. The HNRDR provides a framework for what can be conceptualized as "risk-shifting accounting." An important property is that the total integral of $d(t)$ over the entire time horizon is zero ($\int_0^\infty d(t)dt = 0$). This forces a "zero-sum" narrative of risk redistribution. However, up to a restricted time $\tau$ this will integrate to the $C = F_1(\tau) - F_0(\tau)$. The HNRDR for a net benefit (a negative cumulative difference, C) finds the interval where d(t) is most negative. The corresponding positive portions of d(t) must therefore represent the events that were postponed from this early period to a later time.

### 8.1.4 Methods of constructing Highest Risk Density Regions and Highest Net Risk Difference Regions

Having formally defined the HRDR and HNRDR estimands, we now detail the computational algorithms for their estimation from empirical data.

**HRDRs**

Various method of finding HDRs exist in the literature. In this paper, we will consider two main approaches for the construction of the HRDRs and HNRDRs: an up-down approach, based on the probability density function, and a right-left approach, based on the cumulative distribution function of the time-to-event variable. Both methods have their advantages and drawbacks and are susceptible to some sort of approximation. These algorithms are built on the work of O'Neill (2021), which however considered parametric models estimates of the PDF and CDF.

**Up-down approach**

In the up-down approach If $\hat{f}_a(t)$ is available analytically (from AFT parametric models) or as a smooth estimate (e.g., from a flexible parametric fit like a PWE-spline model), the primary task is to find the level of $f_a$ which defines the HRDR. This often requires numerical integration and root-finding techniques to solve the equation.

A common practical approach, particularly when working with empirical data, is Hyndman's density quantile algorithm. If one can evaluate the estimated density $\hat{f}_a(t_i)$ at a large number of points $t_i$, these density values can be sorted. The $(1-\alpha)F_a(t)$ quantile of these sorted density values can serve as an approximation for $f_\alpha$. Alternatively, if random samples can be

drawn from the distribution characterized by $\hat{f}_a(t)$, one can evaluate the density at these $n$ samples; the $\lfloor(1-\alpha)n\rfloor$-th largest density value among these provides an estimate of $f_\alpha$. In the present work we will consider the Hyndman's density quantile algorithm, also called the "up-down" approach.

Table 8.1: HRDR via "Up-down" Approach

| Step | Description |
|---|---|
| **Foundation** | Operates on the Probability Density Function (PDF), $\hat{f}(t)$. It is general and can identify multiple, non-contiguous high-risk intervals. |
| **1. Evaluate Density** | Estimate the density function $\hat{f}(t)$ and evaluate it at a large number of discrete time points, $t_1, t_2, ..., t_n$, to get a vector of density values $\hat{f}(t_i)$. |
| **2. Sort Density Values** | Create a new list containing the density values $\hat{f}(t_i)$ sorted in descending order. |
| **3. Determine Density Threshold** | Find the density value that corresponds to the $(1-\alpha)$-th quantile of the sorted density values. This value serves as the density threshold, $c_\alpha$. |
| **4. Identify Time Intervals** | The HRDR is the set of all original time points $t_i$ where the unevaluated density $\hat{f}(t_i)$ is greater than or equal to the threshold $c\_\alpha$. $\text{HRDR}_\alpha = t \mid \hat{f}(t) \geq c_\alpha$ |
| **Output** | One or more time intervals representing the periods of highest concentrated risk. |

**Right-left approach**

The right-left approach works under the assumption of a compact set and a quasi-convex PDF, which is often the case of survival data. In such cases, non-parametric or flexible parametric estimates of $\widehat{F}_a(t)$ can be exploited. A HRDR is defined as a compact interval $I = [a, b]$ that contains a specific risk mass, $P_{mass}$, while having the minimum possible length, $L$. The probability mass requirement is given by $F(b) - F(a) = P_{mass}$ . We can reframe the search for the HRDR as an optimization problem with a single unknown variable. The goal is to minimize the interval length, $L = b - a$. First, we express $b$ as a function of $a$. From the probability mass requirement, we have $F(b) = F(a) + P_{mass}$. By applying the inverse cumulative distribution function (the quantile function, $F^{-1}$), we get

$$b = F^{-1}(F(a) + 1 - \alpha) \tag{8.12}$$

this into the length formula gives us the function we need to minimize with respect to $a$:

$$L(a) = F^{-1}(F(a) + 1 - \alpha) - a \tag{8.13}$$

An optimization algorithm can be used to find the value of $a$ that minimizes this function. The search for the optimal $a$ can be restricted to the interval $[0, F^{-1}(\alpha)]$. This optimization strategy is implemented using discrete data points. It begins by taking vectors of time points,

$t_i$, and the corresponding CDF values. First, the algorithm creates continuous, invertible functions from the discrete data points using linear interpolation. This yields an empirical CDF, which we can denote as $\hat{F}(x)$, and its inverse, the empirical quantile function, $\hat{F}^{-1}(p)$.

Next, it defines an objective function that calculates the interval length, $L(a)$, for any given lower bound $a$, exactly as formulated in the continuous problem:

$$L(a) = \hat{F}^{-1}(\hat{F}(a) + 1 - \alpha) - a \tag{8.14}$$

The algorithm then numerically searches for the value of $a$ within a valid search interval that minimizes the objective function, $L(a)$. The result of this optimization is the optimal lower bound, $a^*$.

Finally, the optimal upper bound, $b^*$, is calculated by substituting $a^*$ back into the rearranged constraint equation:

$$b^* = \hat{F}^{-1}(\hat{F}(a^*) + 1 - \alpha) \tag{8.15}$$

The algorithm concludes by returning the interval $[a^*, b^*]$ as the shortest region containing the specified probability coverage.

Table 8.2: HRDR via "Right-Left" Approach (Optimization)

| Step | Description |
|---|---|
| **Foundation** | Operates on the Cumulative Distribution Function (CDF), $\hat{F}(t)$. It is more stable but assumes the HRDR is a single, contiguous interval. |
| **1. Interpolate Empirical CDF** | From the discrete time-to-event data, create a continuous and invertible empirical CDF, $\hat{F}(t)$, and its inverse (the quantile function), $\hat{F}^{-1}(p)$. |
| **2. Define & Minimize Objective Function** | Define an objective function $L(a)$ that calculates the interval length for a given lower bound a: $L(a) = \hat{F}^{-1}(\hat{F}(a) + 1 - \alpha) - a$ Use a numerical optimization algorithm to find the value $a^*$ that minimizes $L(a)$. |
| **3. Calculate Upper Bound** | Calculate the optimal upper bound, $b^*$, by substituting $a^*$ back into the constraint equation: . $b^* = \hat{F}^{-1}(\hat{F}(a^*) + 1 - \alpha)$ |
| **4. Define Final Interval** | The resulting HRDR is the single, shortest interval $[a^*, b^*]$ that contains the probability mass $(1 - \alpha)$. |
| **Output** | A single time interval representing the shortest span for the specified risk mass. |

**HNRDRs**

**Up-down approach**

The up-down and right-left approaches can also be utilized to identify the HNRDRs from the instantaneous risk difference function and the cumulative risk difference function respectively. In the first case, we assume the function $d(t)$ has been evaluated at $n$ equally spaced time points, $t_1, t_2, \ldots, t_n$, with a constant time step $\Delta t = t_{i+1} - t_i$. This gives us a vector of function values, $\mathbf{d} = (d_1, d_2, \ldots, d_n)$, where $d_i = d(t_i)$.

The algorithm begins by sorting these discrete values to identify the most extreme ones. Let $\pi$ be a permutation of the indices $\{1, 2, \ldots, n\}$ that orders the values in $\mathbf{d}$ from most to least extreme, depending on the direction of the effect.

Next, the algorithm finds a threshold index, $k$, by applying the integral constraint. It approximates the integral by calculating the cumulative sum of the sorted values multiplied by the time step, $\Delta t$. It then identifies the smallest integer $k$ for which this cumulative sum reaches the target value $C$. For a risk reduction, this is formulated as finding the minimum $j$ such that:

$$k = \min \left\{ j \in \{1, \ldots, n\} \;\middle|\; \sum_{i=1}^{j} d_{\pi(i)} \Delta t \leq C \right\} \tag{8.16}$$

The threshold value, $d_{thresh}$, is then defined as the function value at this $k$-th sorted position, $d_{thresh} = d_{\pi(k)}$.

Finally, the HNRDR is identified by constructing a set of indices, $I_{HNRDR}$, which includes all of the original time points where the function value $d_i$ is more extreme than or equal to this threshold. For a risk reduction, this set is:

$$I_{HNRDR} = \{i \in \{1, \ldots, n\} \mid d_i \leq d_{thresh}\} \tag{8.17}$$

The algorithm concludes by converting any contiguous blocks of indices within $I_{HNRDR}$ back into time intervals, $[t_{start}, t_{end}]$, which constitute the final HNRDR.

Table 8.3: HNRDR via "Up-Down" Approach

| Step | Description |
|---|---|
| **Foundation** | Operates on the instantaneous risk difference function, $d(t)$, evaluated at discrete time points. |
| **1. Evaluate Risk Difference** | Obtain a vector of risk difference values $d_i = d(t_i)$ by evaluating the function at n equally spaced time points. |
| **2. Sort by Extremity** | Sort the indices $\{1, ..., n\}$ based on the values of di from most to least extreme (e.g., from most negative for a risk reduction). Let this sorted order be $\pi(i)$. |
| **3. Find Threshold via Cumulative Sum** | Approximate the integral by calculating the cumulative sum of the sorted values. Find the smallest index $k$ where this sum reaches the target cumulative difference, $C$. For a risk reduction: $k = \min \left\{ j \mid \sum_{i=1}^{j} d_{\pi(i)} \Delta t \leq C \right\}$ The threshold value is then $d_{thresh} = d_{\pi(k)}$. |

| Step | Description |
|---|---|
| **4. Identify Time Intervals** | The HNRDR is the set of all original time points $t_i$ where the risk difference $d_i$ is more extreme than or equal to the threshold $d_{thresh}$. |
| **Output** | One or more time intervals representing periods of the greatest net risk difference. |

**Right-left approach**

In the right-left approach, if the underlying difference function $d(t)$ is expected to be unimodal, also the HNRDRs would be compact sets. In such cases flexible parametric estimates of $\hat{D}(t)$, i.e., the cumulative risk difference function, can also be exploited. The HNRDR is defined as a compact interval $I = [a, b]$ that contains a specific net risk difference mass, $C_{mass}$, while having the minimum possible length, $L$. The probability mass requirement is given by $D(b) - D(a) = C_{mass}$ . We can reframe the search for the HNRDR as an optimization problem with a single unknown variable. The goal is to minimize the interval length, $L = b-a$. First, we express $b$ as a function of $a$. From the net risk difference mass requirement, we have $D(b) = D(a) + C_{mass}$. By applying the inverse of the cumulative risk difference function, we get

$$b = D^{-1}(D(a) + 1 - \alpha) \tag{8.18}$$

Substituting this into the length formula gives us the function we need to minimize with respect to $a$:

$$L(a) = D^{-1}(D(a) + 1 - \alpha) - a$$

. In order to find the inverse of the cumulative risk difference function, it is important to restrict it up to the maximum net risk difference value, to make it monotonic. An optimization algorithm can be used to find the value of $a$ that minimizes this function. The search for the optimal $a$ can be restricted to the interval $[0, D^{-1}(\alpha)]$. This optimization strategy is again achieved using discrete data points. It begins by taking vectors of time points, $t_i$, and the corresponding cumulative risk difference values, $D(t)$.

First, the algorithm creates continuous, invertible functions from the discrete data points using linear interpolation. This yields an empirical cumulative risk difference function, which we can denote as $\hat{D}(t)$, and its inverse $\hat{D}^{-1}(p)$.

Next, it defines an objective function that calculates the interval length, $L(a)$, for any given lower bound $a$, exactly as formulated in the continuous problem:

$$L(a) = \hat{D}^{-1}(\hat{D}(a) + C) - a \tag{8.19}$$

The algorithm again searches for the value of $a$ within a valid search interval that minimizes the objective function, $L(a)$. The result of this optimization is the optimal lower bound, $a^*$.

Finally, the optimal upper bound, $b^*$, is calculated by substituting $a^*$ back into the rearranged constraint equation:

$$b^* = \hat{D}^{-1}(\hat{D}(a^*) + C) \tag{8.20}$$

The algorithm concludes by returning the interval $[a^*, b^*]$ as the shortest region containing the specified cumulative difference, $C$.

Table 8.4: HNRDR via "Right-Left" Approach (Optimization)

| Step | Description |
|---|---|
| **Foundation** | Operates on the cumulative risk difference function, $\hat{D}(t)$. Assumes the HNRDR is a single, contiguous interval. |
| **1. Interpolate Empirical Function** | Create continuous and invertible functions from the discrete cumulative risk difference data: $\hat{D}(t)$ and its inverse $\hat{D}^{-1}(p)$. The domain of $\hat{D}(t)$ may need to be restricted to ensure it is monotonic for inversion. |
| **2. Define & Minimize Objective Function** | Define an objective function $L(a)$ for the interval length based on a lower bound a: $L(a) = \hat{D}^{-1}(\hat{D}(a) + C) - a$ Use numerical optimization to find the value $a^*$ that minimizes $L(a)$. |
| **3. Calculate Upper Bound** | Calculate the optimal upper bound, $b^*$, using the value of $a^*$: $b^* = \hat{D}^{-1}(\hat{D}(a^*) + C)$ |
| **4. Define Final Interval** | The resulting HNRDR is the single, shortest interval $[a^*, b^*]$ that contains the target cumulative risk difference, $C$. |
| **Output** | A single time interval representing the shortest span for the specified net risk difference. |

The up-down (PDF-based) approach is more general, as it can identify non-contiguous regions, which is a primary advantage of HDRs for multimodal distributions. However, its performance is highly sensitive to the quality of the PDF estimate, which often involves numerical differentiation—a potentially unstable operation that can amplify noise.

The right-left (CDF-based) approach is generally more stable because it operates on the integrated function (CDF), which is inherently smoother than the PDF. Its main limitation is the implicit assumption that the HRDR is a single contiguous interval, which may not be true for complex, multimodal risk patterns (though this is rare in survival analysis).

Table 8.1, Table 8.2, Table 8.3, Table 8.4 summarize the algorithms to obtain the probability intervals of interest.

### 8.1.5 Finding the best estimator of $f_a(t)$ and $F_a(t)$

Regardless of the version of the HRDRs considered and the approach to identify them, their reliability is intrinsically linked to the quality of the underlying PDF and CDF estimate. Thus, finding the best estimator $\hat{f}_a(t)$ and $\hat{F}_a(t)$ is critically important to present coherent intervals in which a specific probability mass is expected to fall.

In the context of right-censored and often highly skewed distributions, the modeling approaches have been mainly focused either on the hazard function or the cumulative distribution function, although powerful parametric AFT models like the Weibull, the log-logistic or the gamma have been also considered in modeling time-to-event data. These are very efficient in retrieving analytically the functions, but they are often too rigid to model follow-up data of clinical and epidemiological studies, as they assume that the hazard, the PDF and the CDF follow a specific, predefined mathematical form.

On the other hand, as already discussed before, the semi-parametric Cox proportional hazards (PH) model does not specify the baseline hazard function, but it relies on the stringent assumption of proportional hazards.

Three interrelated functions describe the distribution of a lifetime variable: the survival function, the hazard rate function, the probability density function. Knowing any one of these allows the other three to be uniquely determined. The interconvertibility of the functions that define the random variable implies that modeling choices can be guided by interpretational needs or analytical convenience.

Several flexible approaches were developed either for modeling the hazard function or the cumulative distribution function of the time-to-event conditional on covariate, where flexibility means being able to account for non-trivial shapes of the baseline hazard/cumulative distribution functions, time-varying covariates and most importantly to relax the assumption of proportionality of the hazards/cumulative distribution functions.

In the present chapter, the piece-wise exponential (PWE) model on $h(t)$ and the transformation model on jackknife pseudo-observations of $F(t)$ will be considered. The choice of the PWE and Pseudo-observation models was aimed at an exploration of two different philosophies in flexible survival modeling.

The PWE estimates the hazard (a rate, akin to a derivative) and then obtains the CDF via integration, which is a smoothing operation. PWE might be superior for PDF estimation, thus, its preferred algorithms of HRDR and HNRDR construction should be the one based on the Up-Down approach, described in Table 8.1 and Table 8.3. On the contrary, the pseudo-observation model directly targets the survival or cumulative distribution function, bypassing the hazard entirely and to obtain a PDF a numerical derivative has to be computed. Numerical differentiation is known to amplify noise. Therefore, it is logical that the Pseudo-observation model would produce a less stable PDF estimate . The model naturally pairs with the Right-Left algorithm described in Table 8.2 and Table 8.4.

### 8.1.6 Visualizing and quantifying the uncertainty around the estimates of the target functions

To quantify the uncertainty across the entire range of the estimated functions (CDF and/or PDF), we construct $(1 - \alpha) \times 100\%$ simultaneous confidence bands (SCBs) using a non-parametric bootstrap approach. The procedure is performed independently for each function and at each specified covariate level. Two distinct bootstrap-based methods were developed to generate the SCBs, following an approach from (Montiel Olea and Plagborg-Møller 2018).

The foundational step for both methods involves generating $B$ bootstrap samples by resampling with replacement from the original dataset. For each bootstrap sample, the target function (e.g., the cumulative distribution function, $F(y)$) is re-estimated over a discrete grid of $k$ points, $\{t_1, t_2, \ldots, t_k\}$. This process yields $B$ bootstrap curves, which can be organized into a $B \times k$ matrix, where each row represents a full estimated curve. The function estimated from the original data serves as the central point estimate, $\hat{f}(t)$ or $\hat{F}(t)$ .

#### 8.1.6.1 Method 1: Critical value approach

This method constructs bands of uniform width in terms of standard error units. The algorithm proceeds as follows:

1. Bootstrap Standard Error Calculation: The standard error for the point estimate at each point $y_j$ is estimated by the standard deviation of the $B$ bootstrap estimates at that specific point: $\text{SE}^*(\hat{F}(t_j)) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{F}_i^*(t_j) - \overline{\hat{F}}^*(t_j))^2}$ where $\hat{F}_i^*(t_j)$ is the estimate from the $i$-th bootstrap sample at prediction point $t_j$.
2. Distribution of Maximum Deviations: For each of the $B$ bootstrap curves, we compute the maximum standardized absolute deviation from the original point estimate across all prediction points: $M_i = \max_{j=1,\ldots,k}\left|\frac{\hat{F}_i^*(t_j) - \hat{F}(t_j)}{\text{SE}^*(\hat{F}(t_j))}\right|$
3. Critical Value Determination: The distribution of these maxima, $\{M_1, M_2, \ldots, M_B\}$, is used to find a single critical value, $\hat{q}_{1-\alpha}$, which is the empirical $(1-\alpha)$-th quantile of this distribution.
4. SCB Construction: The final SCB is constructed by creating a band around the prediction point estimate curve with a width determined by the critical value $\hat{q}_{1-\alpha}$ and the pointwise bootstrap standard errors: $\text{SCB}(t_j) = \hat{F}(t_j) \pm \hat{q}_{1-\alpha} \cdot \text{SE}^*(\hat{F}(t_j))$ This ensures that the entire band simultaneously covers the true function with approximately $(1-\alpha)$ probability.

#### 8.1.6.2 Method 2: Direct Quantile Calibration Method

This method directly finds the appropriate quantiles from the bootstrap distribution of curves to achieve the target simultaneous coverage, allowing the band width to vary more flexibly.

1. **Objective Function Definition:** The goal is to find a quantile probability $\zeta \in (0, \alpha/2)$ such that the band formed by the pointwise $\zeta$-th and $(1-\zeta)$-th quantiles of the bootstrap curves contains exactly $(1-\alpha) \times 100\%$ of the complete bootstrap curves.
2. **Numerical Calibration:** We define an objective function that computes the empirical simultaneous coverage for a given $\zeta$ and returns the difference from the nominal coverage level, $1-\alpha$. A numerical root-finding algorithm (`uniroot` in R) is then employed to solve for the optimal value, $\hat{\zeta}$, which makes this difference zero. The search for $\hat{\zeta}$ is constrained to a plausible interval, typically between a Bonferroni-corrected level $(\alpha/2k)$ and the uncorrected pointwise level $(\alpha/2)$.

3. **SCB Construction:** Once the calibrated value $\hat{\zeta}$ is determined, the final SCBs are constructed using the pointwise quantiles of the bootstrap distribution of function estimates. The lower band is the $\hat{\zeta}$-th quantile at each prediction point $y_j$, and the upper band is the $(1-\hat{\zeta})$-th quantile:

$$\text{Lower SCB}(t_j) = \text{Quantile}(\{\hat{F}_i^*(t_j)\}_{i=1}^B, \text{probs} = \hat{\zeta})$$

$$\text{Upper SCB}(t_j) = \text{Quantile}(\{\hat{F}_i^*(t_j)\}_{i=1}^B, \text{probs} = 1-\hat{\zeta})$$

### 8.1.7 95% confidence intervals for the HRDRs and HNRDRs features

Highest Risk Density Regions are a form of prediction intervals. The 95% Highest Risk Density Region of a hypothetical time-to-event distribution following the normal model would be equivalent to the 95% prediction intervals, due to the symmetry of the normal distribution.

While confidence intervals around prediction intervals are rare in the literature, quantifying the uncertainty around the period where risk is maximally distributed or where treatment has the maximum impact is important.

In this work, we quantify the uncertainty around the estimates of the intervals by focusing on their individual features like the lower boundary (the start of the intervals), the upper boundary (the end of the intervals), the mid-point (the location of the intervals) and the width (the scale). We consider a method based on the bootstrap corrected for the bias and the acceleration factor. The output of the bootstrap procedure is a collection of confidence intervals for specific, pre-defined scalar features of the HDR (e.g., its start, end, or length). The inferential statement is marginal for each feature, e.g., $P(\text{true start} \in [\text{start}_L, \text{start}_U]) = 1-\alpha$ or $P(\text{true end} \in [\text{end}_L, \text{end}_U]) = 1-\alpha$ or $P(\text{true length} \in [\text{length}_L, \text{length}_U]) = 1-\alpha$.

This method is ideal for formal hypothesis testing on specific geometric properties of the HRDR and HNRDR.

Potential issues could arise when bootstrapping the start points of the regions when considering large values of probability mass and heavily skewed distributions (e.g., the exponential distribution). The bootstrap distribution of these features would be characterized by a strong asymmetry due to high probability of obtaining as a starting point or a ending point the first or last value of the domain of x.

We can also quantify the geometric similarity between the original interval estimate and a distribution of intervals derived from bootstrap samples, thereby providing a holistic measure of its stability against sampling variability.

The procedure begins by generating a large number of bootstrap samples from the original dataset. Let the original sample be denoted by $X = \{x_1, x_2, ..., x_n\}$, with a sample size of $n$. We generated $B$ bootstrap samples, $X_b^*$, for $b = 1, ..., B$, by sampling $n$ observations from $X$ with replacement. For each bootstrap sample $X_b^*$, we re-computed the interval of interest, yielding a collection of $B$ bootstrap intervals, denoted as $I_b = [a_b, b_b]$.

The stability of the original interval $I$ was then assessed by comparing it to each bootstrap replicate $I_b$ using the Jaccard index, a well-established metric for set similarity. For two non-empty intervals, the Jaccard index is defined as the ratio of the length of their intersection to the length of their union. The formal definition is:

$$J(I, I_b) = \frac{\text{length}(I \cap I_b)}{\text{length}(I \cup I_b)} \tag{8.21}$$

Where the intersection of the two intervals is given by:

$$I \cap I_b = [\max(a, a_b), \min(b, b_b)] \tag{8.22}$$

And the union of the two intervals is given by:

$$I \cup I_b = [\min(a, a_b), \max(b, b_b)] \tag{8.23}$$

The length of an interval $[x, y]$ is defined as $y - x$, provided $y \geq x$. The Jaccard index is bounded between 0 and 1, where $J = 1$ signifies perfect concordance between the intervals and $J = 0$ indicates that the intervals are disjoint (i.e., have no overlap).

This procedure yields a distribution of $B$ Jaccard indices, $\{J(I, I_1), J(I, I_2), ..., J(I, I_B)\}$. This distribution provides a comprehensive summary of the interval's stability. A distribution tightly concentrated near 1 suggests that the interval is highly robust to sampling variation, as bootstrap replicates consistently exhibit a high degree of overlap with the original estimate. Conversely, a distribution with a lower mean or greater variance indicates that the interval's boundaries and location are sensitive to the specific sample drawn, signifying lower stability.

To provide a single summary statistic for reporting, we calculated the mean of the Jaccard index distribution, $\bar{J}$, to represent the expected similarity of the interval under resampling. This metric serves as an intuitive and quantitative measure of the overall stability of the estimated interval.

### 8.1.8 Demonstration of the methods and simulation Study

#### 8.1.8.1 Artificial causal scenarios

To rigorously evaluate the performance of the estimation methods, we designed four distinct data-generating scenarios. These scenarios were chosen to represent a range of complexities, from simple proportional hazards models to more challenging non-proportional hazards structures with time-varying effects, mimicking situations often encountered in clinical and epidemiological research. In all scenarios, the data structure consists of a binary treatment ($A$), five pre-treatment covariates ($L_1, ..., L_5$) that act as confounders, and a time-to-event outcome ($T$).

The confounding mechanism was held constant across all scenarios. The probability of receiving treatment was determined by the covariates through a logistic model:

$$P(A = 1|L) = \text{logit}^{-1}(\alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_3 + \alpha_4 L_4 + \alpha_5 L_5) \tag{8.24}$$

where the coefficients $(\alpha_0, \ldots, \alpha_5)$ were set to $(-0.5, 0.5, -0.3, 0.2, 0.4, -0.2)$. The covariates themselves consisted of three binary variables $(L_1, L_2, L_3 \sim \text{Bernoulli}(0.5))$ and two continuous standard normal variables $(L_4, L_5 \sim N(0, 1))$.

**The Exponential Scenario**

This scenario represents the simplest case of proportional hazards, where the hazard rate is constant over time. The event times were generated from a Weibull distribution with the shape parameter, $\gamma$, fixed to 1, which simplifies the model to an exponential distribution. The scale parameter, $\lambda$, was determined by an accelerated failure time (AFT) model conditional on the treatment and covariates:

$$\log(T) = \log(\lambda_0) + \beta_A A + \beta_1 L_1 + \cdots + \beta_5 L_5 + \epsilon \tag{8.25}$$

where $\log(\lambda_0) = \log(60)$ is the baseline log-scale parameter, the coefficients $(\beta_A, \ldots, \beta_5)$ were set to $(0.7, -0.3, 0.2, 0.1, -0.25, -0.15)$, and $\epsilon$ is an error term following a Gumbel distribution. This structure results in a constant hazard ratio and a monotonically decreasing probability density function (PDF), where the highest risk of an event occurs at the beginning of the follow-up period.

**The Weibull Scenario**

This scenario also adheres to the proportional hazards assumption but introduces a time-varying hazard rate. Event times were generated from a Weibull distribution with a shape parameter $\gamma = 1.5$, resulting in a monotonically increasing hazard function. The AFT model structure for the scale parameter was identical to the exponential scenario, but with a different baseline log-scale parameter of $\log(\lambda_0) = \log(70)$. The key feature of this scenario is its unimodal, non-monotonic PDF, which contrasts with the ever-increasing hazard function. This design allows us to assess how well the methods identify a period of peak risk that is not at the start of the follow-up.

**The Log-Logistic Scenario**

This scenario was designed to represent a more complex non-proportional hazards setting. Event times were generated from a log-logistic distribution with a shape parameter of 1.5 and a baseline log-scale parameter of log(36). The AFT model structure was modified to include treatment-by-covariate interactions, making the treatment effect dependent on the individual's baseline characteristics. The resulting hazard function is non-monotonic (first increasing, then decreasing), and the hazard ratio is time-varying. This scenario tests the methods' robustness to violations of the proportional hazards assumption and their ability to handle effect modification.

**The Flexible "Vanishing Effect" Scenario**

This final scenario represents a challenging non-proportional hazards case designed to mimic a "vanishing treatment effect," where the initial benefit of a treatment diminishes over time.

The scenario represents a "vanishing treatment effect" scenario, analogous to patterns observed in clinical trials such as the GOG111 ovarian cancer study, where progression-free survival of patients with advanced ovarian cancer treated with paclitaxel–cisplatin (TP) versus cyclophosphamide–cisplatin (CP) drug regimens in a randomized clinical trial was evaluated. To simulate this scenario, we digitalized the survival functions displayed in Figure 1 of the article presenting the results for the first three years of the trial (Piccart 2000) and the utilize the simulation approach already discussed in Section 3.6 .

We defined distinct, non-monotonic baseline hazard functions, $h_{0,A=1}(t)$ and $h_{0,A=0}(t)$, for the treated and control groups, respectively. The individual hazard for a subject $i$ was then calculated as:

$$h_i(t|A_i, L_i) = h_{0,A=A_i}(t) \cdot \exp(\beta_1 L_{1i} + \cdots + \beta_5 L_{5i}) \tag{8.26}$$

where the coefficients $(\beta_1, \ldots, \beta_5)$ were set to $(1.2, 0.8, 1.3, 0.7, 1.6)$. The baseline hazard functions were constructed to be complex and bimodal, leading to a hazard ratio that varies substantially over time. However, the resulting marginal PDFs are relatively simple unimodal functions that are shifted in time. This scenario critically evaluates the ability of the methods to deconvolve complex conditional hazard structures to accurately estimate simpler marginal risk profiles.

#### 8.1.8.2 Simulation

To evaluate the finite-sample performance of the proposed methods, a comprehensive Monte Carlo simulation study was conducted. The primary objective was to assess the ability of the two distinct flexible modeling frameworks to accurately recover true, underlying marginal (counterfactual) PDFs and CDFs and in identifying the regions of interest, with the estimated functions and the algorithmic approaches described above.

We evaluated the performance of the approaches as a function of the complexity of the causal scenario involved, as well as the sample size available. Adopting the causal scenarios described in section 4.1, we varied the sample size considering 250, 500, 750, 1000, 1500 and 2000 subjects.

For each scenario, the true marginal PDFs and CDFs for each treatment arm ($a \in \{0, 1\}$) and their respective difference functions were calculated by simulating a large cohort (N=20,000) and averaging the conditional functions over the empirical distribution of the covariates. These served as the gold standard for performance evaluation.

For each of the five scenarios and sample size, we performed $M = 1000$ Monte Carlo replications. The true event times, $T_{true}$, were generated according to the scenario-specific model. Observed data were then created by applying random censoring, with event times drawn from an exponential distribution at a rate chosen to achieve a proportion of random dropout of about 10 %, and a fixed administrative censoring time at 161 months to achieve an administrative censoring proportion of 15%.

In each replication, the marginal PDFs and CDFs were estimated using the two distinct modeling approaches. Regardless the flexible model employed, both the two distinct algorithms (up-down and right-left approach) were used. The probability mass considered for the computation of the HRDR and the absolute risk difference mass for the HNRDR were $(0.25, 0.4, 0.5, 0.6)$ and $(-0.05, -0.10, -0.15, -0.20, -0.25)$.

A comprehensive set of metrics was used to evaluate the accuracy of both the point-wise function estimates and the interval-based region estimates.

To assess the quality of the estimated marginal functions, we first decomposed the error at each time point $t$ into bias and variance components. For any given function estimate $\hat{\theta}(t)$ (representing $\hat{f}_a(t)$, $\hat{F}_a(t)$, etc.) with a true value of $\theta(t)$, the point-wise bias and point-wise variance were estimated.

The overall point-wise accuracy was quantified using the point-wise Mean Squared Error (MSE), which combines squared bias and variance, and its square root, the Root Mean Squared Error (RMSE). The accuracy of the estimated HRDRs and HNRDRs was evaluated using two complementary approaches.

First, the Jaccard Index was used to measure the degree of overlap between the true region, $R_{true}$, and the estimated region, $\hat{R}_{est}$ (Equation 8.21) .

Second, we evaluated the estimation accuracy of four key features of the intervals: the lower boundary ($L$), the upper boundary ($U$), the width ($W = U - L$), and the mid-point ($M = (L+U)/2$). For each of these features, we calculated the empirical bias, variance, and RMSE across the $M$ replications to provide a detailed decomposition of the estimation error.

Finally, we also calculated the empirical proportion of replications in which the estimation procedure successfully produced a non-empty region for the specified probability coverage or cumulative difference target.

#### 8.1.8.3 Application on progression free survival time

We will show the application of the methods to a real dataset with the aim of evaluating the prognostic impact of lymph nodal status (node-positive N+, node negative N-) on the progression free survival time of patients resected from the primary tumor of the breast in the study "Milan I trial" (Umberto Veronesi et al. 1981). Details of the trial are found in Section 3.7.

In the Milan I trial, the definitive pathological status of the axillary lymph nodes was an unknown variable at the point of randomization. This crucial piece of prognostic information was only obtained as a direct result of the surgical intervention itself—a complete axillary dissection—which was a core component of both the standard and investigational treatment arms. At that time, ALND was viewed primarily as a therapeutic intervention, believed to be essential for achieving regional disease control and preventing the further spread of cancer. Its role in providing prognostic staging information was considered important but secondary to its therapeutic purpose. The fact that the investigators included ALND in both arms indicates that its necessity for regional control was not in question; the central hypothesis of

the trial was whether the breast itself could be preserved, not whether the axilla could be spared dissection.

The analysis is aimed to compute the prognostic impact of the pathological lymph nodal status (pN) of the patients with an outcome model along with the use of the g-formula. The outcome model is a piece-wise exponential model that includes menopausal status of the patients, the tumor dimension and the surgical procedure plus possible radiotherapy administered thereafter as predictor variable, in addition to the pN of the subjects. We will relax the proportionality of the hazards assumption for pN and display how the PDF and the HRDR can easily deal in describing a fairly complex prognostic effect.

Pathological nodal status is a biological state discovered post-randomization, not an intervention that can be assigned. While the g-formula appropriately adjusts for measured baseline confounders, attributing the remaining difference solely to the "causal" effect of nodal status requires strong, untestable assumptions about the absence of unmeasured confounding between the biological determinants of nodal status and the outcome. The result is better interpreted as a confounder-adjusted description of the prognostic trajectories associated with nodal status.

## 8.2 Results

Figure 8.1, Figure 8.2, Figure 8.3 and Figure 8.4 are organized into a standardized set of panels, from (a) to (i), which report key statistical measures including survival functions, cumulative distribution functions, hazard rates, hazard ratios, and functions of the restricted mean survival time (RMST). The final panels are dedicated to the probability density functions (PDFs) and the difference between them for the treatment and control groups. These latter panels are supplemented with Highest Risk Density Regions (HRDRs) and Highest Net Risk Difference Regions (HNRDRs), respectively. To illustrate the utility of these regions for risk communication, we present a series of simulation scenarios that progress in complexity, beginning with an exponential model and concluding with a non-proportional hazards model.

**The exponential scenario**

The exponential scenario, detailed in Figure 8.1, serves as a foundational case. Its defining characteristic is a constant hazard function over time (Panel d), which satisfies the proportional hazards assumption and yields a constant hazard ratio (Panel e). While the hazard ratio offers a convenient single-summary measure of effect, its relative nature obscures the temporal dynamics of absolute risk. Alternative metrics also present interpretational challenges. The risk difference (Panel c), for instance, is time-dependent and non-monotonic, making the choice of a single time-point for reporting ambiguous. Similarly, the difference in RMST (Panel g) is sensitive to the choice of the truncation point, $\tau$, leading to an effect size that varies with the analytical window.

In contrast, the PDFs of the time-to-event variable (Panel h) provide a more granular view by showing the unconditional instantaneous risk of the event. For the exponential model, these functions are monotonically decreasing, revealing that the absolute risk is highest at

the beginning of the follow-up period and diminishes over time—an insight not apparent from the constant hazard ratio. To formalize the interpretation of these densities, HDRs can identify the time-window in which a specified probability mass is most concentrated. For a 50% probability mass, the HDR for the control group is the interval $[0, 42]$ months, whereas for the treatment group it is $[0, 84]$ months, illustrating the treatment's effect of delaying risk accumulation. Given the monotonic nature of the PDFs, these HDRs correspond to quantile intervals of the form $[0, F^{-1}(P_{mass})]$, indicating the treatment effect is primarily on the scale rather than the location of the distribution.

To quantify the treatment effect directly, the HNRDR applied to the instantaneous risk difference function (Panel i) identifies the window of maximal risk reduction. For example, a 20% risk reduction is concentrated in the interval $[0, 37]$ months. Table 1 provides the numerical features of these intervals for a range of probability values. This framework thereby shifts the communication of risk from a single relative measure to absolute measures defined on the time scale of the event.

**The Weibull scenario**

The subsequent Weibull scenario, depicted in Figure 8.2, illustrates a case where the hazard functions are not constant but are monotonically increasing for both groups (Panel d). Despite this, their ratio remains constant over time, thus satisfying the proportional hazards assumption (Panel e). While the hazard ratio remains a convenient summary, the interpretation of the hazard functions themselves can be misleading; their increasing nature is a consequence of conditioning on survival and does not reflect the timing of the highest absolute risk. A more direct representation is offered by the probability density functions (Panel h), which show the unconditional instantaneous risk. In stark contrast to the exponential case, these PDFs are non-monotonic, each with a distinct mode that indicates the period of highest risk—approximately 30 months for the control group and 60 months for the treatment group. This demonstrates that the peak risk occurs well before the end of the follow-up, an insight obscured by both the hazard functions and the summary hazard ratio.

To quantify these temporal shifts in risk concentration, we compute the Highest Density Regions. The 50% HDRs indicate that the control group accumulates half its risk in the time-window of $[14, 64]$ months, while for the treatment group, this period is delayed and broadened to $[27, 128]$ months. Here, the treatment effect is more pronounced on the scale (width) of the interval than on its location. However, examining an HDR with a smaller probability mass, such as 25%, can provide a more focused comparison. The 25% HDR for the control group is $[23, 46]$ months, whereas for the treatment group it is $[46, 94]$ months. This reveals a clear 23-month delay in the onset of this concentrated risk period, as shown by the difference in the lower interval boundaries. To identify the period of maximum treatment efficacy, the Highest Net Risk Difference Region is applied to the risk difference function. A 20% absolute risk reduction is concentrated in the time-window of months $[12, 46]$ , with the peak effect occurring around 25 months. A smaller 5% absolute risk reduction is similarly concentrated in a narrower window of $[22, 30]$ months.

**The Log-logistic scenario**

The log-logistic model, presented in Figure 8.3, also produces a unimodal, non-monotonic event-time density similar to the Weibull case (Panel h). However, it introduces greater com-

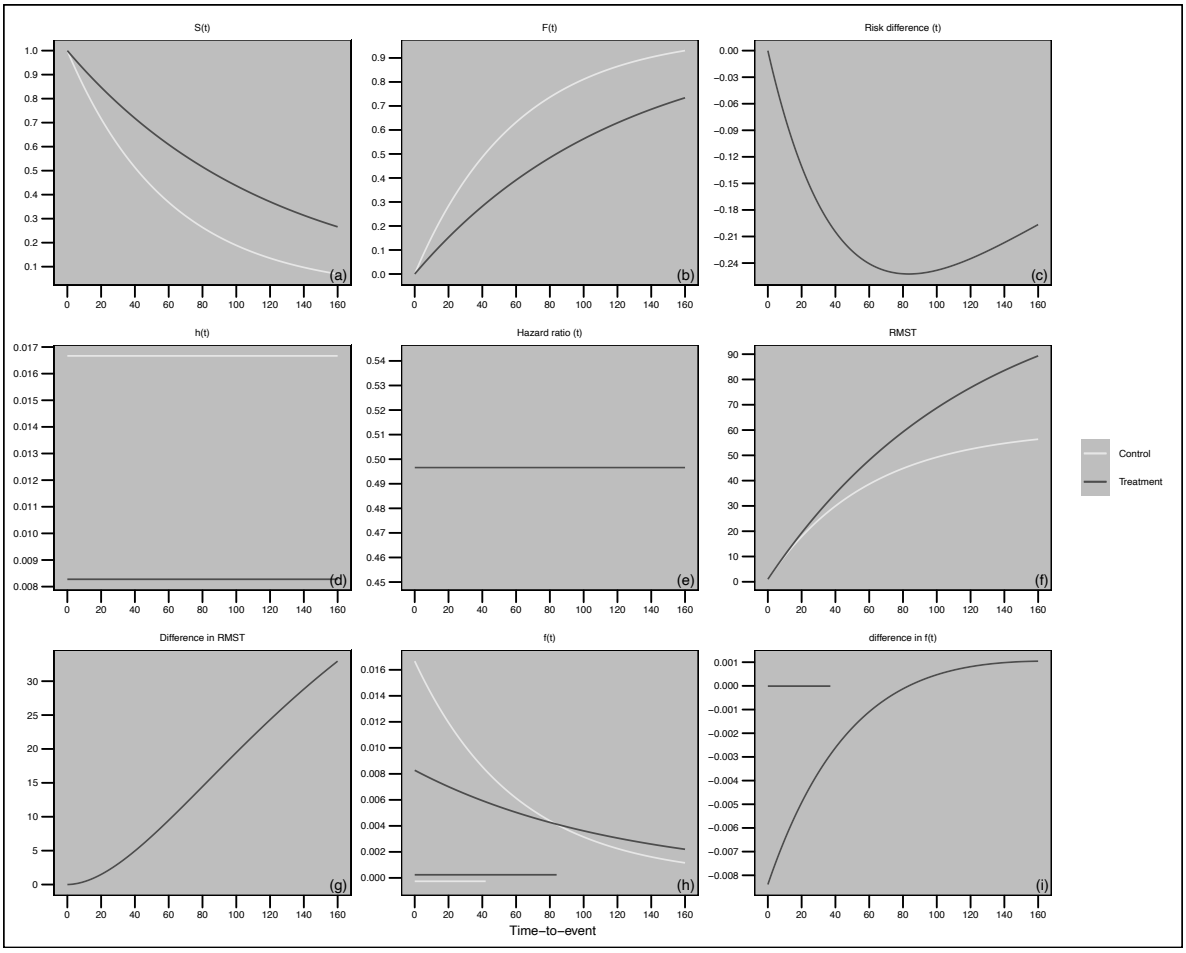



Figure 8.1: Graphical representation of the exponential simulation scenario, defined by a constant hazard rate over time. Its nine panels display the following functions: (a) Survival, S(t); (b) Cumulative Distribution Function, F(t); (c) Risk Difference, D(t); (d) Hazard, h(t); (e) Hazard Ratio, HR(t); (f) Restricted Mean Survival Time, RMST( ); (g) Difference in RMST, $\Delta$RMST( ); (h) Probability Density Function, f(t); and (i) Difference in PDFs. Throughout the figure, the control group is represented by a white solid line and the treatment group by a black solid line.

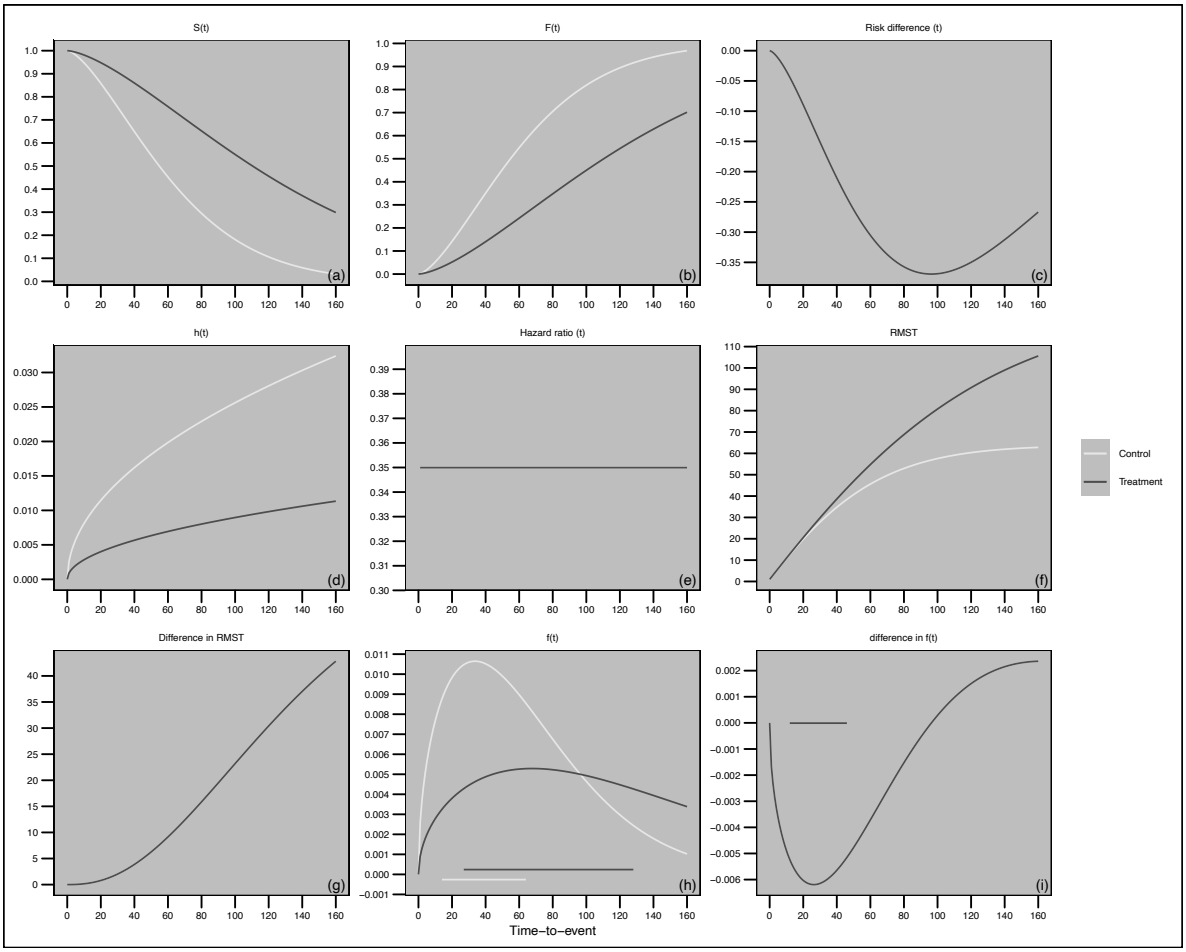



Figure 8.2: Results from a Weibull proportional hazards scenario with monotonically increasing hazard rates. Its nine panels show the: (a) Survival, S(t); (b) Cumulative Distribution Function, F(t); (c) Risk Difference, D(t); (d) Hazard, h(t); (e) Hazard Ratio, HR(t); (f) Restricted Mean Survival Time, RMST( ); (g) Difference in RMST, ΔRMST( ); (h) Probability Density Function, f(t); and (i) Difference in PDFs. The control group is represented by a white solid line and the treatment group by a black solid line.

plexity as it is characterized by non-proportional and non-monotonic hazards. This complexity presents a significant challenge for communicating the treatment effect via conventional metrics, particularly with a time-varying hazard ratio as seen in Panel (e). In contrast, the Highest Risk Density Region provides a robust tool for summarizing treatment effects that is not dependent on the proportional hazards assumption.

The 50% HRDR reveals that the control group's primary risk window is $[3, 38]$ months, whereas the treatment group exhibits a slower accumulation of risk over a wider window of $[5, 76]$ months. Examining the 25% HRDRs for the control group, $[7, 22]$ months, and the treatment group, $[13, 43]$ months, further illustrates that the treatment's primary impact is on the scale of risk accumulation rather than its location. This is evidenced by a 6-month difference in the lower boundaries compared to a 15-month difference in the interval widths. Consistent with the exponential scenario, the treatment effect is most pronounced early in the follow-up period; the 20% HNRDR indicates that an absolute risk reduction of this magnitude is most concentrated in the time-window of $[2, 26]$ months.

**The 'vanishing effect' scenario**

While the survival and cumulative distribution functions present a simple visual pattern (Panels a-b), the underlying hazard functions and the resultant hazard ratio are considerably more complex. The hazards exhibit non-monotonic, bimodal shapes, and the hazard ratio varies substantially over time, making its interpretation exceptionally challenging.

In contrast, the corresponding probability density functions (Panel h) are far simpler, presenting as unimodal functions of similar shape that are primarily shifted in time. The analysis of these densities confirms that the treatment's effect is more on the location than on the scale of the risk period. This is substantiated by the 50% Highest Risk Density Regions: the control group's interval is $[5, 48]$ months, while the treatment group's is $[11, 71]$ months. A comparison reveals that the interval width for the treatment group is only 1.4 times that of the control group, whereas its lower boundary is 2.2 times greater, confirming the dominant effect on location. While the overall magnitude of the treatment effect is smaller in this scenario, with a maximum risk reduction not exceeding 16%, the Highest Net Risk Density Regions demonstrate that the treatment's efficacy is still concentrated early in the follow-up. An absolute risk reduction of 10% is concentrated in the window of $[6, 23]$ months, and a 15% reduction is found between $[0, 30]$ months.

### 8.2.1 Simulation results

#### 8.2.1.1 General performance of the estimation methods in terms of F(t) and f(t) and their respective difference-functions

Figure 8.5 illustrates the performance of two flexible modeling approaches—a piecewise exponential model and a pseudo-observation model—in estimating several target functions. The analysis, based on root mean squared error, evaluates the estimation of probability density functions, cumulative distribution functions, and the differences between them across various sample sizes.

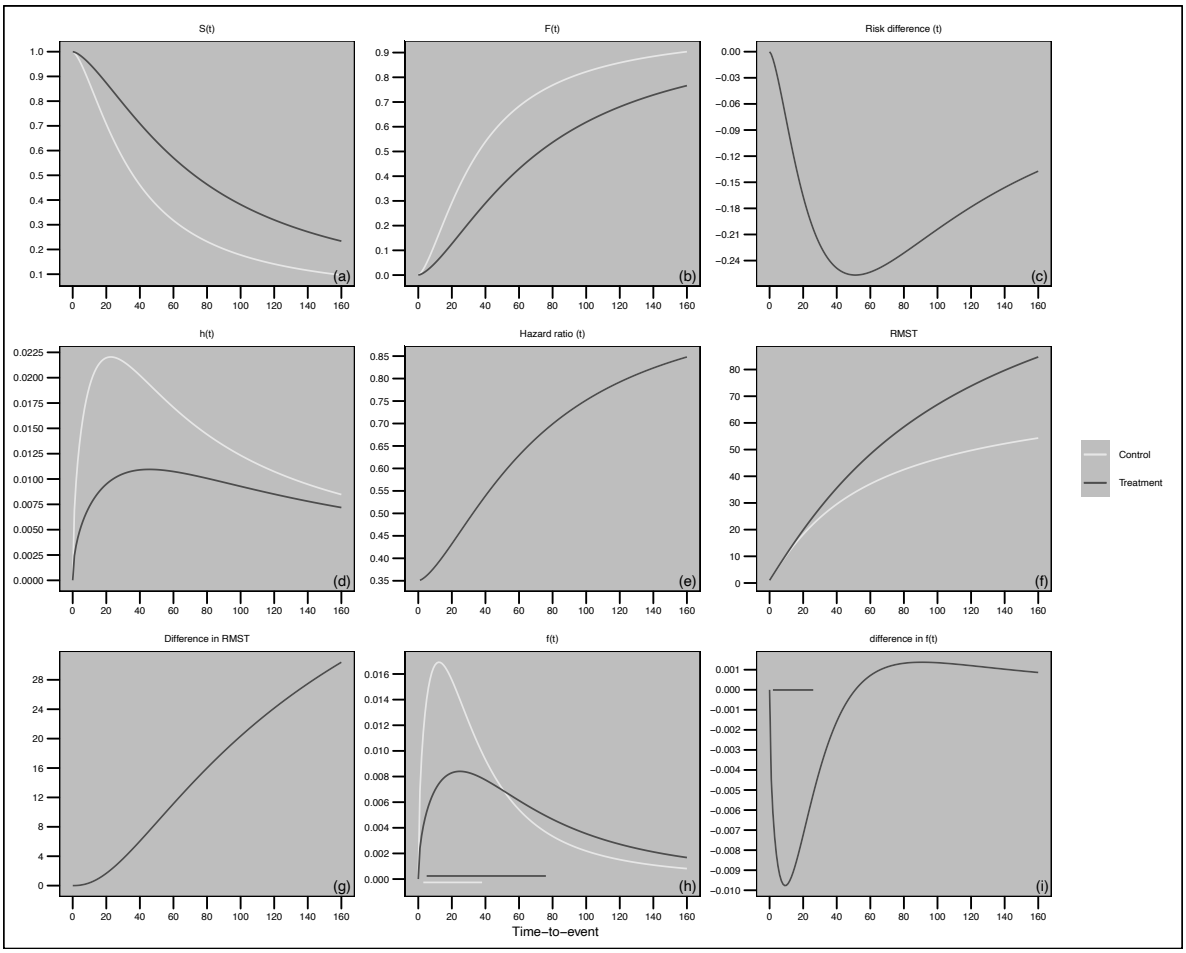



Figure 8.3: This figure illustrates a non-proportional hazards scenario based on a log-logistic model, which is characterized by a time-varying hazard ratio. Its panels display the: (a) Survival, S(t); (b) Cumulative Distribution Function, F(t); (c) Risk Difference, D(t); (d) Hazard, h(t); (e) Hazard Ratio, HR(t); (f) Restricted Mean Survival Time, RMST( ); (g) Difference in RMST, $\Delta$RMST( ); (h) Probability Density Function, f(t); and (i) Difference in PDFs, for both a control (white solid line) and treatment (black solid line) group.

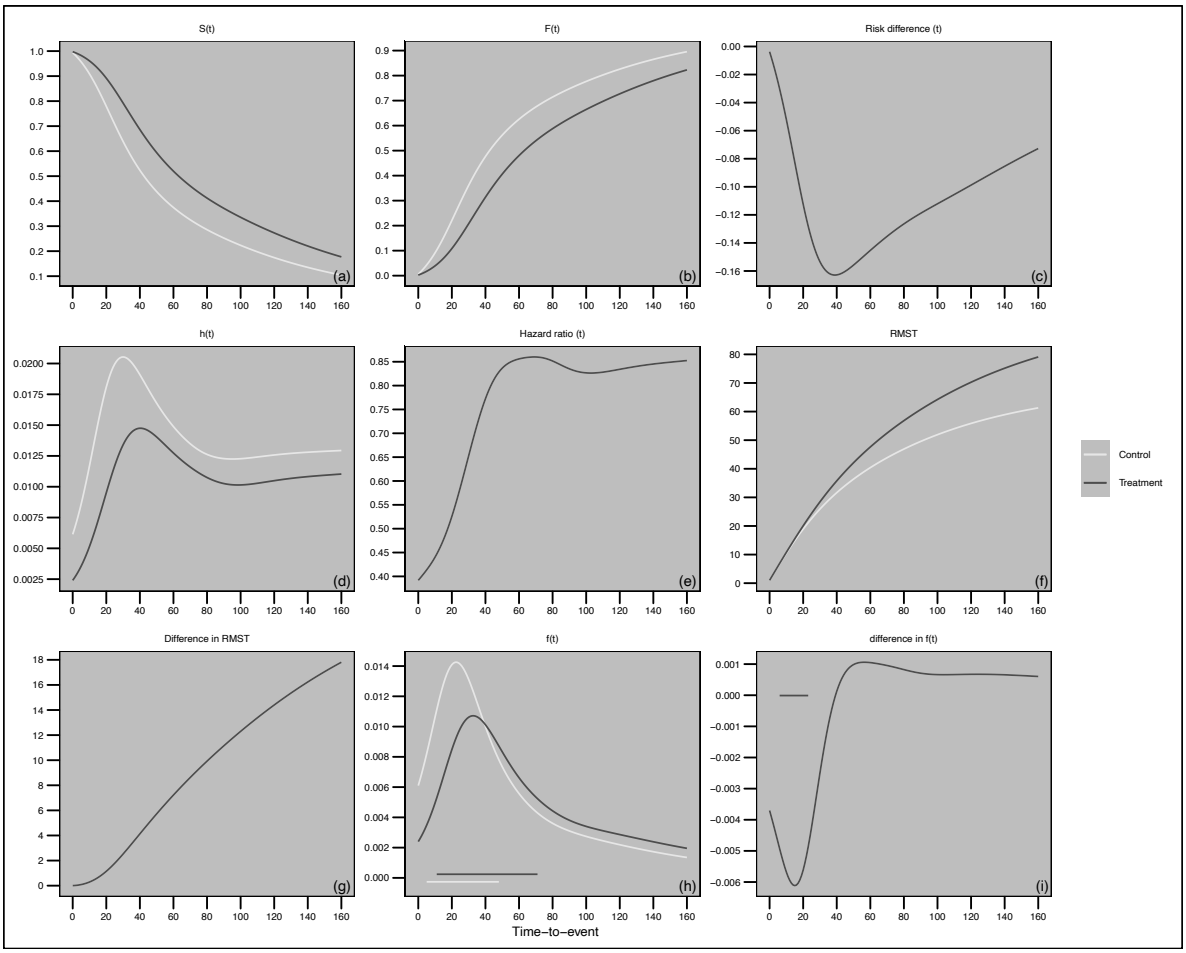



Figure 8.4: This figure represents a complex non-proportional hazards scenario featuring a vanishing treatment effect. Its nine panels report the: (a) Survival, S(t); (b) Cumulative Distribution Function, F(t); (c) Risk Difference, D(t); (d) Hazard, h(t); (e) Hazard Ratio, HR(t); (f) Restricted Mean Survival Time, RMST( ); (g) Difference in RMST, ΔRMST( ); (h) Probability Density Function, f(t); and (i) Difference in PDFs for a control (white solid line) and treatment (black solid line) group.

Table 8.5: This table presents a detailed summary of the numerical features of the Highest Risk Density Regions (HRDRs) and the Highest Net Risk Difference Regions (HNRDRs) for each of the four simulation scenarios. For a range of specified probability mass values, the table reports the lower boundary, upper boundary, mid-point, and width for each interval. Values are provided for the HRDRs corresponding to the control and treatment groups, as well as for the HNRDRs corresponding to the risk difference function. The values in this table provide the quantitative basis for the intervals visualized in Figures 2 through 5 and for the interpretations discussed in the text. Values for control and treatment groups are shown as `value (value)`

| | Highest Net Risk Difference Region | | | | Highest Risk Density Region | | | |
|---|---|---|---|---|---|---|---|---|
| | -0.2 | -0.15 | -0.1 | -0.05 | 0.25 | 0.4 | 0.5 | 0.6 |
| Exponential | | | | | | | | |
| Lower boundary | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 (0.0) | 0.0 (0.0) | 0.0 (0.0) | 0.0 (0.0) |
| Upper boundary | 37.0 | 23.0 | 14.0 | 6.0 | 17.0 (35.0) | 31.0 (62.0) | 42.0 (84.0) | 55.0 (111.0) |
| Mid-point | 18.5 | 11.5 | 7.0 | 3.0 | 8.5 (17.5) | 15.5 (31.0) | 21.0 (42.0) | 27.5 (55.5) |
| Width | 37.0 | 23.0 | 14.0 | 6.0 | 17.0 (35.0) | 31.0 (62.0) | 42.0 (84.0) | 55.0 (111.0) |
| Weibull | | | | | | | | |
| Lower boundary | 12.0 | 16.0 | 19.0 | 22.0 | 23.0 (46.0) | 17.0 (35.0) | 14.0 (27.0) | 10.0 (20.0) |
| Upper boundary | 46.0 | 40.0 | 35.0 | 30.0 | 46.0 (94.0) | 56.0 (113.0) | 64.0 (128.0) | 72.0 (146.0) |
| Mid-point | 29.0 | 28.0 | 27.0 | 26.0 | 34.5 (70.0) | 36.5 (74.0) | 39.0 (77.5) | 41.0 (83.0) |
| Width | 34.0 | 24.0 | 16.0 | 8.0 | 23.0 (48.0) | 39.0 (78.0) | 50.0 (101.0) | 62.0 (126.0) |
| Log-logistic | | | | | | | | |
| Lower boundary | 2.0 | 3.0 | 5.0 | 7.0 | 7.0 (13.0) | 4.0 (8.0) | 3.0 (5.0) | 2.0 (3.0) |
| Upper boundary | 26.0 | 19.0 | 15.0 | 12.0 | 22.0 (43.0) | 30.0 (60.0) | 38.0 (76.0) | 48.0 (97.0) |
| Mid-point | 14.0 | 11.0 | 10.0 | 9.5 | 14.5 (28.0) | 17.0 (34.0) | 20.5 (40.5) | 25.0 (50.0) |
| Width | 24.0 | 16.0 | 10.0 | 5.0 | 15.0 (30.0) | 26.0 (52.0) | 35.0 (71.0) | 46.0 (94.0) |
| NPH scenario | | | | | | | | |
| Lower boundary | | 0.0 | 6.0 | 11.0 | 14.0 (22.0) | 9.0 (16.0) | 5.0 (11.0) | 0.0 (7.0) |
| Upper boundary | | 30.0 | 23.0 | 19.0 | 32.0 (46.0) | 40.0 (59.0) | 48.0 (71.0) | 57.0 (90.0) |
| Mid-point | | 15.0 | 14.5 | 15.0 | 23.0 (34.0) | 24.5 (37.5) | 26.5 (41.0) | 28.5 (48.5) |
| Width | | 30.0 | 17.0 | 8.0 | 18.0 (24.0) | 31.0 (43.0) | 43.0 (60.0) | 57.0 (83.0) |

As anticipated, both methods demonstrate acceptable performance, with estimation accuracy improving substantially as the sample size increases toward 1,000 observations. However, a key distinction emerges when comparing the estimation of instantaneous versus cumulative functions. For instantaneous functions, such as the PDF and its difference, performance was sensitive to the complexity of the underlying data-generating scenario, with the highest accuracy achieved in the exponential and Weibull scenarios. In this context, the piecewise exponential model showed markedly superior performance compared to the pseudo-observation model.

In contrast, a different pattern was observed for cumulative functions. Here, the estimation of the CDF and its difference was far more robust across the various scenarios, and critically, the performance advantage of the piecewise exponential model over the pseudo-observation model was no longer evident. This finding has important practical implications, suggesting that methods based on cumulative functions for the construction of Highest Risk Density Regions (HRDRs), may be more stable and reliable than the one based on the PDF. This is particularly relevant for researchers using the pseudo-observation approach, as focusing on cumulative targets appears to mitigate the performance deficits seen with instantaneous functions.

Regarding the non-proportional hazards (NPH) scenario, while both methods struggled to accurately estimate the marginal CDFs under NPH, their ability to estimate the difference-functions remained remarkably robust, achieving an accuracy comparable to that seen in less complex scenarios. This suggests that even when the underlying marginal distributions are challenging to model, the contrast between them might be captured effectively.

Table 8.6 displays numerically the estimation performance of the flexible models as a function of the sample size, complexity of the scenario and the type of the target function.

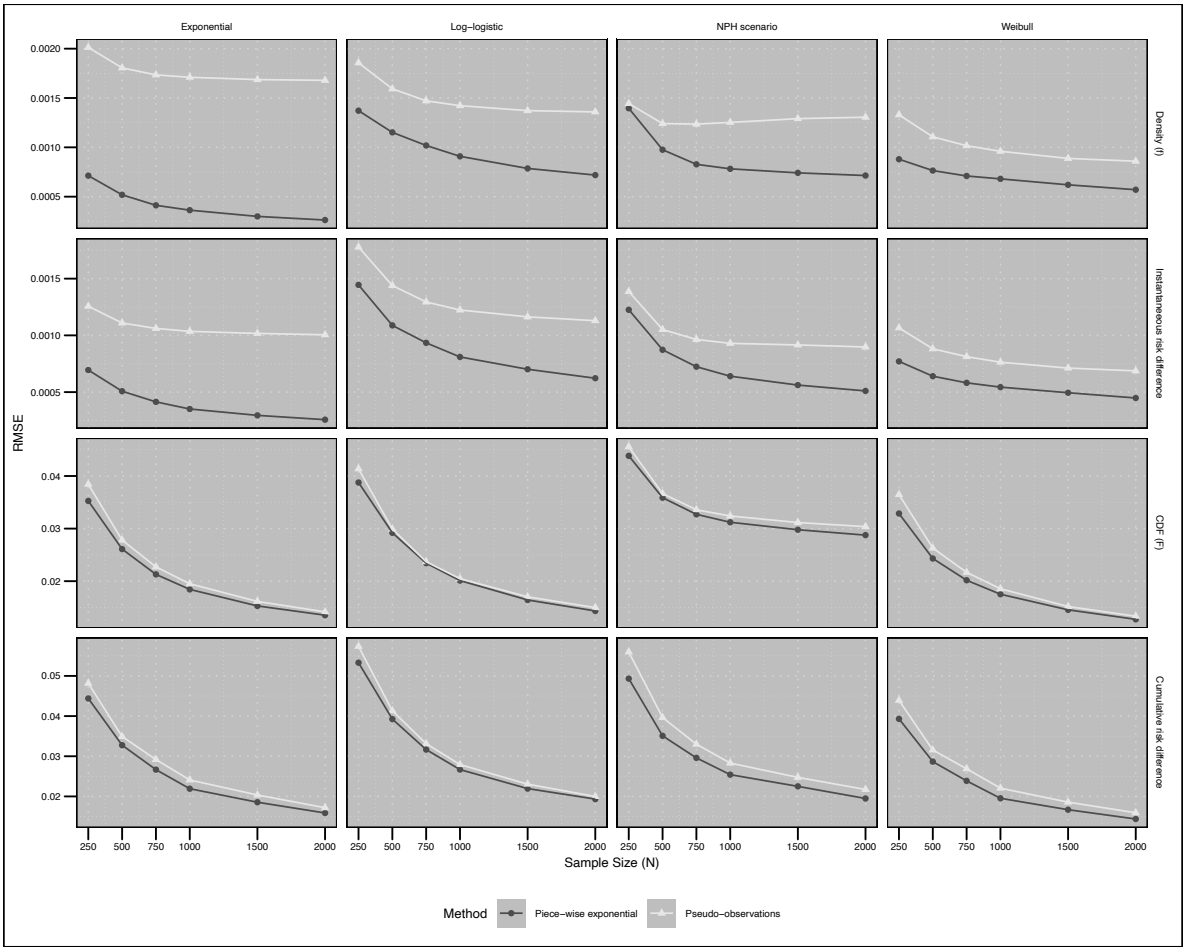



Figure 8.5: Overall Performance: RMSE vs. Sample Size, for the target functions f(t), F(t) and their respective difference functions d(t) and D(t). RMSE is aggregated over time and treatment categories

Table 8.6: Comparison of the performance of the flexible methods for the estimation of the target functions. Root Mean Squared Error (RMSE) across different target functions and sample sizes

| | Density (f) | | | | | | CDF (F) | | | | | | Instantaneous risk difference | | | | | | Cumulative risk difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | N = 250 | N = 500 | N = 750 | N = 1000 | N = 1500 | N = 2000 | N = 250 | N = 500 | N = 750 | N = 1000 | N = 1500 | N = 2000 | N = 250 | N = 500 | N = 750 | N = 1000 | N = 1500 | N = 2000 | N = 250 | N = 500 | N = 750 | N = 1000 | N = 1500 | N = 2000 |
| **Exponential** | | | | | | | | | | | | | | | | | | | | | | | | |
| Piece-wise exponential | 0.00071 | 0.00052 | 0.00041 | 0.00036 | 0.00030 | 0.00026 | 0.03526 | 0.02610 | 0.02129 | 0.01844 | 0.01529 | 0.01353 | 0.00069 | 0.00051 | 0.00041 | 0.00035 | 0.00030 | 0.00026 | 0.04439 | 0.03275 | 0.02666 | 0.02192 | 0.01855 | 0.01588 |
| Pseudo-observations | 0.00201 | 0.00181 | 0.00173 | 0.00171 | 0.00169 | 0.00168 | 0.03843 | 0.02782 | 0.02272 | 0.01951 | 0.01615 | 0.01416 | 0.00126 | 0.00111 | 0.00106 | 0.00104 | 0.00102 | 0.00101 | 0.04817 | 0.03490 | 0.02919 | 0.02410 | 0.02035 | 0.01719 |
| **Log-logistic** | | | | | | | | | | | | | | | | | | | | | | | | |
| Piece-wise exponential | 0.00137 | 0.00115 | 0.00102 | 0.00091 | 0.00079 | 0.00072 | 0.03877 | 0.02921 | 0.02343 | 0.02011 | 0.01643 | 0.01434 | 0.00145 | 0.00109 | 0.00093 | 0.00081 | 0.00070 | 0.00062 | 0.05329 | 0.03925 | 0.03166 | 0.02668 | 0.02195 | 0.01933 |
| Pseudo-observations | 0.00186 | 0.00159 | 0.00147 | 0.00142 | 0.00137 | 0.00136 | 0.04137 | 0.02992 | 0.02370 | 0.02048 | 0.01702 | 0.01497 | 0.00178 | 0.00144 | 0.00129 | 0.00122 | 0.00116 | 0.00113 | 0.05735 | 0.04132 | 0.03314 | 0.02793 | 0.02305 | 0.01998 |
| **NPH scenario** | | | | | | | | | | | | | | | | | | | | | | | | |
| Piece-wise exponential | 0.00140 | 0.00098 | 0.00083 | 0.00078 | 0.00074 | 0.00071 | 0.04384 | 0.03587 | 0.03271 | 0.03120 | 0.02979 | 0.02876 | 0.00123 | 0.00087 | 0.00072 | 0.00064 | 0.00056 | 0.00051 | 0.04932 | 0.03508 | 0.02959 | 0.02542 | 0.02250 | 0.01948 |
| Pseudo-observations | 0.00144 | 0.00124 | 0.00124 | 0.00125 | 0.00129 | 0.00131 | 0.04557 | 0.03665 | 0.03359 | 0.03240 | 0.03116 | 0.03037 | 0.00139 | 0.00105 | 0.00096 | 0.00093 | 0.00092 | 0.00090 | 0.05596 | 0.03964 | 0.03299 | 0.02831 | 0.02474 | 0.02173 |
| **Weibull** | | | | | | | | | | | | | | | | | | | | | | | | |
| Piece-wise exponential | 0.00088 | 0.00076 | 0.00071 | 0.00068 | 0.00062 | 0.00057 | 0.03287 | 0.02431 | 0.02018 | 0.01752 | 0.01455 | 0.01274 | 0.00077 | 0.00064 | 0.00058 | 0.00054 | 0.00049 | 0.00045 | 0.03931 | 0.02866 | 0.02387 | 0.01953 | 0.01669 | 0.01438 |
| Pseudo-observations | 0.00133 | 0.00111 | 0.00102 | 0.00096 | 0.00089 | 0.00086 | 0.03647 | 0.02630 | 0.02168 | 0.01856 | 0.01518 | 0.01333 | 0.00107 | 0.00088 | 0.00081 | 0.00076 | 0.00071 | 0.00069 | 0.04392 | 0.03156 | 0.02688 | 0.02206 | 0.01857 | 0.01593 |

RMSE is aggregated over time and treatment categories.

#### 8.2.1.2 Performance on HRDR and HNRDR computation

In the following section we will evaluate the performance of estimation of the HRDR and HNRDR. We will first focus on the individual features of the intervals, such as the Lower boundary, the Mid-point, the Upper boundary and the Width. Then we will consider as a measure of performance the Jaccard index, which quantifies the similarity between two sets. In this case, the two sets are the estimated and the true interval. For the interval features we will focus particularly on the Mid-point and the Width, since they represent somehow their location and scale, but we also report in a dashboard and the table all the other features.

Figure 8.6 presents the estimation accuracy for the mid-point of Highest Density Regions (HDRs), as quantified by the root mean squared error. A primary observation is the superior performance of the piecewise exponential model over the pseudo-observation model across nearly all scenarios. The notable exception is the non-proportional hazards scenario, where both models exhibit comparable performance. Furthermore, the optimal HRDR construction method was found to interact with the choice of modeling approach. When the piecewise exponential model is utilized, deriving the HRDR from either the instantaneous (PDF) or the cumulative (CDF) function yields similar results. In contrast, when using the pseudo-observation model, constructing the HRDR from the cumulative function is preferable, resulting in more accurate estimates.

The analysis also explored the relationship between the HRDR's probability content and the stability of its mid-point, revealing counterintuitive results. While one might expect higher variability for smaller-probability intervals, this pattern only emerged in specific cases: the Weibull scenario, and the exponential scenario when analyzed with the pseudo-observation model. In the exponential scenario, this issue stems from numerical artifacts; the PDF, obtained by differentiating the estimated CDF, fails to remain monotonically decreasing and instead exhibits spurious modes that incorrectly shift the lower bound of the estimated HDR away from zero. For the Weibull scenario, the challenge is more fundamental and appears to represent an inherent limitation of HRDR estimation itself. The low curvature of the true probability density function of the treatment group makes the location of the region of highest density difficult to pinpoint. This problem persists when using the cumulative function, as it manifests as a nearly linear shape, again complicating the precise identification of the interval.

The analysis presented in **?@fig-perf** turns to the estimation performance for the Width of HRDR intervals, measured by RMSE. Here again, the superiority of the piecewise exponential model over the pseudo-observation model is evident across all scenarios, irrespective of the method used for HRDR computation. The choice between HRDR construction methods shows little impact when using the piecewise exponential model, although the CDF-based approach is favored in the Weibull scenario. When the pseudo-observation model is used, the method relying on the cumulative distribution function proves to be better. The reason for this is that the density function, when computed by differentiating the estimated CDF, tends to form artifactual peaks that misrepresent the probability distribution, thereby increasing the RMSE of the width. Notably, the estimation problem observed in the Weibull scenario is less severe for the interval width, which suggests the primary challenge is not defining the span of the time-window, but rather locating its center.

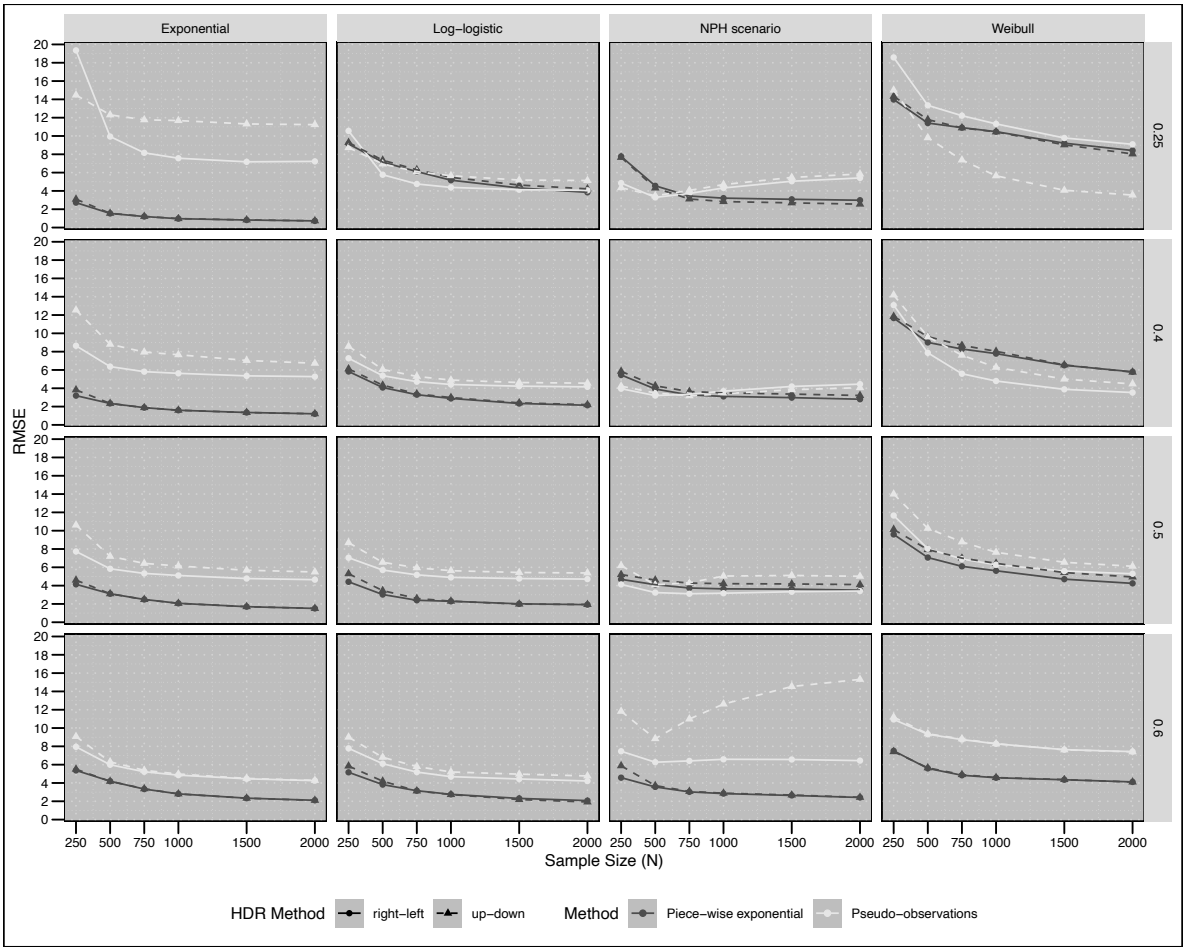



Figure 8.6: Overall Performance: RMSE vs. Sample Size, for the mid-point of the HRDR intervals for the different scenarios and different risk mass. RMSE is aggregated over treatment categories

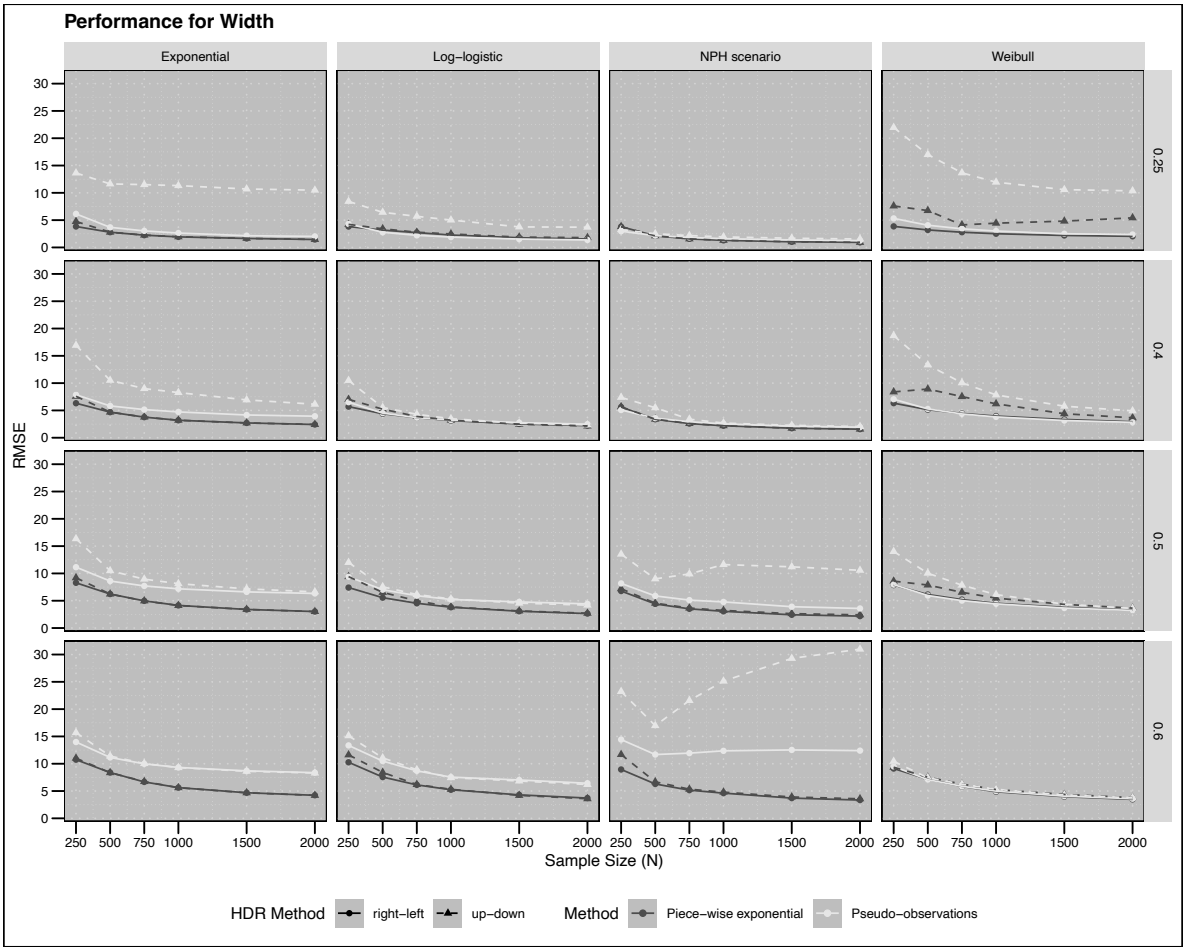



Figure 8.7: Overall Performance: RMSE vs. Sample Size, for the width of the HRDR intervals for the different scenarios and different risk mass. RMSE is aggregated over treatment categories

Table 8.7 reports the results of the RMSE above discussed for the Mid-point and the Width of the interval along with the Lower Boundary and Upper Boundary.

Table 8.7: Performances of estimation of the interval features of the HRDR. Root Mean Squared Error (RMSE) for Piece-wise exponential (Pseudo-observations) methods.

| | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N = 250 | N = 250 | N = 250 | N = 250 | N = 500 | N = 500 | N = 500 | N = 500 | N = 750 | N = 750 | N = 750 | N = 750 | N = 1000 | N = 1000 | N = 1000 | N = 1000 | N = 1500 | N = 1500 | N = 1500 | N = 1500 | N = 2000 | N = 2000 | N = 2000 | N = 2000 |
| **Lower boundary** | | | | | | | | | | | | | | | | | | | | | | | | |
| 0.25 - right-left | 1.67 (17.32) | 8.78 ( 9.92) | 8.39 ( 4.60) | 14.01 (18.80) | 0.62 ( 9.10) | 6.97 ( 5.82) | 4.72 ( 3.42) | 11.47 (13.38) | 0.38 ( 7.46) | 6.02 ( 4.86) | 3.54 ( 3.98) | 10.94 (12.19) | 0.07 ( 6.92) | 5.13 ( 4.52) | 3.30 ( 4.55) | 10.52 (11.38) | 0.00 ( 6.63) | 4.31 ( 4.27) | 3.16 ( 5.21) | 9.25 ( 9.88) | 0.00 ( 6.68) | 3.84 ( 4.22) | 3.06 ( 5.52) | 8.46 ( 9.18) |
| 0.25 - up-down | 1.73 (11.20) | 9.05 ( 7.94) | 8.29 ( 4.48) | 14.14 (12.82) | 0.65 ( 8.86) | 7.24 ( 6.27) | 4.44 ( 3.83) | 11.53 ( 8.72) | 0.40 ( 8.01) | 6.31 ( 5.48) | 3.15 ( 4.55) | 10.89 ( 7.75) | 0.08 ( 7.81) | 5.51 ( 5.19) | 2.84 ( 5.16) | 10.40 ( 7.06) | 0.00 ( 7.62) | 4.71 ( 4.99) | 2.68 ( 5.86) | 9.03 ( 6.55) | 0.00 ( 7.60) | 4.30 ( 4.93) | 2.54 ( 6.19) | 8.15 ( 6.32) |
| 0.4 - right-left | 0.42 ( 6.23) | 5.46 ( 6.61) | 6.39 ( 4.13) | 11.83 (13.19) | 0.04 ( 4.57) | 4.00 ( 5.01) | 4.11 ( 3.38) | 9.23 ( 7.87) | 0.02 ( 4.02) | 3.40 ( 4.45) | 3.39 ( 3.49) | 8.51 ( 5.53) | 0.00 ( 3.90) | 3.02 ( 4.27) | 3.22 ( 3.76) | 8.02 ( 4.74) | 0.00 ( 3.71) | 2.50 ( 4.04) | 3.06 ( 4.11) | 6.70 ( 3.67) | 0.00 ( 3.65) | 2.30 ( 3.99) | 2.89 ( 4.29) | 5.96 ( 3.34) |
| 0.4 - up-down | 0.43 ( 6.52) | 5.50 ( 6.55) | 6.90 ( 4.00) | 11.48 ( 9.76) | 0.04 ( 5.22) | 4.15 ( 5.25) | 4.49 ( 3.26) | 8.89 ( 6.48) | 0.02 ( 4.73) | 3.50 ( 4.70) | 3.77 ( 3.34) | 8.12 ( 5.31) | 0.00 ( 4.63) | 3.09 ( 4.54) | 3.61 ( 3.58) | 7.61 ( 4.62) | 0.00 ( 4.47) | 2.50 ( 4.34) | 3.45 ( 3.92) | 6.31 ( 3.72) | 0.00 ( 4.44) | 2.26 ( 4.29) | 3.30 ( 4.07) | 5.55 ( 3.39) |
| 0.5 - right-left | 0.04 ( 3.43) | 3.23 ( 5.07) | 5.27 ( 3.69) | 9.46 (11.18) | 0.00 ( 2.52) | 2.22 ( 4.08) | 4.11 ( 2.94) | 7.01 ( 7.49) | 0.00 ( 2.20) | 1.83 ( 3.77) | 3.73 ( 2.94) | 6.11 ( 6.24) | 0.00 ( 2.10) | 1.76 ( 3.69) | 3.64 ( 3.23) | 5.65 ( 5.66) | 0.00 ( 1.90) | 1.56 ( 3.50) | 3.58 ( 3.25) | 4.68 ( 4.90) | 0.00 ( 1.83) | 1.53 ( 3.50) | 3.49 ( 3.17) | 4.17 ( 4.68) |
| 0.5 - up-down | 0.04 ( 3.99) | 3.35 ( 5.75) | 5.83 ( 3.62) | 9.60 (10.28) | 0.00 ( 3.22) | 2.32 ( 4.86) | 4.60 ( 2.87) | 7.12 ( 7.79) | 0.00 ( 2.95) | 1.95 ( 4.59) | 4.21 ( 2.74) | 6.32 ( 6.89) | 0.00 ( 2.90) | 1.89 ( 4.53) | 4.14 ( 2.88) | 5.89 ( 6.38) | 0.00 ( 2.71) | 1.66 ( 4.36) | 4.08 ( 2.95) | 4.98 ( 5.66) | 0.00 ( 2.66) | 1.67 ( 4.37) | 4.01 ( 2.92) | 4.50 ( 5.45) |
| 0.6 - right-left | 0.00 ( 2.09) | 1.48 ( 3.64) | 3.39 ( 3.30) | 7.28 (11.03) | 0.00 ( 1.06) | 0.91 ( 2.57) | 2.97 ( 2.51) | 5.45 ( 8.96) | 0.00 ( 0.66) | 0.77 ( 2.20) | 2.75 ( 1.99) | 4.69 ( 8.11) | 0.00 ( 0.54) | 0.79 ( 2.05) | 2.78 ( 1.69) | 4.48 ( 7.63) | 0.00 ( 0.34) | 0.73 ( 1.78) | 2.77 ( 1.39) | 4.23 ( 6.88) | 0.00 ( 0.29) | 0.71 ( 1.74) | 2.69 ( 1.29) | 3.80 ( 6.69) |
| 0.6 - up-down | 0.00 ( 2.42) | 1.54 ( 4.25) | 3.83 ( 3.49) | 7.20 (10.23) | 0.00 ( 1.45) | 0.97 ( 3.25) | 3.37 ( 2.59) | 5.55 ( 8.67) | 0.00 ( 1.01) | 0.83 ( 2.93) | 3.12 ( 1.97) | 4.87 ( 7.95) | 0.00 ( 0.88) | 0.85 ( 2.81) | 3.15 ( 1.65) | 4.66 ( 7.54) | 0.00 ( 0.60) | 0.76 ( 2.55) | 3.12 ( 1.33) | 4.42 ( 6.81) | 0.00 ( 0.52) | 0.75 ( 2.53) | 3.03 ( 1.22) | 3.90 ( 6.63) |
| **Mid-Point** | | | | | | | | | | | | | | | | | | | | | | | | |
| 0.25 - right-left | 2.72 (19.35) | 9.13 (10.56) | 7.76 ( 4.86) | 13.99 (18.58) | 1.54 ( 9.94) | 7.14 ( 5.77) | 4.56 ( 3.29) | 11.41 (13.34) | 1.20 ( 8.16) | 6.12 ( 4.76) | 3.44 ( 3.72) | 10.91 (12.22) | 0.97 ( 7.57) | 5.17 ( 4.38) | 3.21 ( 4.32) | 10.47 (11.32) | 0.82 ( 7.18) | 4.32 ( 4.13) | 3.08 ( 5.06) | 9.21 ( 9.78) | 0.73 ( 7.22) | 3.85 ( 4.06) | 2.98 ( 5.42) | 8.41 ( 9.09) |
| 0.25 - up-down | 3.09 (14.47) | 9.28 ( 8.73) | 7.66 ( 4.35) | 14.33 (15.01) | 1.56 (12.31) | 7.35 ( 6.94) | 4.32 ( 3.47) | 11.79 ( 9.81) | 1.21 (11.77) | 6.31 ( 6.10) | 3.12 ( 4.06) | 10.89 ( 7.38) | 0.98 (11.68) | 5.48 ( 5.66) | 2.84 ( 4.70) | 10.45 ( 5.65) | 0.83 (11.32) | 4.64 ( 5.19) | 2.70 ( 5.47) | 9.04 ( 4.07) | 0.73 (11.24) | 4.22 ( 5.14) | 2.56 ( 5.85) | 8.04 ( 3.55) |
| 0.4 - right-left | 3.20 ( 8.64) | 5.83 ( 7.28) | 5.46 ( 3.96) | 11.67 (13.08) | 2.34 ( 6.37) | 4.08 ( 5.38) | 3.92 ( 3.20) | 9.01 ( 7.87) | 1.89 ( 5.81) | 3.30 ( 4.72) | 3.28 ( 3.34) | 8.29 ( 5.59) | 1.60 ( 5.63) | 2.89 ( 4.41) | 3.11 ( 3.69) | 7.78 ( 4.79) | 1.36 ( 5.36) | 2.34 ( 4.23) | 2.98 ( 4.20) | 6.51 ( 3.88) | 1.21 ( 5.27) | 2.15 ( 4.16) | 2.81 ( 4.45) | 5.80 ( 3.53) |
| 0.4 - up-down | 3.82 (12.55) | 6.12 ( 8.58) | 5.85 ( 4.23) | 11.84 (14.19) | 2.35 ( 8.81) | 4.34 ( 6.06) | 4.27 ( 3.49) | 9.64 ( 9.62) | 1.89 ( 7.95) | 3.39 ( 5.29) | 3.64 ( 3.17) | 8.69 ( 7.61) | 1.60 ( 7.67) | 2.99 ( 4.90) | 3.50 ( 3.39) | 8.04 ( 6.29) | 1.35 ( 7.04) | 2.41 ( 4.63) | 3.36 ( 3.84) | 6.55 ( 5.00) | 1.21 ( 6.73) | 2.21 ( 4.56) | 3.21 ( 4.07) | 5.76 ( 4.47) |
| 0.5 - right-left | 4.14 ( 7.73) | 4.42 ( 7.05) | 4.69 ( 4.13) | 9.59 (11.66) | 3.11 ( 5.84) | 3.03 ( 5.70) | 4.10 ( 3.24) | 7.07 ( 8.03) | 2.49 ( 5.33) | 2.40 ( 5.18) | 3.74 ( 3.11) | 6.11 ( 6.91) | 2.07 ( 5.11) | 2.29 ( 4.90) | 3.64 ( 3.17) | 5.62 ( 6.28) | 1.70 ( 4.78) | 2.01 ( 4.77) | 3.62 ( 3.33) | 4.71 ( 5.59) | 1.51 ( 4.65) | 1.92 ( 4.71) | 3.51 ( 3.41) | 4.25 ( 5.29) |
| 0.5 - up-down | 4.60 (10.62) | 5.30 ( 8.69) | 5.21 ( 6.18) | 10.14 (14.00) | 3.11 ( 7.20) | 3.45 ( 6.55) | 4.61 ( 4.09) | 7.93 (10.26) | 2.48 ( 6.43) | 2.60 ( 5.95) | 4.28 ( 4.32) | 7.01 ( 8.77) | 2.06 ( 6.14) | 2.30 ( 5.63) | 4.21 ( 5.04) | 6.44 ( 7.64) | 1.70 ( 5.68) | 1.98 ( 5.44) | 4.18 ( 5.11) | 5.43 ( 6.55) | 1.51 ( 5.49) | 1.94 ( 5.38) | 4.11 ( 5.02) | 4.94 ( 6.11) |
| 0.6 - right-left | 5.37 ( 7.95) | 5.17 ( 7.78) | 4.58 ( 7.47) | 7.51 (10.94) | 4.19 ( 5.99) | 3.82 ( 6.11) | 3.57 ( 6.27) | 5.59 ( 9.32) | 3.33 ( 5.22) | 3.15 ( 5.19) | 3.03 ( 6.41) | 4.84 ( 8.74) | 2.81 ( 4.86) | 2.76 ( 4.69) | 2.84 ( 6.59) | 4.57 ( 8.27) | 2.34 ( 4.45) | 2.31 ( 4.44) | 2.64 ( 6.57) | 4.35 ( 7.64) | 2.11 ( 4.26) | 2.07 ( 4.22) | 2.42 ( 6.44) | 4.12 ( 7.43) |
| 0.6 - up-down | 5.51 ( 9.08) | 5.85 ( 9.00) | 5.88 (11.83) | 7.42 (11.19) | 4.18 ( 6.31) | 4.20 ( 6.81) | 3.71 ( 8.84) | 5.65 ( 9.37) | 3.37 ( 5.38) | 3.13 ( 5.80) | 3.08 (10.99) | 4.91 ( 8.81) | 2.81 ( 4.98) | 2.73 ( 5.20) | 2.89 (12.62) | 4.59 ( 8.29) | 2.34 ( 4.50) | 2.19 ( 4.96) | 2.69 (14.54) | 4.37 ( 7.62) | 2.10 ( 4.28) | 1.93 ( 4.77) | 2.43 (15.31) | 4.10 ( 7.40) |
| **Upper boundary** | | | | | | | | | | | | | | | | | | | | | | | | |
| 0.25 - right-left | 4.40 (21.63) | 9.85 (11.56) | 7.58 ( 5.52) | 14.23 (18.75) | 2.87 (11.02) | 7.64 ( 6.05) | 4.65 ( 3.51) | 11.58 (13.62) | 2.30 ( 9.06) | 6.50 ( 4.91) | 3.52 ( 3.69) | 11.06 (12.49) | 1.93 ( 8.38) | 5.46 ( 4.45) | 3.26 ( 4.24) | 10.58 (11.46) | 1.65 ( 7.84) | 4.52 ( 4.13) | 3.10 ( 5.00) | 9.31 ( 9.85) | 1.45 ( 7.87) | 4.00 ( 4.03) | 2.98 ( 5.38) | 8.49 ( 9.16) |
| 0.25 - up-down | 5.24 (19.65) | 9.97 (11.18) | 7.49 ( 4.80) | 15.49 (22.97) | 2.89 (17.09) | 7.84 ( 8.82) | 4.46 ( 3.54) | 12.95 (16.15) | 2.31 (16.71) | 6.61 ( 7.79) | 3.29 ( 3.82) | 11.26 (11.92) | 1.95 (16.62) | 5.72 ( 7.05) | 2.98 ( 4.39) | 10.95 ( 9.25) | 1.66 (15.98) | 4.76 ( 6.00) | 2.81 ( 5.19) | 9.66 ( 6.81) | 1.46 (15.81) | 4.33 ( 5.94) | 2.65 ( 5.59) | 8.81 ( 6.24) |
| 0.4 - right-left | 6.34 (11.90) | 7.38 ( 9.14) | 5.87 ( 5.21) | 12.35 (13.89) | 4.69 ( 8.79) | 5.19 ( 6.55) | 4.42 ( 3.94) | 9.50 ( 8.71) | 3.77 ( 8.05) | 4.13 ( 5.62) | 3.65 ( 3.79) | 8.66 ( 6.43) | 3.20 ( 7.71) | 3.54 ( 5.09) | 3.38 ( 4.03) | 8.04 ( 5.54) | 2.71 ( 7.23) | 2.81 ( 4.82) | 3.14 ( 4.54) | 6.75 ( 4.65) | 2.42 ( 7.08) | 2.52 ( 4.66) | 2.94 ( 4.80) | 6.00 ( 4.22) |
| 0.4 - up-down | 7.60 (20.38) | 8.33 (12.62) | 6.08 ( 6.87) | 13.57 (21.96) | 4.69 (13.53) | 5.84 ( 7.82) | 4.70 ( 5.36) | 12.11 (15.23) | 3.78 (12.02) | 4.28 ( 6.57) | 3.96 ( 3.84) | 10.65 (11.75) | 3.20 (11.41) | 3.67 ( 5.78) | 3.73 ( 3.71) | 9.51 ( 9.41) | 2.71 (10.16) | 2.90 ( 5.25) | 3.50 ( 4.10) | 7.47 ( 7.27) | 2.43 ( 9.47) | 2.64 ( 5.09) | 3.32 ( 4.31) | 6.49 ( 6.38) |
| 0.5 - right-left | 8.28 (13.04) | 7.49 (10.85) | 6.25 ( 7.34) | 11.21 (13.36) | 6.23 ( 9.94) | 5.37 ( 8.54) | 5.15 ( 5.45) | 8.34 ( 9.52) | 4.98 ( 9.04) | 4.30 ( 7.55) | 4.51 ( 4.88) | 7.13 ( 8.32) | 4.14 ( 8.58) | 3.83 ( 6.93) | 4.24 ( 4.59) | 6.45 ( 7.53) | 3.40 ( 7.99) | 3.23 ( 6.69) | 4.03 ( 4.39) | 5.44 ( 6.73) | 3.02 ( 7.74) | 2.93 ( 6.48) | 3.85 ( 4.43) | 4.91 ( 6.27) |
| 0.5 - up-down | 9.20 (18.53) | 9.45 (13.78) | 6.80 (12.44) | 12.26 (19.61) | 6.22 (12.18) | 6.30 ( 9.52) | 5.63 ( 8.11) | 10.31 (14.15) | 4.97 (10.68) | 4.70 ( 8.28) | 5.06 ( 8.88) | 8.93 (11.72) | 4.12 (10.00) | 3.80 ( 7.56) | 4.85 (10.49) | 7.95 ( 9.74) | 3.41 ( 9.11) | 3.16 ( 7.10) | 4.68 (10.32) | 6.60 ( 7.95) | 3.02 ( 8.67) | 2.90 ( 6.90) | 4.55 ( 9.89) | 5.94 ( 7.18) |
| 0.6 - right-left | 10.75 (14.82) | 10.19 (14.04) | 8.40 (14.31) | 10.06 (12.76) | 8.39 (11.53) | 7.55 (11.09) | 6.03 (11.85) | 7.64 (10.89) | 6.67 (10.20) | 6.17 ( 9.32) | 4.91 (12.23) | 6.49 (10.21) | 5.62 ( 9.50) | 5.32 ( 8.26) | 4.37 (12.66) | 5.80 ( 9.55) | 4.68 ( 8.80) | 4.41 ( 7.80) | 3.62 (12.76) | 5.28 ( 8.81) | 4.22 ( 8.43) | 3.87 ( 7.29) | 3.17 (12.58) | 5.04 ( 8.49) |
| 0.6 - up-down | 11.01 (16.80) | 11.56 (16.07) | 11.08 (23.18) | 10.10 (14.15) | 8.36 (11.94) | 8.35 (11.97) | 6.21 (17.14) | 7.84 (11.32) | 6.73 (10.35) | 6.17 ( 9.91) | 4.85 (21.68) | 6.65 (10.57) | 5.61 ( 9.59) | 5.32 ( 8.60) | 4.27 (25.13) | 5.86 ( 9.73) | 4.67 ( 8.79) | 4.23 ( 8.13) | 3.51 (29.15) | 5.31 ( 8.88) | 4.21 ( 8.41) | 3.64 ( 7.63) | 3.01 (30.78) | 5.07 ( 8.52) |
| **Width** | | | | | | | | | | | | | | | | | | | | | | | | |
| 0.25 - right-left | 3.83 ( 6.14) | 3.83 ( 4.30) | 3.80 ( 2.97) | 3.86 ( 5.32) | 2.78 ( 3.68) | 3.17 ( 2.78) | 2.13 ( 2.17) | 3.20 ( 4.07) | 2.25 ( 3.03) | 2.72 ( 2.21) | 1.56 ( 1.83) | 2.78 ( 3.41) | 1.93 ( 2.63) | 2.27 ( 1.90) | 1.30 ( 1.60) | 2.51 ( 3.00) | 1.65 ( 2.16) | 1.77 ( 1.53) | 1.04 ( 1.37) | 2.20 ( 2.51) | 1.45 ( 2.02) | 1.55 ( 1.37) | 0.93 ( 1.25) | 2.04 ( 2.37) |
| 0.25 - up-down | 4.75 (13.63) | 4.23 ( 8.43) | 3.81 ( 3.25) | 7.59 (21.97) | 2.79 (11.63) | 3.41 ( 6.43) | 2.13 ( 2.50) | 6.75 (17.01) | 2.26 (11.50) | 2.78 ( 5.70) | 1.57 ( 2.18) | 4.13 (13.66) | 1.95 (11.32) | 2.47 ( 5.02) | 1.30 ( 1.93) | 4.44 (11.95) | 1.66 (10.70) | 1.89 ( 3.76) | 1.04 ( 1.68) | 4.82 (10.59) | 1.46 (10.49) | 1.79 ( 3.71) | 0.93 ( 1.54) | 5.42 (10.36) |
| 0.4 - right-left | 6.32 ( 7.87) | 5.69 ( 6.51) | 5.62 ( 5.08) | 6.33 ( 7.05) | 4.69 ( 5.83) | 4.39 ( 4.51) | 3.39 ( 3.59) | 5.14 ( 5.28) | 3.77 ( 5.16) | 3.69 ( 3.68) | 2.60 ( 2.92) | 4.45 ( 4.36) | 3.20 ( 4.72) | 3.14 ( 3.24) | 2.20 ( 2.54) | 3.97 ( 3.82) | 2.71 ( 4.17) | 2.53 ( 2.76) | 1.74 ( 2.12) | 3.34 ( 3.15) | 2.42 ( 3.94) | 2.22 ( 2.48) | 1.57 ( 1.94) | 2.93 ( 2.82) |
| 0.4 - up-down | 7.58 (16.92) | 7.02 (10.49) | 5.67 ( 7.39) | 8.39 (18.69) | 4.69 (10.51) | 5.23 ( 5.54) | 3.40 ( 5.49) | 8.92 (13.34) | 3.78 ( 9.00) | 3.89 ( 4.29) | 2.60 ( 3.39) | 7.53 (10.05) | 3.20 ( 8.27) | 3.21 ( 3.48) | 2.20 ( 2.69) | 6.21 ( 7.84) | 2.71 ( 6.93) | 2.48 ( 2.67) | 1.74 ( 2.27) | 4.39 ( 5.78) | 2.43 ( 6.13) | 2.15 ( 2.34) | 1.57 ( 2.06) | 3.65 ( 4.92) |
| 0.5 - right-left | 8.28 (11.14) | 7.41 ( 9.38) | 6.78 ( 8.17) | 7.87 ( 7.95) | 6.23 ( 8.61) | 5.56 ( 7.02) | 4.41 ( 5.88) | 6.14 ( 5.95) | 4.98 ( 7.71) | 4.54 ( 5.92) | 3.52 ( 5.11) | 5.21 ( 5.02) | 4.14 ( 7.18) | 3.81 ( 5.23) | 3.06 ( 4.78) | 4.53 ( 4.43) | 3.40 ( 6.62) | 3.08 ( 4.78) | 2.43 ( 3.91) | 3.79 ( 3.66) | 3.02 ( 6.33) | 2.65 ( 4.43) | 2.18 ( 3.58) | 3.28 ( 3.23) |
| 0.5 - up-down | 9.20 (16.36) | 9.41 (12.01) | 7.19 (13.53) | 8.59 (14.01) | 6.22 (10.50) | 6.52 ( 7.55) | 4.55 ( 9.01) | 7.88 (10.04) | 4.97 ( 8.95) | 4.99 ( 6.14) | 3.66 ( 9.91) | 6.54 ( 7.85) | 4.12 ( 8.13) | 3.86 ( 5.32) | 3.22 (11.62) | 5.48 ( 6.12) | 3.41 ( 7.17) | 3.13 ( 4.56) | 2.63 (11.21) | 4.34 ( 4.37) | 3.02 ( 6.62) | 2.71 ( 4.19) | 2.41 (10.57) | 3.65 ( 3.60) |
| 0.6 - right-left | 10.75 (13.98) | 10.27 (13.35) | 8.95 (14.43) | 9.12 ( 9.50) | 8.39 (11.15) | 7.56 (10.49) | 6.29 (11.68) | 7.15 ( 7.08) | 6.67 ( 9.98) | 6.13 ( 8.70) | 5.16 (11.94) | 5.86 ( 5.87) | 5.62 ( 9.32) | 5.23 ( 7.55) | 4.62 (12.36) | 4.87 ( 4.99) | 4.68 ( 8.69) | 4.30 ( 7.02) | 3.71 (12.53) | 3.96 ( 4.06) | 4.22 ( 8.37) | 3.71 ( 6.42) | 3.35 (12.40) | 3.47 ( 3.59) |
| 0.6 - up-down | 11.01 (15.70) | 11.62 (15.12) | 11.69 (23.23) | 9.36 (10.44) | 8.36 (11.41) | 8.41 (11.07) | 6.69 (16.98) | 7.53 ( 7.43) | 6.73 (10.03) | 6.20 ( 8.90) | 5.34 (21.57) | 6.27 ( 6.26) | 5.61 ( 9.29) | 5.31 ( 7.45) | 4.80 (25.13) | 5.27 ( 5.32) | 4.67 ( 8.61) | 4.20 ( 6.84) | 3.91 (29.29) | 4.38 ( 4.28) | 4.21 ( 8.29) | 3.58 ( 6.20) | 3.57 (30.98) | 3.80 ( 3.77) |

RMSE is aggregated over time and treatment categories.

The evaluation of the HNRDR estimation performance was performed considering different net risk difference content of the intervals (-0.25, -0.2, -0.15, -0.1, and -0.05). However, the evaluation strictly depends on the maximum cumulative risk difference between the marginal functions achieved in each scenario. Namely, while the Weibull scenario showed a maximum cumulative risk difference of ~ -0.35, the maximum cumulative risk difference of the NPH scenario reached ~0.15. We felt confident in evaluating the performances of the flexible estimation methods for probability levels easily achieavable for a large amount of M replications. This is the reason why for the Weibull scenario, reuslts up to -0.25 HNRDR intervals will be shown, whereas for the NPH scenario we will show the results up to -0.1 HNRDR intervals.

Regarding the performance of the estimation of the mid-point HNRDR in Figure 8.8, the piece-wise exponential model is again performing better than the pseudo-observation models in locating the interval. The choice between HRDR construction methods shows little impact when using the piecewise exponential model or the pseudo-observation model. Notably, the problem for the Weibull scenario of not being able to locate correctly the center of the intervals does not apply for HNRDR. Even in this case, the lower is the amount of net risk difference the (slightly) higher is the RMSE around the mid-point, regardeless the flexible estimation method or the method of constructin of the HNRDR.

The analysis presented in Figure 8.9 turns to the estimation performance for the width of HNRDR intervals. Here, the superiority of the piecewise exponential model over the pseudo-observation model is less evident across all scenarios, irrespective of the method used for HRDR computation. Again, the estimation problem observed in the Weibull scenario for the HRDR mid-point and width does not apply for the interval Mid-point and Width of the HNRDR.

Table 8.8 reports the results of the RMSE above discussed for the Mid-point and the Width of the interval along with the Lower Boundary and Upper Boundary.

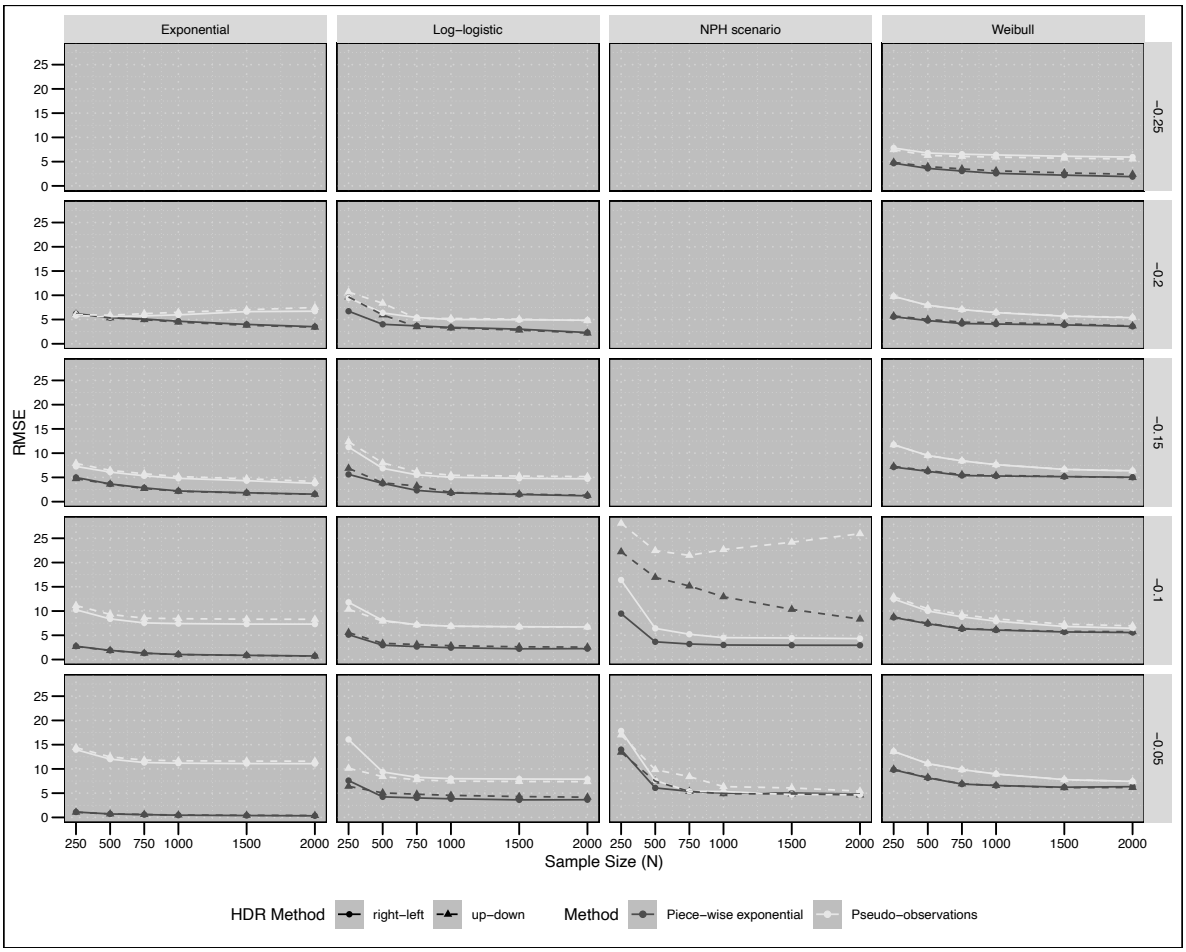



Figure 8.8: Overall Performance: RMSE vs. Sample Size, for the mid-point of the HNRDR intervals for the different scenarios and different risk mass.

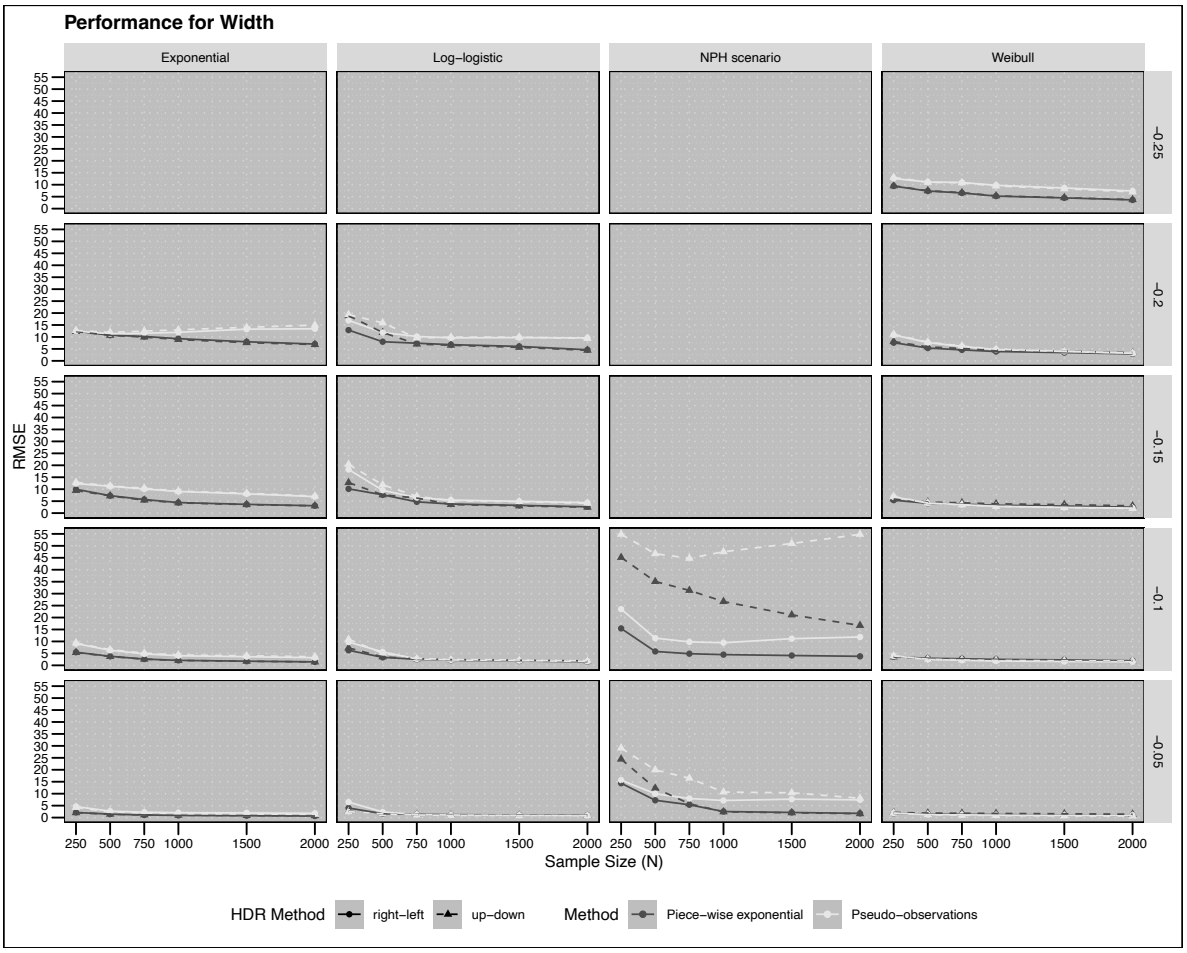



Figure 8.9: Overall Performance: RMSE vs. Sample Size, for the width of the HNRDR intervals for the different scenarios and different risk mass.

Table 8.8: Performances of estimation of the interval features of the HNRDR. Root Mean Squared Error (RMSE) for Piece-wise exponential (Pseudo-observations) methods.

| | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N = 250 | N = 250 | N = 250 | N = 250 | N = 500 | N = 500 | N = 500 | N = 500 | N = 750 | N = 750 | N = 750 | N = 750 | N = 1000 | N = 1000 | N = 1000 | N = 1000 | N = 1500 | N = 1500 | N = 1500 | N = 1500 | N = 2000 | N = 2000 | N = 2000 | N = 2000 |
| **Lower boundary** | | | | | | | | | | | | | | | | | | | | | | | | |
| -0.05 - right-left | 0.28 (12.47) | 6.36 (13.77) | 11.57 (16.57) | 9.72 (13.41) | 0.00 (11.02) | 4.14 ( 8.95) | 5.88 ( 5.89) | 8.09 (11.18) | 0.16 (10.43) | 3.95 ( 7.95) | 5.60 ( 4.50) | 6.85 (10.04) | 0.00 (10.34) | 3.83 ( 7.71) | 5.51 ( 4.29) | 6.56 ( 9.21) | 0.00 (10.27) | 3.68 ( 7.63) | 5.51 ( 3.13) | 6.20 ( 8.11) | 0.00 (10.29) | 3.70 ( 7.61) | 5.23 ( 2.62) | 6.28 ( 7.81) |
| -0.05 - up-down | 0.32 (13.15) | 5.85 ( 9.65) | 8.19 (10.57) | 9.90 (13.18) | 0.00 (11.72) | 4.83 ( 8.19) | 5.71 ( 4.82) | 8.24 (10.97) | 0.19 (11.19) | 4.60 ( 7.68) | 5.35 ( 4.40) | 6.92 ( 9.78) | 0.00 (11.08) | 4.45 ( 7.44) | 5.22 ( 4.00) | 6.57 ( 8.96) | 0.00 (11.03) | 4.29 ( 7.36) | 5.17 ( 3.46) | 6.03 ( 7.88) | 0.00 (11.04) | 4.20 ( 7.35) | 4.86 ( 2.99) | 5.94 ( 7.55) |
| -0.1 - right-left | 0.19 ( 7.79) | 3.57 ( 9.39) | 5.32 (11.91) | 8.66 (11.94) | 0.00 ( 6.66) | 2.48 ( 7.07) | 2.74 ( 4.25) | 7.46 (10.00) | 0.06 ( 6.14) | 2.35 ( 6.69) | 2.61 ( 3.27) | 6.50 ( 8.94) | 0.00 ( 6.08) | 2.30 ( 6.46) | 2.72 ( 2.94) | 6.23 ( 8.19) | 0.00 ( 5.97) | 2.24 ( 6.34) | 2.79 ( 2.52) | 5.81 ( 7.20) | 0.00 ( 6.05) | 2.28 ( 6.31) | 2.85 ( 2.44) | 5.52 ( 6.92) |
| -0.1 - up-down | 0.19 ( 8.46) | 3.96 ( 8.21) | 5.06 ( 9.93) | 8.70 (12.62) | 0.00 ( 7.38) | 3.11 ( 6.83) | 2.76 ( 4.47) | 7.51 (10.71) | 0.06 ( 6.88) | 3.00 ( 6.43) | 2.62 ( 3.58) | 6.52 ( 9.62) | 0.00 ( 6.85) | 2.91 ( 6.21) | 2.70 ( 3.19) | 6.27 ( 8.89) | 0.00 ( 6.74) | 2.86 ( 6.08) | 2.76 ( 2.76) | 5.91 ( 7.92) | 0.00 ( 6.80) | 2.77 ( 6.07) | 2.82 ( 2.67) | 5.75 ( 7.66) |
| -0.15 - right-left | 0.06 ( 3.88) | 2.28 ( 6.25) | - (-) | 7.14 (10.98) | 0.00 ( 2.58) | 1.34 ( 4.57) | - (-) | 6.44 ( 9.41) | 0.00 ( 2.01) | 1.20 ( 4.04) | - (-) | 5.77 ( 8.47) | 0.00 ( 1.73) | 1.17 ( 3.73) | - (-) | 5.66 ( 7.87) | 0.00 ( 1.47) | 1.11 ( 3.47) | - (-) | 5.52 ( 7.01) | 0.00 ( 1.31) | 1.05 ( 3.36) | - (-) | 5.18 ( 6.81) |
| -0.15 - up-down | 0.06 ( 4.39) | 2.35 ( 6.83) | - (-) | 7.48 (10.74) | 0.00 ( 3.08) | 1.38 ( 5.15) | - (-) | 6.81 ( 9.20) | 0.00 ( 2.50) | 1.26 ( 4.69) | - (-) | 6.16 ( 8.25) | 0.00 ( 2.22) | 1.23 ( 4.38) | - (-) | 5.98 ( 7.64) | 0.00 ( 1.95) | 1.17 ( 4.14) | - (-) | 5.65 ( 6.79) | 0.00 ( 1.79) | 1.13 ( 4.03) | - (-) | 5.08 ( 6.58) |
| -0.2 - right-left | 0.00 ( 1.40) | 1.43 ( 4.57) | - (-) | 5.53 ( 8.77) | 0.00 ( 0.62) | 0.45 ( 2.72) | - (-) | 5.17 ( 7.39) | 0.00 ( 0.24) | 0.29 ( 1.97) | - (-) | 4.90 ( 6.59) | 0.00 ( 0.14) | 0.18 ( 1.42) | - (-) | 4.82 ( 6.09) | 0.00 ( 0.00) | 0.13 ( 1.00) | - (-) | 4.64 ( 5.42) | 0.00 ( 0.00) | 0.09 ( 0.66) | - (-) | 4.16 ( 5.22) |
| -0.2 - up-down | 0.00 ( 1.67) | 1.46 ( 4.99) | - (-) | 6.02 ( 8.59) | 0.00 ( 0.75) | 0.48 ( 3.11) | - (-) | 5.74 ( 7.17) | 0.00 ( 0.29) | 0.30 ( 2.38) | - (-) | 5.55 ( 6.38) | 0.00 ( 0.19) | 0.20 ( 1.77) | - (-) | 5.40 ( 5.94) | 0.00 ( 0.00) | 0.14 ( 1.26) | - (-) | 5.09 ( 5.22) | 0.00 ( 0.00) | 0.13 ( 0.86) | - (-) | 4.37 ( 5.05) |
| -0.25 - right-left | - (-) | - (-) | - (-) | 3.38 ( 6.52) | - (-) | - (-) | - (-) | 3.11 ( 5.07) | - (-) | - (-) | - (-) | 2.97 ( 4.52) | - (-) | - (-) | - (-) | 2.91 ( 4.20) | - (-) | - (-) | - (-) | 2.74 ( 3.75) | - (-) | - (-) | - (-) | 2.51 ( 3.69) |
| -0.25 - up-down | - (-) | - (-) | - (-) | 3.95 ( 6.40) | - (-) | - (-) | - (-) | 3.84 ( 5.00) | - (-) | - (-) | - (-) | 3.78 ( 4.44) | - (-) | - (-) | - (-) | 3.71 ( 4.12) | - (-) | - (-) | - (-) | 3.45 ( 3.67) | - (-) | - (-) | - (-) | 3.17 ( 3.57) |
| **Mid-Point** | | | | | | | | | | | | | | | | | | | | | | | | |
| -0.05 - right-left | 1.11 (13.97) | 7.60 (16.06) | 13.98 (17.83) | 9.82 (13.61) | 0.72 (12.02) | 4.24 ( 9.38) | 6.08 ( 7.61) | 8.16 (11.09) | 0.57 (11.32) | 4.02 ( 8.28) | 5.38 ( 5.39) | 6.87 ( 9.86) | 0.46 (11.18) | 3.84 ( 7.99) | 5.01 ( 5.33) | 6.58 ( 8.95) | 0.39 (11.10) | 3.62 ( 7.92) | 5.03 ( 4.72) | 6.22 ( 7.79) | 0.33 (11.08) | 3.65 ( 7.89) | 4.82 ( 4.68) | 6.34 ( 7.43) |
| -0.05 - up-down | 1.07 (14.40) | 6.42 (10.12) | 13.43 (17.01) | 9.94 (13.60) | 0.71 (12.47) | 5.06 ( 8.42) | 7.46 ( 9.85) | 8.24 (11.10) | 0.60 (11.83) | 4.78 ( 7.76) | 5.36 ( 8.41) | 6.92 ( 9.84) | 0.49 (11.67) | 4.55 ( 7.48) | 4.88 ( 6.36) | 6.59 ( 8.94) | 0.43 (11.61) | 4.30 ( 7.40) | 4.85 ( 6.08) | 6.13 ( 7.79) | 0.39 (11.59) | 4.17 ( 7.37) | 4.61 ( 5.42) | 6.17 ( 7.43) |
| -0.1 - right-left | 2.75 (10.26) | 5.12 (11.80) | 9.47 (16.40) | 8.68 (12.47) | 1.89 ( 8.38) | 2.95 ( 8.06) | 3.65 ( 6.42) | 7.38 (10.02) | 1.31 ( 7.55) | 2.66 ( 7.17) | 3.20 ( 5.21) | 6.33 ( 8.80) | 1.04 ( 7.41) | 2.44 ( 6.88) | 3.00 ( 4.51) | 6.08 ( 7.91) | 0.87 ( 7.31) | 2.21 ( 6.76) | 2.95 ( 4.44) | 5.69 ( 6.81) | 0.73 ( 7.30) | 2.25 ( 6.72) | 2.94 ( 4.32) | 5.57 ( 6.45) |
| -0.1 - up-down | 2.69 (11.13) | 5.56 (10.36) | 22.22 (28.09) | 8.78 (12.92) | 1.85 ( 9.31) | 3.34 ( 7.93) | 16.94 (22.46) | 7.48 (10.47) | 1.30 ( 8.54) | 3.10 ( 7.16) | 15.16 (21.48) | 6.40 ( 9.26) | 1.03 ( 8.43) | 2.85 ( 6.86) | 12.94 (22.67) | 6.15 ( 8.38) | 0.85 ( 8.32) | 2.62 ( 6.75) | 10.34 (24.19) | 5.79 ( 7.28) | 0.72 ( 8.30) | 2.56 ( 6.71) | 8.33 (25.95) | 5.74 ( 6.96) |
| -0.15 - right-left | 4.99 ( 7.27) | 5.62 (11.26) | - (-) | 7.13 (11.74) | 3.65 ( 6.10) | 3.77 ( 6.87) | - (-) | 6.26 ( 9.51) | 2.84 ( 5.33) | 2.32 ( 5.53) | - (-) | 5.39 ( 8.41) | 2.20 ( 4.81) | 1.81 ( 5.03) | - (-) | 5.29 ( 7.62) | 1.83 ( 4.33) | 1.49 ( 4.82) | - (-) | 5.16 ( 6.65) | 1.54 ( 3.77) | 1.21 ( 4.71) | - (-) | 5.03 ( 6.35) |
| -0.15 - up-down | 4.74 ( 7.88) | 6.87 (12.36) | - (-) | 7.27 (11.73) | 3.61 ( 6.46) | 3.87 ( 8.00) | - (-) | 6.41 ( 9.52) | 2.72 ( 5.76) | 3.23 ( 6.15) | - (-) | 5.56 ( 8.43) | 2.14 ( 5.17) | 1.88 ( 5.45) | - (-) | 5.42 ( 7.63) | 1.78 ( 4.72) | 1.57 ( 5.28) | - (-) | 5.19 ( 6.67) | 1.52 ( 4.16) | 1.31 ( 5.17) | - (-) | 4.97 ( 6.37) |
| -0.2 - right-left | 6.22 ( 5.82) | 6.72 ( 9.46) | - (-) | 5.57 ( 9.76) | 5.38 ( 5.59) | 4.01 ( 6.36) | - (-) | 4.79 ( 7.93) | 5.10 ( 5.76) | 3.69 ( 5.42) | - (-) | 4.17 ( 7.06) | 4.64 ( 5.96) | 3.38 ( 5.01) | - (-) | 4.07 ( 6.42) | 3.99 ( 6.63) | 3.02 ( 4.95) | - (-) | 3.89 ( 5.69) | 3.52 ( 6.71) | 2.32 ( 4.85) | - (-) | 3.59 ( 5.37) |
| -0.2 - up-down | 6.14 ( 6.02) | 9.65 (10.66) | - (-) | 5.74 ( 9.75) | 5.34 ( 5.89) | 5.95 ( 8.35) | - (-) | 5.03 ( 7.92) | 4.98 ( 6.21) | 3.51 ( 5.23) | - (-) | 4.45 ( 7.07) | 4.46 ( 6.49) | 3.25 ( 5.22) | - (-) | 4.34 ( 6.47) | 3.81 ( 7.02) | 2.82 ( 5.08) | - (-) | 4.10 ( 5.73) | 3.41 ( 7.48) | 2.20 ( 4.77) | - (-) | 3.67 ( 5.45) |
| -0.25 - right-left | - (-) | - (-) | - (-) | 4.68 ( 7.81) | - (-) | - (-) | - (-) | 3.61 ( 6.78) | - (-) | - (-) | - (-) | 3.07 ( 6.54) | - (-) | - (-) | - (-) | 2.58 ( 6.38) | - (-) | - (-) | - (-) | 2.21 ( 6.11) | - (-) | - (-) | - (-) | 1.90 ( 5.94) |
| -0.25 - up-down | - (-) | - (-) | - (-) | 4.86 ( 7.47) | - (-) | - (-) | - (-) | 3.98 ( 6.28) | - (-) | - (-) | - (-) | 3.50 ( 6.07) | - (-) | - (-) | - (-) | 3.11 ( 5.93) | - (-) | - (-) | - (-) | 2.69 ( 5.71) | - (-) | - (-) | - (-) | 2.38 ( 5.53) |
| **Upper boundary** | | | | | | | | | | | | | | | | | | | | | | | | |
| -0.05 - right-left | 2.17 (15.67) | 9.11 (18.64) | 18.99 (22.00) | 10.00 (13.85) | 1.43 (13.07) | 4.49 ( 9.92) | 8.10 (11.44) | 8.28 (11.04) | 1.11 (12.24) | 4.20 ( 8.64) | 6.39 ( 8.35) | 6.97 ( 9.70) | 0.92 (12.05) | 3.93 ( 8.30) | 4.81 ( 8.00) | 6.66 ( 8.72) | 0.78 (11.95) | 3.63 ( 8.22) | 4.75 ( 8.02) | 6.30 ( 7.48) | 0.66 (11.89) | 3.66 ( 8.18) | 4.54 ( 8.04) | 6.44 ( 7.07) |
| -0.05 - up-down | 2.07 (15.84) | 7.19 (10.72) | 24.34 (29.83) | 10.09 (14.07) | 1.42 (13.28) | 5.40 ( 8.70) | 12.42 (19.23) | 8.35 (11.27) | 1.15 (12.51) | 5.03 ( 7.88) | 6.76 (16.06) | 7.02 ( 9.93) | 0.98 (12.28) | 4.72 ( 7.55) | 4.83 (11.04) | 6.72 ( 8.94) | 0.87 (12.21) | 4.36 ( 7.45) | 4.71 (10.76) | 6.33 ( 7.71) | 0.79 (12.16) | 4.18 ( 7.41) | 4.50 ( 9.09) | 6.47 ( 7.33) |
| -0.1 - right-left | 5.48 (13.82) | 7.72 (15.42) | 16.43 (25.94) | 9.09 (13.29) | 3.78 (10.76) | 4.11 ( 9.78) | 6.01 (11.37) | 7.59 (10.18) | 2.62 ( 9.38) | 3.44 ( 7.86) | 5.05 ( 9.58) | 6.45 ( 8.80) | 2.09 ( 9.00) | 2.98 ( 7.44) | 4.54 ( 8.76) | 6.17 ( 7.74) | 1.73 ( 8.82) | 2.54 ( 7.28) | 4.25 ( 9.76) | 5.80 ( 6.50) | 1.45 ( 8.67) | 2.44 ( 7.20) | 4.04 (10.08) | 5.79 ( 6.03) |
| -0.1 - up-down | 5.37 (14.82) | 8.53 (14.38) | 44.51 (54.61) | 9.20 (13.51) | 3.70 (11.85) | 4.35 ( 9.59) | 34.37 (45.60) | 7.70 (10.39) | 2.59 (10.56) | 3.74 ( 8.09) | 30.73 (43.73) | 6.52 ( 9.02) | 2.06 (10.22) | 3.27 ( 7.66) | 26.15 (46.33) | 6.25 ( 7.97) | 1.71 (10.06) | 2.80 ( 7.53) | 20.71 (49.61) | 5.87 ( 6.72) | 1.44 ( 9.89) | 2.69 ( 7.42) | 16.44 (53.30) | 5.87 ( 6.32) |
| -0.15 - right-left | 9.97 (12.91) | 10.46 (19.52) | - (-) | 8.09 (13.34) | 7.29 (11.44) | 7.48 (10.92) | - (-) | 6.72 (10.06) | 5.68 (10.17) | 4.52 ( 8.11) | - (-) | 5.63 ( 8.71) | 4.41 ( 9.17) | 3.53 ( 7.18) | - (-) | 5.41 ( 7.64) | 3.65 ( 8.23) | 2.90 ( 6.83) | - (-) | 5.21 ( 6.49) | 3.09 ( 7.11) | 2.30 ( 6.53) | - (-) | 5.18 ( 6.04) |
| -0.15 - up-down | 9.48 (13.69) | 13.05 (21.60) | - (-) | 8.19 (13.57) | 7.21 (11.77) | 7.57 (13.09) | - (-) | 6.81 (10.28) | 5.44 (10.69) | 6.27 ( 8.85) | - (-) | 5.75 ( 8.95) | 4.28 ( 9.55) | 3.46 ( 7.32) | - (-) | 5.52 ( 7.89) | 3.56 ( 8.68) | 2.85 ( 7.06) | - (-) | 5.31 ( 6.74) | 3.03 ( 7.53) | 2.27 ( 6.76) | - (-) | 5.31 ( 6.29) |
| -0.2 - right-left | 12.43 (11.97) | 13.08 (17.35) | - (-) | 7.78 (13.13) | 10.75 (11.24) | 8.00 (12.00) | - (-) | 5.79 (10.04) | 10.21 (11.53) | 7.37 (10.34) | - (-) | 4.61 ( 8.63) | 9.29 (11.94) | 6.77 ( 9.72) | - (-) | 4.19 ( 7.52) | 7.99 (13.26) | 6.04 ( 9.77) | - (-) | 3.83 ( 6.56) | 7.03 (13.42) | 4.65 ( 9.63) | - (-) | 3.57 ( 5.96) |
| -0.2 - up-down | 12.28 (12.26) | 19.05 (19.74) | - (-) | 7.88 (13.36) | 10.69 (11.83) | 11.87 (16.00) | - (-) | 5.90 (10.23) | 9.95 (12.43) | 7.01 ( 9.73) | - (-) | 4.71 ( 8.87) | 8.92 (12.99) | 6.52 (10.04) | - (-) | 4.31 ( 7.77) | 7.62 (14.03) | 5.64 ( 9.96) | - (-) | 3.98 ( 6.82) | 6.83 (14.96) | 4.41 ( 9.43) | - (-) | 3.72 ( 6.26) |
| -0.25 - right-left | - (-) | - (-) | - (-) | 8.77 (12.61) | - (-) | - (-) | - (-) | 6.60 (11.32) | - (-) | - (-) | - (-) | 5.61 (11.19) | - (-) | - (-) | - (-) | 4.31 (10.55) | - (-) | - (-) | - (-) | 3.52 ( 9.90) | - (-) | - (-) | - (-) | 2.75 ( 9.18) |
| -0.25 - up-down | - (-) | - (-) | - (-) | 8.79 (12.51) | - (-) | - (-) | - (-) | 6.71 (10.69) | - (-) | - (-) | - (-) | 5.69 (10.62) | - (-) | - (-) | - (-) | 4.44 ( 9.92) | - (-) | - (-) | - (-) | 3.60 ( 9.28) | - (-) | - (-) | - (-) | 2.89 ( 8.55) |
| **Width** | | | | | | | | | | | | | | | | | | | | | | | | |
| -0.05 - right-left | 2.15 ( 4.62) | 3.99 ( 6.50) | 14.39 (15.67) | 1.77 ( 1.85) | 1.43 ( 2.67) | 1.63 ( 2.35) | 7.24 ( 9.98) | 1.42 ( 1.25) | 1.09 ( 2.26) | 1.33 ( 1.24) | 5.36 ( 7.98) | 1.35 ( 1.14) | 0.92 ( 2.04) | 1.20 ( 1.07) | 2.53 ( 7.15) | 1.23 ( 1.03) | 0.78 ( 1.92) | 1.10 ( 0.96) | 2.14 ( 7.69) | 1.11 ( 0.97) | 0.66 ( 1.77) | 0.99 ( 0.85) | 1.76 ( 7.43) | 0.94 ( 0.97) |
| -0.05 - up-down | 2.04 ( 4.27) | 2.62 ( 2.49) | 24.44 (29.07) | 2.18 ( 1.93) | 1.42 ( 2.33) | 1.58 ( 1.55) | 12.30 (19.95) | 1.88 ( 1.19) | 1.13 ( 1.91) | 1.25 ( 1.09) | 5.79 (16.49) | 1.81 ( 1.00) | 0.98 ( 1.62) | 1.08 ( 0.93) | 2.41 (10.68) | 1.71 ( 0.85) | 0.87 ( 1.51) | 0.93 ( 0.80) | 2.01 (10.38) | 1.58 ( 0.71) | 0.79 ( 1.36) | 0.75 ( 0.67) | 1.63 ( 8.10) | 1.40 ( 0.63) |
| -0.1 - right-left | 5.48 ( 9.09) | 6.30 ( 9.75) | 15.43 (23.53) | 3.68 ( 3.98) | 3.78 ( 6.31) | 3.36 ( 5.61) | 5.81 (11.38) | 2.94 ( 2.49) | 2.62 ( 4.80) | 2.52 ( 2.66) | 4.87 ( 9.81) | 2.76 ( 2.14) | 2.09 ( 4.02) | 2.12 ( 2.24) | 4.49 ( 9.45) | 2.50 ( 1.82) | 1.73 ( 3.62) | 1.83 ( 1.95) | 4.11 (11.16) | 2.26 ( 1.58) | 1.45 ( 3.20) | 1.48 ( 1.65) | 3.78 (11.85) | 1.93 ( 1.49) |
| -0.1 - up-down | 5.37 ( 9.32) | 7.29 (10.92) | 45.15 (54.82) | 3.53 ( 3.93) | 3.70 ( 6.56) | 3.55 ( 5.08) | 35.07 (46.70) | 2.77 ( 2.58) | 2.59 ( 5.13) | 2.74 ( 2.88) | 31.36 (44.79) | 2.53 ( 2.28) | 2.06 ( 4.35) | 2.39 ( 2.50) | 26.69 (47.53) | 2.28 ( 2.07) | 1.71 ( 4.02) | 2.13 ( 2.23) | 21.10 (50.97) | 2.07 ( 1.92) | 1.44 ( 3.60) | 1.87 ( 1.94) | 16.69 (54.80) | 1.76 ( 1.82) |
| -0.15 - right-left | 9.97 (12.34) | 10.15 (18.26) | - (-) | 5.42 ( 6.76) | 7.29 (11.24) | 7.66 ( 9.55) | - (-) | 4.09 ( 4.19) | 5.68 (10.08) | 4.71 ( 6.47) | - (-) | 3.71 ( 3.51) | 4.41 ( 9.04) | 3.81 ( 5.46) | - (-) | 3.28 ( 2.85) | 3.65 ( 8.06) | 3.24 ( 4.94) | - (-) | 2.95 ( 2.38) | 3.09 ( 6.90) | 2.64 ( 4.38) | - (-) | 2.48 ( 2.10) |
| -0.15 - up-down | 9.48 (12.85) | 12.76 (20.38) | - (-) | 5.88 ( 6.94) | 7.21 (11.37) | 7.65 (11.83) | - (-) | 4.61 ( 4.26) | 5.44 (10.42) | 6.32 ( 7.05) | - (-) | 4.32 ( 3.50) | 4.28 ( 9.23) | 3.59 ( 5.19) | - (-) | 3.87 ( 2.82) | 3.56 ( 8.33) | 3.02 ( 4.75) | - (-) | 3.55 ( 2.31) | 3.03 ( 7.12) | 2.44 ( 4.09) | - (-) | 3.05 ( 1.92) |
| -0.2 - right-left | 12.43 (12.44) | 12.88 (16.91) | - (-) | 7.61 (10.83) | 10.75 (11.34) | 8.01 (11.87) | - (-) | 5.37 ( 7.68) | 10.21 (11.55) | 7.38 (10.21) | - (-) | 4.60 ( 6.04) | 9.29 (11.95) | 6.79 ( 9.62) | - (-) | 3.91 ( 4.75) | 7.99 (13.26) | 6.05 ( 9.73) | - (-) | 3.47 ( 3.92) | 7.03 (13.42) | 4.65 ( 9.61) | - (-) | 2.89 ( 3.17) |
| -0.2 - up-down | 12.28 (12.71) | 18.91 (19.36) | - (-) | 8.07 (11.13) | 10.69 (11.92) | 11.88 (15.89) | - (-) | 5.84 ( 7.82) | 9.95 (12.45) | 7.02 ( 9.56) | - (-) | 5.16 ( 6.20) | 8.92 (13.00) | 6.54 ( 9.94) | - (-) | 4.49 ( 4.87) | 7.62 (14.03) | 5.65 ( 9.92) | - (-) | 4.04 ( 4.03) | 6.83 (14.96) | 4.42 ( 9.39) | - (-) | 3.44 ( 3.26) |
| -0.25 - right-left | - (-) | - (-) | - (-) | 9.43 (12.62) | - (-) | - (-) | - (-) | 7.37 (11.13) | - (-) | - (-) | - (-) | 6.56 (10.94) | - (-) | - (-) | - (-) | 5.24 ( 9.75) | - (-) | - (-) | - (-) | 4.50 ( 8.63) | - (-) | - (-) | - (-) | 3.66 ( 7.38) |
| -0.25 - up-down | - (-) | - (-) | - (-) | 9.54 (13.10) | - (-) | - (-) | - (-) | 7.50 (11.00) | - (-) | - (-) | - (-) | 6.66 (10.83) | - (-) | - (-) | - (-) | 5.32 ( 9.48) | - (-) | - (-) | - (-) | 4.58 ( 8.29) | - (-) | - (-) | - (-) | 3.74 ( 7.01) |

RMSE is aggregated over time and treatment categories.

#### 8.2.1.3 Jaccard index for the HRDR intervals

To have a comprehensive viewpoint of the estimation performances of the intervals, we computed the mean Jaccard index across the M montecarlo replications that are displayed in Figure 8.10 and Figure 8.11 for the HRDR and the HNRDR intervals respectively. The higher the Jaccard index, the higher is the similarity of the estimated intervals with the target ones. The piece-wise exponential model definitely performs better than the model on pseudo-observation, although the NPH and Weibull scenarios do not show this discrepancy. Again, as the probability amount integrated over the intervals decreases, the performance of the estimation method decreases as well, except for the exponential scenario. The same is true for the HNRDR that however show overall a slightly decreased performance of the whole estimation, as lower values of the mean jaccard index indicate. Table 8.9 and Table 8.10 report numerically the results visualized in the figures.

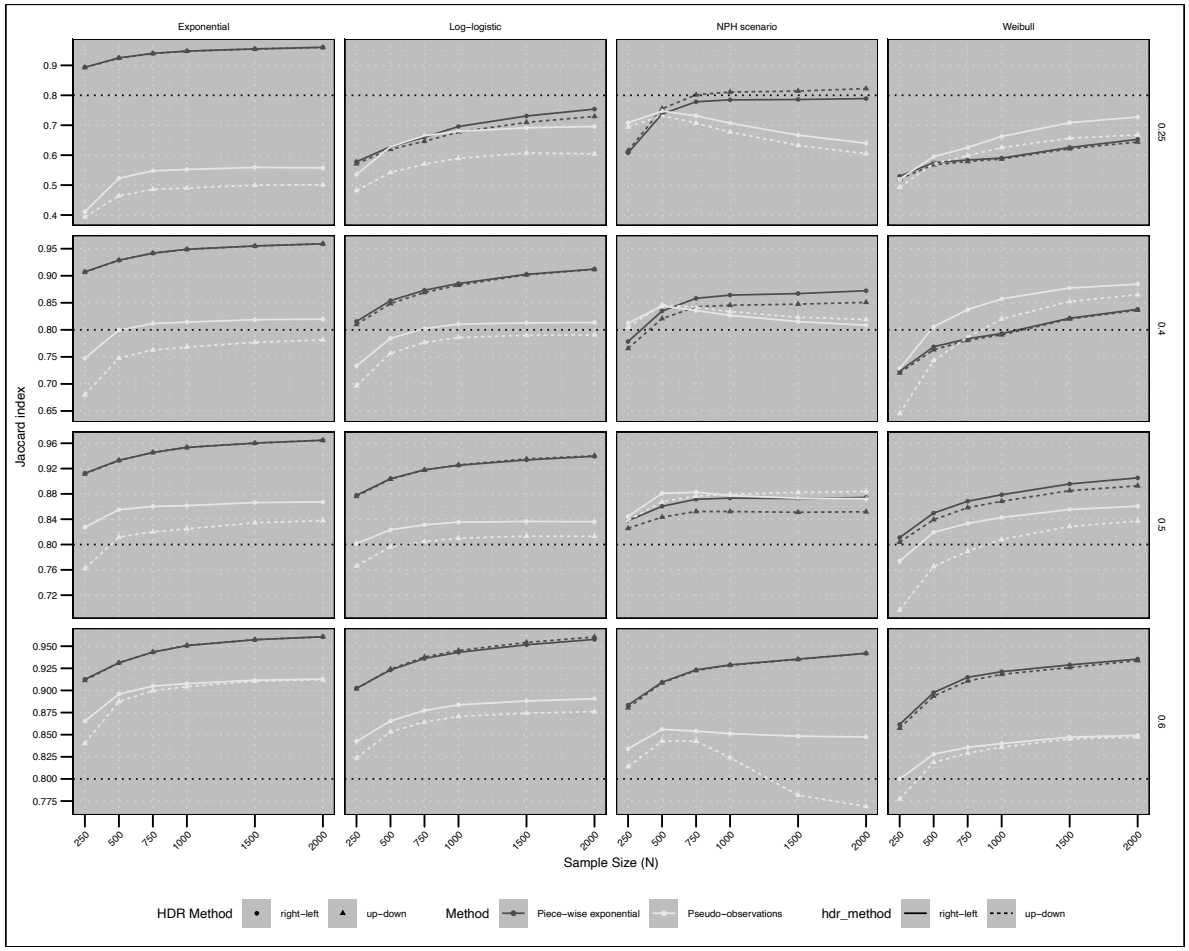



Figure 8.10: Performance of HRDR estimation in terms of Jaccard index computed considering the estimated vs the true intervals as a function of the scenario complexity and the sample size available.

Table 8.9: Performances of estimation of HRDR in terms of Jaccard index. Results are shown as a function of the scenario complexity, sample size available, method of interval construction and flexible model adopted.

| | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N = 250 | N = 250 | N = 250 | N = 250 | N = 500 | N = 500 | N = 500 | N = 500 | N = 750 | N = 750 | N = 750 | N = 750 | N = 1000 | N = 1000 | N = 1000 | N = 1000 | N = 1500 | N = 1500 | N = 1500 | N = 1500 | N = 2000 | N = 2000 | N = 2000 | N = 2000 |
| 0.25 - right-left | 0.89 (0.41) | 0.58 (0.54) | 0.61 (0.71) | 0.53 (0.52) | 0.92 (0.52) | 0.63 (0.63) | 0.74 (0.75) | 0.57 (0.60) | 0.94 (0.55) | 0.66 (0.66) | 0.78 (0.73) | 0.58 (0.63) | 0.95 (0.55) | 0.70 (0.68) | 0.78 (0.71) | 0.59 (0.66) | 0.95 (0.56) | 0.73 (0.69) | 0.79 (0.67) | 0.63 (0.71) | 0.96 (0.56) | 0.75 (0.70) | 0.79 (0.64) | 0.65 (0.73) |
| 0.25 - up-down | 0.89 (0.39) | 0.57 (0.48) | 0.61 (0.69) | 0.52 (0.49) | 0.92 (0.46) | 0.62 (0.54) | 0.76 (0.73) | 0.57 (0.57) | 0.94 (0.49) | 0.65 (0.57) | 0.80 (0.71) | 0.58 (0.60) | 0.95 (0.49) | 0.68 (0.59) | 0.81 (0.68) | 0.59 (0.63) | 0.95 (0.50) | 0.71 (0.61) | 0.81 (0.63) | 0.62 (0.66) | 0.96 (0.50) | 0.73 (0.60) | 0.82 (0.60) | 0.64 (0.67) |
| 0.4 - right-left | 0.91 (0.75) | 0.82 (0.73) | 0.78 (0.81) | 0.72 (0.73) | 0.93 (0.80) | 0.85 (0.78) | 0.83 (0.84) | 0.77 (0.81) | 0.94 (0.81) | 0.87 (0.80) | 0.86 (0.84) | 0.78 (0.84) | 0.95 (0.81) | 0.89 (0.81) | 0.86 (0.83) | 0.79 (0.86) | 0.96 (0.82) | 0.90 (0.81) | 0.87 (0.82) | 0.82 (0.88) | 0.96 (0.82) | 0.91 (0.81) | 0.87 (0.81) | 0.84 (0.88) |
| 0.4 - up-down | 0.91 (0.68) | 0.81 (0.70) | 0.77 (0.80) | 0.72 (0.65) | 0.93 (0.75) | 0.85 (0.76) | 0.82 (0.85) | 0.76 (0.74) | 0.94 (0.76) | 0.87 (0.78) | 0.84 (0.84) | 0.78 (0.79) | 0.95 (0.77) | 0.88 (0.79) | 0.85 (0.83) | 0.79 (0.82) | 0.96 (0.78) | 0.90 (0.79) | 0.85 (0.82) | 0.82 (0.85) | 0.96 (0.78) | 0.91 (0.79) | 0.85 (0.82) | 0.84 (0.87) |
| 0.5 - right-left | 0.91 (0.83) | 0.88 (0.80) | 0.84 (0.84) | 0.81 (0.77) | 0.93 (0.86) | 0.90 (0.82) | 0.86 (0.88) | 0.85 (0.82) | 0.95 (0.86) | 0.92 (0.83) | 0.87 (0.88) | 0.87 (0.83) | 0.95 (0.86) | 0.93 (0.84) | 0.87 (0.88) | 0.88 (0.84) | 0.96 (0.87) | 0.93 (0.84) | 0.87 (0.87) | 0.90 (0.86) | 0.96 (0.87) | 0.94 (0.84) | 0.87 (0.87) | 0.91 (0.86) |
| 0.5 - up-down | 0.91 (0.76) | 0.88 (0.77) | 0.83 (0.84) | 0.80 (0.70) | 0.93 (0.81) | 0.90 (0.80) | 0.84 (0.87) | 0.84 (0.77) | 0.95 (0.82) | 0.92 (0.80) | 0.85 (0.88) | 0.86 (0.79) | 0.95 (0.82) | 0.93 (0.81) | 0.85 (0.88) | 0.87 (0.81) | 0.96 (0.83) | 0.94 (0.81) | 0.85 (0.88) | 0.89 (0.83) | 0.96 (0.84) | 0.94 (0.81) | 0.85 (0.88) | 0.89 (0.84) |
| 0.6 - right-left | 0.91 (0.87) | 0.90 (0.84) | 0.88 (0.83) | 0.86 (0.80) | 0.93 (0.90) | 0.92 (0.87) | 0.91 (0.86) | 0.90 (0.83) | 0.94 (0.90) | 0.94 (0.88) | 0.92 (0.85) | 0.91 (0.84) | 0.95 (0.91) | 0.94 (0.88) | 0.93 (0.85) | 0.92 (0.84) | 0.96 (0.91) | 0.95 (0.89) | 0.94 (0.85) | 0.93 (0.85) | 0.96 (0.91) | 0.96 (0.89) | 0.94 (0.85) | 0.94 (0.85) |
| 0.6 - up-down | 0.91 (0.84) | 0.90 (0.82) | 0.88 (0.81) | 0.86 (0.78) | 0.93 (0.89) | 0.92 (0.85) | 0.91 (0.84) | 0.89 (0.82) | 0.94 (0.90) | 0.94 (0.86) | 0.92 (0.84) | 0.91 (0.83) | 0.95 (0.90) | 0.95 (0.87) | 0.93 (0.82) | 0.92 (0.84) | 0.96 (0.91) | 0.95 (0.87) | 0.94 (0.78) | 0.93 (0.85) | 0.96 (0.91) | 0.96 (0.88) | 0.94 (0.77) | 0.93 (0.85) |

Mean Jaccard is aggregated over time and treatment categories.

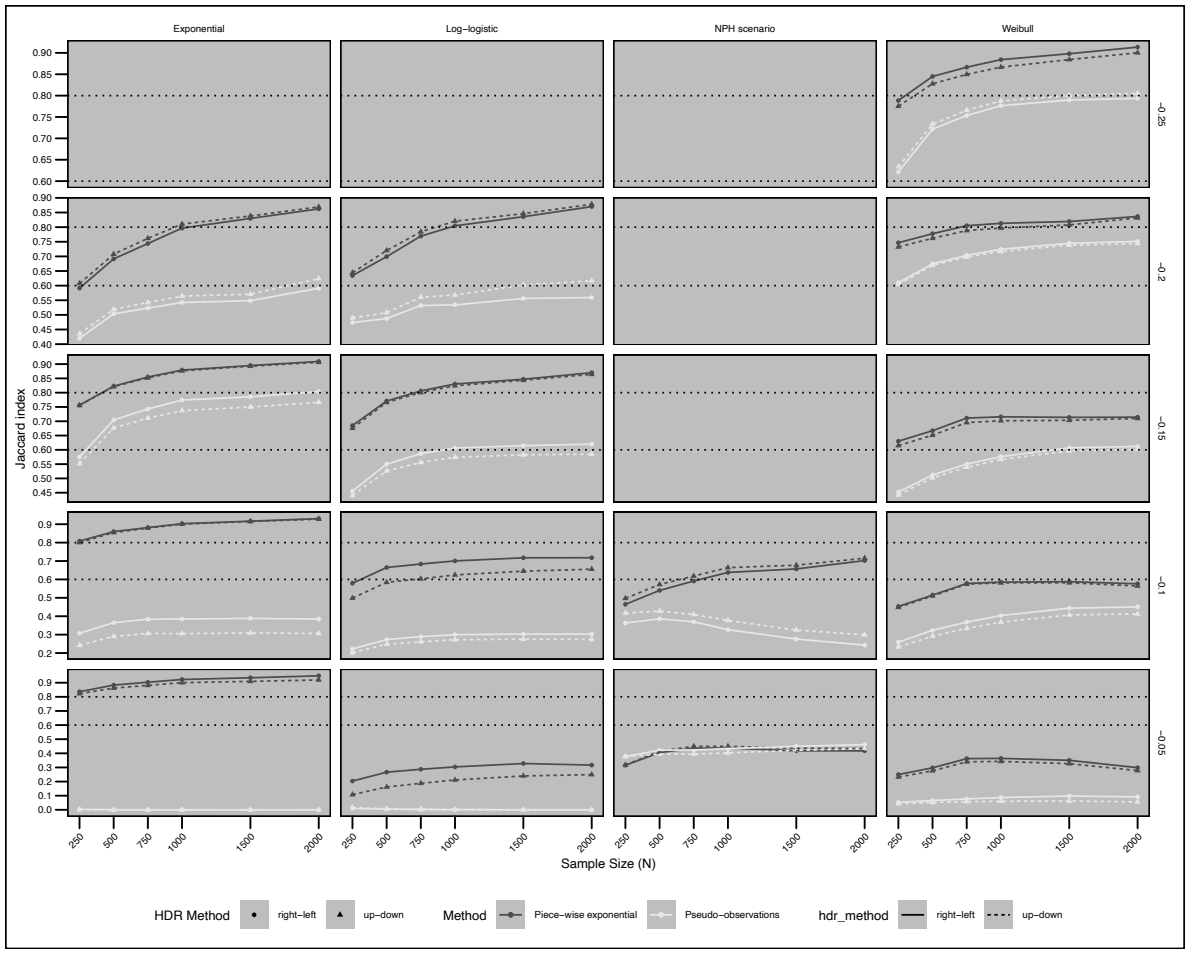



Figure 8.11: Performance of HNRDR estimation in terms of Jaccard index computed considering the estimated vs the true intervals as a function of the scenario complexity and the sample size available.

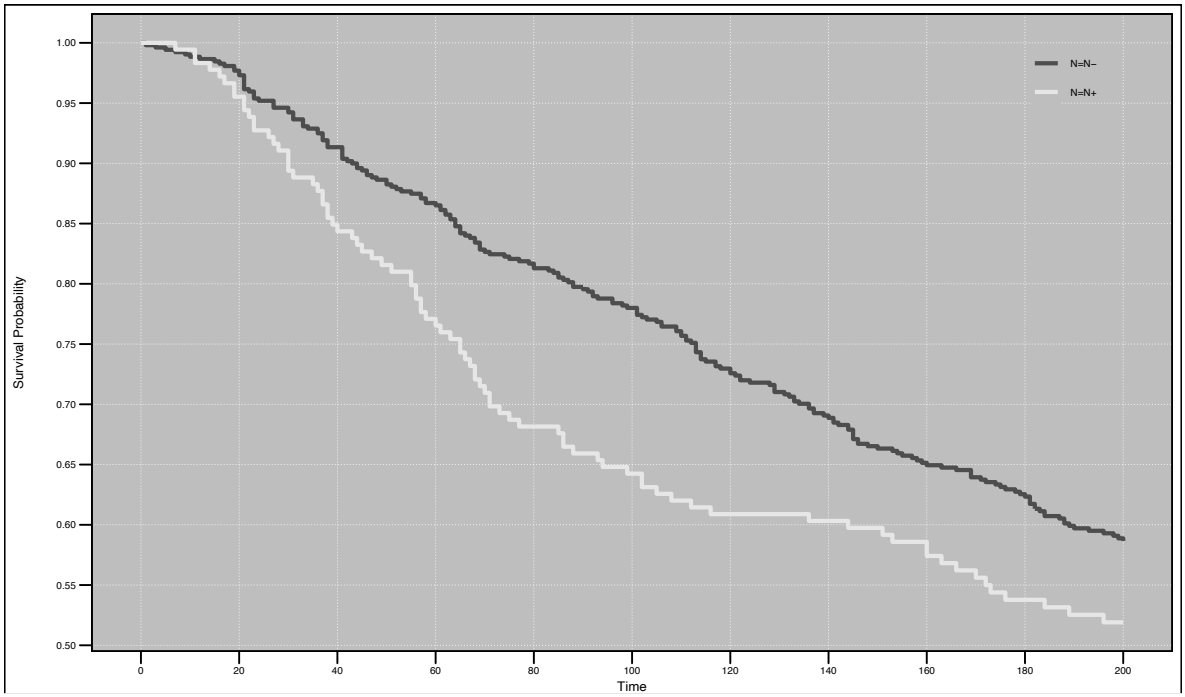
Table 8.10: Performances of estimation of HNRDR in terms of Jaccard index. Results are shown as a function of the scenario complexity, sample size available, method of interval construction and flexible model adopted.

| | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull | Exponential | Log-logistic | NPH scenario | Weibull |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N = 250 | N = 250 | N = 250 | N = 250 | N = 500 | N = 500 | N = 500 | N = 500 | N = 750 | N = 750 | N = 750 | N = 750 | N = 1000 | N = 1000 | N = 1000 | N = 1000 | N = 1500 | N = 1500 | N = 1500 | N = 1500 | N = 2000 | N = 2000 | N = 2000 | N = 2000 |
| -0.05 - right-left | 0.84 (0.00) | 0.20 (0.01) | 0.32 (0.38) | 0.25 (0.05) | 0.88 (0.00) | 0.27 (0.00) | 0.40 (0.42) | 0.30 (0.07) | 0.90 (0.00) | 0.29 (0.00) | 0.44 (0.42) | 0.36 (0.08) | 0.92 (0.00) | 0.30 (0.00) | 0.44 (0.43) | 0.36 (0.09) | 0.94 (0.00) | 0.33 (0.00) | 0.42 (0.45) | 0.35 (0.10) | 0.95 (0.00) | 0.32 (0.00) | 0.42 (0.46) | 0.30 (0.09) |
| -0.05 - up-down | 0.82 (0.00) | 0.11 (0.02) | 0.32 (0.38) | 0.23 (0.04) | 0.86 (0.00) | 0.16 (0.01) | 0.41 (0.39) | 0.28 (0.05) | 0.88 (0.00) | 0.19 (0.00) | 0.45 (0.39) | 0.34 (0.06) | 0.90 (0.00) | 0.21 (0.00) | 0.45 (0.40) | 0.34 (0.06) | 0.91 (0.00) | 0.24 (0.00) | 0.44 (0.42) | 0.33 (0.06) | 0.92 (0.00) | 0.25 (0.00) | 0.43 (0.43) | 0.28 (0.06) |
| -0.1 - right-left | 0.81 (0.31) | 0.58 (0.22) | 0.46 (0.36) | 0.45 (0.26) | 0.86 (0.36) | 0.66 (0.27) | 0.54 (0.39) | 0.51 (0.32) | 0.88 (0.38) | 0.68 (0.29) | 0.59 (0.37) | 0.58 (0.37) | 0.90 (0.38) | 0.70 (0.30) | 0.64 (0.33) | 0.59 (0.40) | 0.92 (0.39) | 0.72 (0.30) | 0.66 (0.28) | 0.59 (0.44) | 0.93 (0.38) | 0.72 (0.30) | 0.70 (0.24) | 0.58 (0.45) |
| -0.1 - up-down | 0.80 (0.24) | 0.50 (0.20) | 0.50 (0.42) | 0.45 (0.23) | 0.85 (0.29) | 0.58 (0.25) | 0.57 (0.43) | 0.51 (0.29) | 0.88 (0.31) | 0.60 (0.26) | 0.62 (0.41) | 0.57 (0.33) | 0.90 (0.31) | 0.62 (0.27) | 0.66 (0.38) | 0.58 (0.37) | 0.91 (0.31) | 0.64 (0.27) | 0.68 (0.32) | 0.58 (0.41) | 0.93 (0.31) | 0.66 (0.27) | 0.72 (0.30) | 0.56 (0.41) |
| -0.15 - right-left | 0.76 (0.57) | 0.68 (0.46) | - (-) | 0.63 (0.45) | 0.82 (0.70) | 0.77 (0.55) | - (-) | 0.67 (0.51) | 0.85 (0.74) | 0.81 (0.59) | - (-) | 0.71 (0.55) | 0.88 (0.77) | 0.83 (0.61) | - (-) | 0.72 (0.58) | 0.89 (0.79) | 0.85 (0.61) | - (-) | 0.71 (0.61) | 0.91 (0.80) | 0.87 (0.62) | - (-) | 0.71 (0.61) |
| -0.15 - up-down | 0.76 (0.55) | 0.68 (0.44) | - (-) | 0.61 (0.44) | 0.82 (0.68) | 0.77 (0.53) | - (-) | 0.65 (0.50) | 0.85 (0.71) | 0.80 (0.56) | - (-) | 0.70 (0.54) | 0.88 (0.74) | 0.82 (0.57) | - (-) | 0.70 (0.57) | 0.89 (0.75) | 0.84 (0.58) | - (-) | 0.70 (0.60) | 0.91 (0.77) | 0.86 (0.58) | - (-) | 0.71 (0.60) |
| -0.2 - right-left | 0.59 (0.42) | 0.63 (0.47) | - (-) | 0.75 (0.61) | 0.69 (0.50) | 0.70 (0.49) | - (-) | 0.78 (0.67) | 0.74 (0.52) | 0.77 (0.53) | - (-) | 0.81 (0.70) | 0.80 (0.54) | 0.80 (0.53) | - (-) | 0.81 (0.72) | 0.83 (0.55) | 0.84 (0.56) | - (-) | 0.82 (0.74) | 0.86 (0.59) | 0.87 (0.56) | - (-) | 0.84 (0.75) |
| -0.2 - up-down | 0.61 (0.44) | 0.64 (0.49) | - (-) | 0.73 (0.60) | 0.71 (0.52) | 0.72 (0.51) | - (-) | 0.76 (0.67) | 0.76 (0.54) | 0.78 (0.56) | - (-) | 0.79 (0.70) | 0.81 (0.56) | 0.82 (0.57) | - (-) | 0.80 (0.72) | 0.84 (0.57) | 0.85 (0.60) | - (-) | 0.81 (0.74) | 0.87 (0.62) | 0.88 (0.62) | - (-) | 0.83 (0.74) |
| -0.25 - right-left | - (-) | - (-) | - (-) | 0.79 (0.62) | - (-) | - (-) | - (-) | 0.84 (0.72) | - (-) | - (-) | - (-) | 0.87 (0.75) | - (-) | - (-) | - (-) | 0.88 (0.78) | - (-) | - (-) | - (-) | 0.90 (0.79) | - (-) | - (-) | - (-) | 0.91 (0.79) |
| -0.25 - up-down | - (-) | - (-) | - (-) | 0.78 (0.63) | - (-) | - (-) | - (-) | 0.83 (0.73) | - (-) | - (-) | - (-) | 0.85 (0.77) | - (-) | - (-) | - (-) | 0.87 (0.79) | - (-) | - (-) | - (-) | 0.88 (0.80) | - (-) | - (-) | - (-) | 0.90 (0.81) |

Mean Jaccard is aggregated over treatment categories.

### 8.2.2 Application to Milan 1 trial data

The survival functions in Figure 8.12 describe the cumulative probability of being free from any event for the patients involved in Milan 1 trial study, conditional on their lymph-nodal status. Here, disease free survival is a composite outcome that consists of distant recurrences, other cancers not related to the disease and death without evidence of recurrence. Node-positive women are clearly characterized by an increased risk of any event, mainly driven by distant recurrences, as the two survival functions are well vertically separated and the N+ one is below the N-.




| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| N=N– | | | | | | | | | | | |
| At Risk | 520 | 508 | 474 | 449 | 422 | 403 | 376 | 353 | 329 | 310 | 275 |
| Events | 0 | 14 | 45 | 70 | 97 | 114 | 142 | 161 | 181 | 194 | 212 |
| N=N+ | | | | | | | | | | | |
| At Risk | 180 | 171 | 152 | 138 | 122 | 115 | 108 | 107 | 100 | 87 | NA |
| Events | 0 | 8 | 28 | 42 | 57 | 64 | 70 | 71 | 76 | 82 | 85 |

Figure 8.12: Survival functions estimated with the Kaplan Meier method for Node-positive and negative patients.

Our purpose is to describe what is the effect of nodal status on the distribution of the disease-free survival time in terms of the time-scale. To such end we will compute the 10%, 20% and 30% Highest Risk Density Regions and the 5% and 10% Highest Net Risk Difference Region, considering estimates of the PDF and CDF obtained with a flexible Piece-wise exponential model. The model accounts for the surgical operation adopted, the menopausal status and the tumor dimension at the time of resection as possible confounding factor. From the survival function plot, we hypothesized the relaxation to non-proportional hazards for the lymph

Table 8.11: HRDRs and HNRDRs for Milan 1 trial data

**HRDRs and HNRDRs for Milan 1 trial data**

| Nodal Status | method | HRDR_0.1 | HRDR_0.2 | HRDR_0.3 | HNRDR_0.05 | HNRDR_0.1 |
|---|---|---|---|---|---|---|
| N+ | up-down | [0, 26] | [0, 53] | [0, 84] | | |
| N- | up-down | [51, 93] | [28, 115] | [2, 137] | | |
| N+ | right-left | [0, 26] | [0, 53] | [0, 83] | | |
| N- | right-left | [50, 93] | [27, 115] | [1, 137] | | |
| | up-down | | | | [0, 28] | [0, 65] |
| | right-left | | | | [0, 29] | [0, 66] |

nodal status of the patients. The baseline hazard function was smoothed with natural splines specifying 2 knots at quantiles of the event times, whereas the time-varying effect just with one knot at their median. The AIC selected the model specifying the interaction between the baseline hazard and the covariate. This is in line with results obtained from a previous analysis focused on the modeling of the crude cumulative incidence of distant recurrences. We adopted the g-formula to obtain the marginal PDFs and CDFs for N+ and N- patients that are displayed in Figure 8.13 (a) and (b) respectively.The HRDRs constructed with the up-down approach are shown in (a) whereas the ones with right-left approach in (b). The risk profile of N+ women highly differ from the N- women, as they are characterized by the highest risk density of events at early follow-up times, whereas N- women are characterized by a delayed risk for any event. Namely, 10% of the total events in N+ women is mostly expected to be in a time-window of between the start of the follow-up and 26 months ([0,26] months), whereas in N- women in a time-window of [51, 93] months. This means that being N- delays the period of the highest risk by 51 months. Other than the effect on the location of risk, we can also focus on the effect on the scale, i.e, the rate of accumulation of the events. N- women show a wider interval compared to N+ women (interval width 26 vs 42). The remaining HRDRs show again this difference, as the 20% HRDRs for N+ and N- patients are [0,53] and [28, 115] respectively and the 30% HRDRs are [0,84] and [2, 137] respectively. The effect on the location on the intervals is visible also for the 20% HRDRs whereas the effect on the scale is well characterized by the intervals widths that display a constant ratio of ~1.6, indicating that the being N- slows down the dynamic of event or increases the disease free survival time by ~60%. To characterize better the impact of nodal status on the redistribution of risk, we computed the 5% and 10% HNRDRs that are displayed in Figure 9 (c) and (d) below the instantaneous risk difference and the cumulative risk difference function respectively. The cumulative risk difference function (d) displays that N+ patients accumulate a risk for the event higher by ~12% than the risk of N- women. Difficult to understand is that this risk-increase is evident at early times of the follow-up. Indeed, the instantaneous risk difference function display that the N+ patients show a higher risk than N- patients since the start of the follow-up, reflecting what we have seen with the HRDRs above. From the 5% and 10% HNRDRs we understand that in a time period of [0,28] N+ women show a higher risk of 5% (i.e, 20 number needed to harm) and of 10% in a time-period of [0, 65] months.

Table 8.11 summarizes the relevant HRD and HNRD regions for N+ and N- patients.

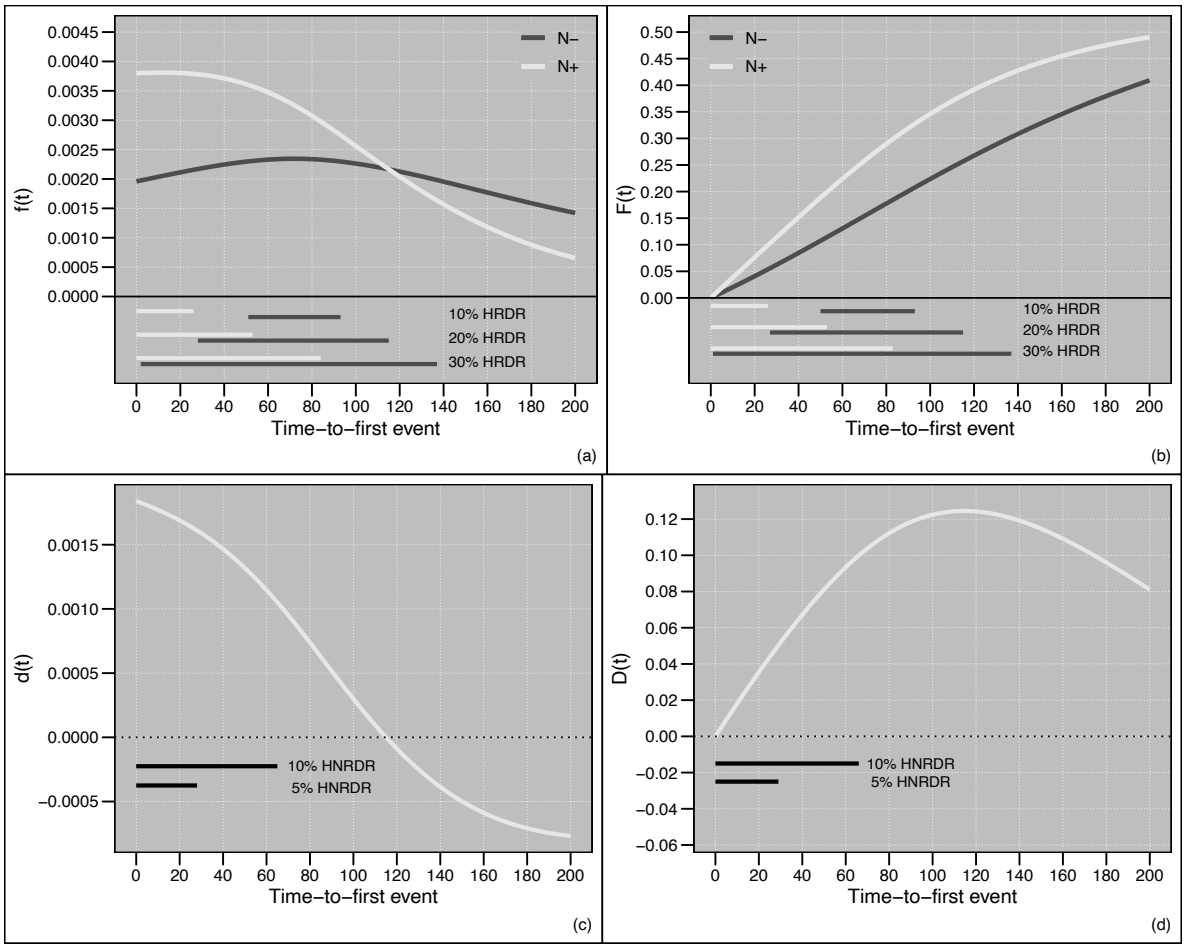



Figure 8.13: PDF and CDF estimated with the flexible PWE models for time-to-first event (a,b) for N+ and N- patients. Below are the 10%, 20% and 30% HRDR intervals computed with the Up-down approach and the Right-left approach respectively. In (c) and (d) the instantaneous risk difference and the cumulative risk difference functions along with the 5% and 10% HNRDR intervals computed with the Up-down approach and the Right-left approach respectively

## 8.3 Discussion

This paper introduces a novel framework for time-to-event (TTE) analysis to address the limitations of conventional metrics. Standard measures like the hazard ratio (HR), restricted mean survival time (RMST), and accelerated failure time (AFT) models are often constrained by restrictive assumptions, sensitivity to user-specified parameters, or challenges in causal interpretation. We propose two new estimands: the Highest Risk Density Region (HRDR), defined as the narrowest time interval containing a pre-specified probability of events, and the Highest Net Risk Difference Region (HNRDR), the narrowest interval in which a pre-specified net absolute risk difference is achieved. Together, these estimands shift the analytical focus from a single summary statistic to characterizing the temporal location, shape, and dynamics of the underlying risk distribution, thereby answering the direct clinical questions of "when is risk highest?" and "when is the treatment effect most pronounced?".

The HRDR and HNRDR are grounded in a formal causal inference framework, defined as features of potential outcome distributions and standardized via the g-formula to permit adjustment for confounding. A key strength of this approach is its communicability; by expressing results in natural units of time, it translates complex statistical findings into actionable clinical insights. For example, in the Milan I trial data, the HRDRs revealed distinct "temporal risk fingerprints" for node-positive patients (high, concentrated, early risk) versus node-negative patients (lower, dispersed, persistent risk), directly informing tailored surveillance strategies. Furthermore, the framework can disentangle a treatment's mechanism, separating effects on the timing (location) from the duration (scale) of risk. This capacity was demonstrated in simulations where the HRDRs clearly distinguished between treatments that delayed the onset of risk versus those that dispersed risk over a longer period.

This work fits into a broader movement within biostatistics to develop more patient-centric and clinically relevant effect measures. The HNRDR, in particular, can be viewed as a tool for "benefit accounting." A core principle of TTE analysis for non-curative therapies is that risk is not eliminated but redistributed over time; the integral of the difference in PDFs over an infinite horizon is necessarily zero. The HNRDR for a net benefit identifies the specific time-window where events were most intensely *prevented* or *postponed*. This implicitly acknowledges that those postponed events may occur later, providing a more complete and honest narrative of a treatment's effect. This concept aligns closely with other modern patient-oriented measures like "The Net Chance of a Longer Survival," which also seeks to quantify benefit in a more direct and probabilistic manner (Péron et al. 2016).

As summarized in the following Table, the HRDR/HNRDR framework occupies a unique and valuable niche among these measures by focusing specifically on the temporal location of risk and benefit. While the Win Ratio (Wang et al. 2011) prioritizes a hierarchy of different outcomes (e.g., death over hospitalization) and RMST quantifies the average time gained, the HRDR/HNRDR specifies *when* the risk is highest and *when* that time is gained most efficiently, providing complementary, not competing, information.

A critical advantage of our framework is its inherent robustness to non-proportional hazards (NPH), a common scenario where the Cox model is unreliable. By operating directly on unconditional probability and cumulative density functions, the HRDR and HNRDR remain

interpretable regardless of the underlying hazard behavior. Monte Carlo simulations provided practical guidance for implementation, revealing that a Piecewise Exponential (PWE) model is superior for estimating the instantaneous risk density (PDF) and should be paired with a PDF-based algorithm for constructing the regions. Conversely, a Pseudo-observation approach, which directly models the cumulative distribution (CDF), should be paired with a CDF-based algorithm to avoid the numerical instability of differentiation. While marginal estimates can be challenging in complex NPH settings, the estimation of the difference between groups for the HNRDR remains remarkably robust.

The primary limitations of this framework are its dependence on the adequacy of the underlying flexible parametric model and the subjectivity in pre-specifying the probability mass or risk difference for the regions; we recommend sensitivity analyses across a range of values to ensure transparency. Despite these considerations, the framework provides a robust and intuitive tool for TTE analysis. Future work will focus on extending these methods to competing risks settings and to accommodate time-varying exposures and dynamic treatment regimes.

**A Comparative Overview of Time-to-Event Summary Measures**

| Feature | Hazard Ratio (HR) | Restricted Mean Survival Time (RMST) | Accelerated Failure Time (AFT) Models | Win Ratio | The Net Chance of a Longer Survival | Proposed HRDR / HNRDR |
|---|---|---|---|---|---|---|
| **Primary Interpretation** | Relative instantaneous event rate | Absolute difference in mean event-free time up to horizon $\tau$ | Multiplicative factor on time scale (time ratio) | Prioritized comparison of composite outcomes | Net probability of a patient in one group surviving longer than a patient in another | Absolute time interval of highest risk concentration or maximal treatment effect |
| **Key Assumption(s)** | Proportional Hazards (PH) | None (non-parametric) | Constant time ratio; specified error distribution | Pre-specified clinical hierarchy of outcomes | None (non-parametric) | None (relies on quality of PDF/CDF estimate) |
| **Robustness to NPH** | Poor; yields a difficult-to-interpret average | High | Moderate; time ratio assumption may be violated | High (by design) | High | High |
| **Primary Clinical Question Answered** | What is the relative rate of events at any given time? | How much average event-free time is gained/lost over a specific period? | By what factor does treatment acceler-ate/decelerate the time to event? | Does the treatment group have a better outcome profile, considering a hierarchy of clinical importance? | What are the net odds that a treated patient will live longer than a control patient? | When is the risk highest? When is the treatment effect most concentrated? |
| **Key Limitation** | PH assumption often violated; poor causal interpretability; not intuitive | Sensitive to choice of time horizon $\tau$; can mask dynamic effects | Risk of model misspecification from incorrect distributional choice | Can be driven by a single component in the hierarchy; ties do not contribute to the ratio | Requires pairwise comparisons; interpretation of magnitude can be complex | Relies on accurate density estimation; requires user to specify probability mass or net difference |

## 8.4 Conclusion

In conclusion, this paper has introduced the Highest Risk Density Region (HRDR) and the Highest Net Risk Difference Region (HNRDR) as novel, intuitive, and robust tools for the analysis of time-to-event data. By shifting the analytical focus from single-number summaries of effect magnitude to a direct characterization of the temporal distribution of risk, this framework provides a richer and more clinically relevant understanding of how treatments and exposures influence patient outcomes over time. The extensive simulation study not only validates the methods across a range of scenarios, from simple proportional hazards to

complex non-proportional hazards, but also provides clear, practical guidance on their optimal implementation.

The HRDR/HNRDR framework should not be viewed as a replacement for all existing time-to-event measures, but rather as a powerful and essential complement. It fills a critical niche by providing clear answers to the fundamental clinical questions of "when is risk highest?" and "when is a treatment's benefit most concentrated?". In doing so, it serves as a vital bridge between the statistical sophistication of flexible parametric survival models and the practical need for clear, causal, and communicable evidence. This work contributes to a broader, ongoing paradigm shift in biostatistics and epidemiology—a movement towards a more holistic, descriptive, and interpretable approach to quantifying treatment effects. By empowering researchers and clinicians to better visualize, understand, and communicate the temporal dynamics of risk, this framework represents a meaningful advancement in the field of survival analysis.

# 9 Describing dose-response relationships with probability density and cumulative distribution functions: an approach based on the generalized linear model framework

Building on the limitations of traditional mean-based regression discussed in Section 3.10, this chapter exploits the use of the flexible-parametric methods adopted in this thesis for estimating the full conditional probability density function of continuous, non-censored, health outcomes in dose-response analysis. This method operates within the generalized linear (additive) model framework, offering both flexibility and interpretability.

We obtain smooth estimates of the density and the cumulative distribution functions with only a few parameters considering splines. This enables the derivation of quantile functions for the outcome distribution, allowing for direct percentile predictions based on exposure variables within a single unified model.

Additionally, exploiting the 'chain rule' of probability, we estimate the joint PDF and CDF of two outcome variables conditioned on exposure, facilitating the evaluation of exposure effects on two variables equivalently relevant to describe an health outcome, like for example plasma glucose and blood pressure for metabolic heath. We emphasize clear and effective visualization method to enhance the interpretability of the results, adopting highest density regions (HDR) and upper bound equal tailed intervals (ETI) to easily communicate the impact of an exposure covariate on the distribution of an outcome.

This chapter is structured as follows. First off, we formalize the adoption of the flexible models on the hazard function and on the cumulative distribution function to the analysis of non-censored outcomes. Then, we describe the two-stage modeling process along with the chain rule of probability to obtain smoothed estimates of joint probability density functions of bivariate (multivariate) outcomes. Then, we summarize possible methods to translate PDF and CDF into interpretable statements about probability mass shifts, given the exposure. We describe the simulation adopted to demonstrate the visualization techniques and to assess the performances of the methods. Finally, a small application on the publicly available PIMA Indians Study dataset (Venables and Ripley 2002) is presented.

## 9.1 Methods

### 9.1.1 The relationship between the hazard, the cumulative distribution function and the probability density function

The survival function is proper of any random variable. In time-to-event analysis, it has a straightforward interpretation –having a time to event $T$ greater than $t$ it means surviving at that time-point-. Also in non-censored data contexts, a survival function of a random variable such as plasma glucose $G$ describes the proportion of subjects in a distribution that have a value of plasma glucose greater than a certain value $g$.

The same is true for the cumulative distribution function, often visualized along with box-plots or histograms, since it simply describes the probability of having a value of glucose $G$ below or equal a certain value $g$.

Equivalently, the hazard function is proper of any random variable. However, its direct interpretation in contexts different from the time-to-event is particularly difficult. The hazard function of plasma glucose would describe the probability of having a value of $G$ in an 'infinitesimal interval' given that it is not of a lower value. However, the hazard function is mathematically valid and it is convenient for the analysis of any, possibly non-censored, outcome.

Following the logic of Chapter 8, we aim for the best smoothed estimates of $h(y|X)$ or $F(y|X)$ to then compute a PDF $f(y|X)$ considering the fundamental relationships displayed in Section 4.1 . A great advantage of non-censored data is the fact that it is not necessary to to restrict the support of the outcome variable.

As in the paper before, we are using models of the form:

$$g[Q|X)] = \beta X + \sum_{k=1}^{K} (\gamma_k + \delta_k X) B_k(Q) \tag{9.1}$$

where, $Q$ is either the hazard or the CDF of the random variable $Y$; $\beta X$ is the coefficient associated to the exposure covariate $X$, imposing a constant shift on $g(Q)$; $B_k(Q)$ is the spline bases matrix to specify the function $g(Q)$ for a subject at reference values of the covariates; $\gamma_k$ are the parameters associated to the spline bases, whereas $\delta_k X$ are the possible interaction terms which relax the constant shift defined by $\beta X$.

### 9.1.2 Joint probability density and joint cumulative distribution function

To model the joint probability density function, a two step modeling procedure and the chain rule of probability are adopted.

Suppose $Y_1$ and $Y_2$ are outcome variables whose joint distribution is of interest. We want to obtain estimates of $f(Y_1, Y_2|X)$. We will use the relation: $f(Y_1, Y_2|X) = f(Y_1|Y_2, X) \times f(Y_2|X)$.

To such end, we fit separate models for $f(Y_2|X)$ and for $f(Y_1|Y_2, X)$ to aim for their best smoothed estimates. For example, adopting a flexible PWE model on the hazard, the first model is

$$\log[h(y_2|X)] = \beta X + \sum_{k=1}^{K}(\gamma_k + \delta_k X)B_k(y_2) \tag{9.2}$$

and the second model is

$$\log[h(y_1|\mathbf{X})] = \mathbf{X}\beta + \sum_{k=1}^{K}(\gamma_k + \mathbf{X}\delta_k)B_k(y_1) \tag{9.3}$$

where $\mathbf{X}$ is a matrix containing values of $X$ and $Y_2$.

### 9.1.3 Translating model results into interpretable statements about probability mass shifts

For time-to-event data, it is often acceptable to express associations exponentiating the model coefficients to obtain hazard ratios, risk ratios, or risk differences, although, as previously discussed, this method comes with significant interpretative challenges. In contrast, applying this approach to non-censored outcomes by presenting an estimated time-varying hazard ratio or a ratio of cumulative distribution functions is entirely inappropriate.

However, building from the logic displayed in Chapter 8, we are able to translate the qualitatively sound estimated PDFs and the CDFs into quantitatively sound statements of probability mass shifts of the distribution conditional on covariates. In particular, other than the Highest Density Regions, we can obtain the Upper Bound Equal Tailed Intervals, that merely correspond to the quantiles of the conditional distributions. In particular, the upper bound ETI correspond to the percentiles of the conditional distribution, similar to the conditional quantile function plots of quantile regression (Koenker and Hallock 2001).

The method of HDRs identifies the smallest region of the outcome space where a given proportion (e.g., ) of the $(1-\alpha)\%$ population is expected to lie (e.g., where do most people's outcomes cluster?). For a fixed probability level (e.g., 95%), the HDR shows how the most probable range of outcomes evolves with exposure. This highlights where the distribution is most concentrated (e.g., narrowing/widening of high-probability intervals).

The plot of the conditional quantile functions focuses on cumulative probabilities, showing the outcome value below which a given proportion (e.g., $(1-\alpha)\%$) of the population falls. Unlike HDRs, percentile plots emphasize thresholds (e.g., median, 90th percentile) and how these cutoffs shift with exposure. This is particularly useful for understanding extremes (e.g., tail behavior) or policy-relevant thresholds (e.g., what exposure level ensures $(1-\alpha)\%$ of outcomes remain below a critical value?).

### 9.1.4 Simulated scenario

Considered the large simulation study conducted in Chapter 8, we performed a smaller Monte Carlo simulation study to assess the properties of our method under different conditions than

before, which involved continuous exposure covariates -not just a binary 'treatment'- and the additional estimation of a joint distribution.

The data-generating mechanism was based on one continuous exposure covariate (Physical Activity measured in metabolic equivalents, `METs`) and two distinct outcomes (plasma glucose and plasma triglycerides, `glu` and `tgy`). First, the exposure covariate, $X$, was randomly sampled for all observations from a log-normal distribution, specifically $X \sim \text{Lognormal}(\mu_{\log(X)} = 2, \sigma_{\log(X)} = 0.4)$.

Then `glu` and `tgy` were generated to be conditionally dependent on $X$. The relationship for both outcomes was defined by a linear model on the natural logarithmic scale. For each observation $i$ and for each outcome $k \in \{1, 2\}$, the mean of the log-transformed outcome was determined by the equation:

$$\mu_{\log(Y_{ik})} = \beta_{0k} + \beta_{1k}(X_i - 5) \tag{9.4}$$

Here, $\beta_{0k}$ represents the baseline mean of $\log(Y_k)$ when the covariate $X$ is equal to 5, and $\beta_{1k}$ is the coefficient quantifying the effect of $X$ on $\log(Y_k)$. This structure allows each outcome to have its own unique baseline and effect size. Finally, the outcome values were drawn from log-normal distributions using these dynamically calculated means and a specific residual standard deviation for each outcome,

$$\text{Y}_{ik} \sim \text{Lognormal}(\mu_{\log(Y_{ik})}, \sigma_{\log(Y_k)}) \tag{9.5}$$

In particular:

- $\text{METs} \sim \text{Log-Normal}(\mu = 2, \sigma = 0.4)$
- $\text{glu} \mid \text{METs} \sim \text{Log-Normal}(\mu = \beta_0 - 0.04 \cdot \text{METs}, \sigma = 0.2)$
- $\text{tgy} \mid \text{METs} \sim \text{Log-Normal}(\mu = \beta_0 - 0.1 \cdot \text{METs}, \sigma = 0.3)$

In the first section of the results, we show the results of a very large simulated dataset ($N = 10000$) to demonstrate the methods and to highlight the closeness of the simulation to a real world context. For sake of simplicity we show just the results from the piece-wise exponential model, targeting the hazard function of the outcomes `glu` and `tgy`.

We measure the performances of the method with the already adopted Jaccard indexes for the conditional regions estimated and compared with the true ones. In particular, for the univariate case, we will evaluate the average $J(\text{est}, \text{true}) = \frac{|\text{est} \cap \text{true}|}{|\text{est} \cup \text{true}|}$, for the multivariate case the average $J(\text{est}, \text{true}) = \frac{\text{Area}(\text{est} \cap \text{true})}{\text{Area}(\text{est} \cup \text{true})}$.

### 9.1.5 Pima Indian Study

The Pima Indian Study investigates diabetes risk factors in a population with high type 2 diabetes incidence. The dataset includes 532 complete records of women aged 21 (median age: 28, IQR: 23–38) from the Pima community near Phoenix, Arizona, with measurements

of glucose, BMI, blood pressure, and triceps skinfold thickness (TSF, mm) (R. G. Nelson et al. 1988; Ripley 1996).

Our analysis evaluates how adiposity, represented by Triceps Skin fold thickness (mm), TSF, modulates fasting glucose (FG, mg/dL) and BP (mmHg), estimating their conditional densities. We further assess the impact of TSF on their joint distribution.

## 9.2 Results

### 9.2.1 Simulation (large N)

(A) and (B) of Figure 9.1 compare the estimated conditional PDFs and CDFs of FG against their true counterparts.

The PWE model derived PDFs closely approximate the true distributions, though minor discrepancies in peak intensities arise when conditioning on METs values distant from its location parameter. Importantly, the positions of distributional peaks remain accurately estimated. As physical activity increases, the FG distribution sharpens toward values below 100 mg/dL, a trend effectively captured by the GLM. Similarly, the CDFs shift leftward with higher METs, reflecting lower glucose concentrations. For example, the median (50th percentile) decreases from 115 mg/dL to 95 mg/dL across the METs spectrum.

(C) visualizes the conditional PDF surface of FG across METs, with shaded high-density regions (HDRs) representing 95%, 75%, and 50% probability masses. Reference dashed lines denote glucose thresholds for prediabetes (100 mg/dL) and diabetes (125 mg/dL). The 50% HDR narrows from [95–130 mg/dL] to [85–95 mg/dL] as METs increase, while the distribution mode shifts from 110 mg/dL to 90 mg/dL with a 6-MET improvement in PA. In this plot, instead of just an expected value, we are showing how the entire distribution is modified conditioned on the values of PA. Even though there is a shift of the entire distribution towards higher values of FG, this shift is not constant for all the quantiles of the distribution, because the dispersion increases as the PA decreases.

(D) displays the conditional CDF surface, highlighting key upper bound ETI. Quantile-focused interpretations (via white trend-lines) reveal, for instance, a decline in the 75th percentile from 135 mg/dL to 100 mg/dL with a 6-MET increase. Conversely, proportion-focused analysis demonstrates that 75% of high-PA individuals (12 METs) fall below 100 mg/dL, compared to 50% at moderate PA (9.5 METs) and 25% at low PA (6 METs).

Figure 9.2 represents the estimates of the bivariate joint probability density function of FG and TGY summarized with HDRs and upper bound-ETI.

The plot includes two horizontal lines representing critical thresholds for glucose metabolism (100 mg/dL and 125 mg/dL) and two vertical lines for triglyceride metabolism (150 mg/dL and 200 mg/dL). Shaded regions indicate the 95%, 75%, and 50% highest density regions (HDRs) for FG and TGY. The symmetry of the HDRs and their alignment parallel to the axes demonstrate the conditional independence of FG and TGY.

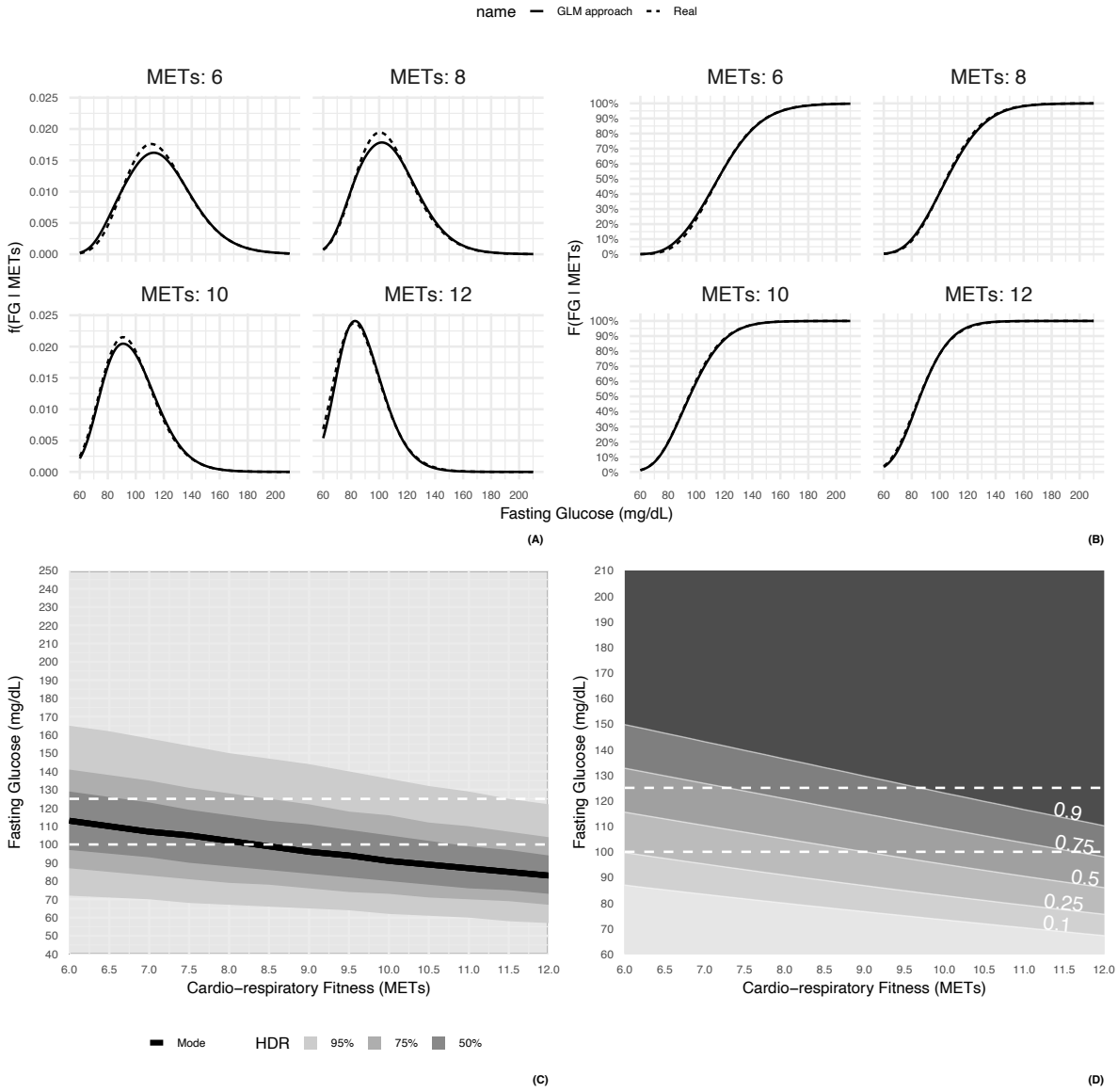



Figure 9.1: Results of the large N artificial scenario to show what the scenarios in the Montecarlo simulation procedure are. The conditional PDF and CDF of glucose given values of Physical Activity are reported (A,B), along with HDR and Upper Bound ETI (C,D)

The plot provides valuable insights into how the likelihood of specific FG and TGY joint values varies with PA. As expected, increasing physical activity shifts the probability masses from values above the critical thresholds (Panels A and B) to lower, healthier values of FG and TGY (Panels C and D), and no residual dependence is present between FG and TGY.

The shifts in joint percentiles across PA levels reflect changes in the relationship between TGY and FG as physical fitness improves. At low PA levels, the joint CDF shows higher cumulative probabilities for elevated TGY and FG values, indicating a greater probability of poorer metabolic health. As PA increase, the joint CDFs shift toward lower TGY and FG values, reflecting a reduced probability of high levels of both variables. This shift signifies improved metabolic control, with the joint distribution concentrating in the lower-left quadrant (low TGY and low FG). At the highest METs levels, the steepest rise in the joint CDF in this quadrant indicates a high probability of favorable metabolic outcomes.

### 9.2.2 Montecarlo Simulation Results

The Monte Carlo simulation confirmed that our method is reliable with moderate sample size. Overall, both modeling approaches performed well. Jaccard indexes were consistently high (mean above 0.7) across nearly all tested scenarios. However, the models' performance declined in two specific, predictable situations. Accuracy decreased at very high exposure levels (e.g., METs $> 12$). This occurred because the METs were generated from a log-normal distribution, making data points in this upper range extremely rare and insufficient for robust estimation. Moreover, Estimating HDRs for small probability levels (e.g., 25%) was challenging. These regions correspond to the peak of the PDF (the steepest part of the CDF), which naturally is characterized by higher variability. This difficulty was amplified in areas with sparse data, such as at the extreme covariate values mentioned above.

Figure 9.5 and Table 9.3 show that the two stage modeling approach combined with the chain rule of probability to obtain two dimensional HDR have a slightly lower but similar performance. In this case, results are shown just for the algorithm 'Up-down approach' for finding HDRs, as we don't have an algorithm that works on the bivariate CDF.

### 9.2.3 Pima Indian Study

We expected that, as the TSF of the women of the PIMA Indian Study increased, their metabolic health, identified by the concentration of fasting plasma glucose and diastolic blood pressure, would decrease. In the following analysis, we will consider just these three covariates. Namely we fit a piecewise exponential model of the form

$$\log[h(\text{glu}|\text{TSF})] = \beta\text{TSF} + \sum_{k=1}^{K}(\gamma_k + \text{TSF}\delta_k)B_k(\text{glu})$$

and another model for

$$\log[h(\text{bp}|\text{TSF, glu})] = (\text{TSF, glu})\beta^\tau + \sum_{k=1}^{K}(\gamma_k + \text{TSF}\delta_k)B_k(\text{bp})$$

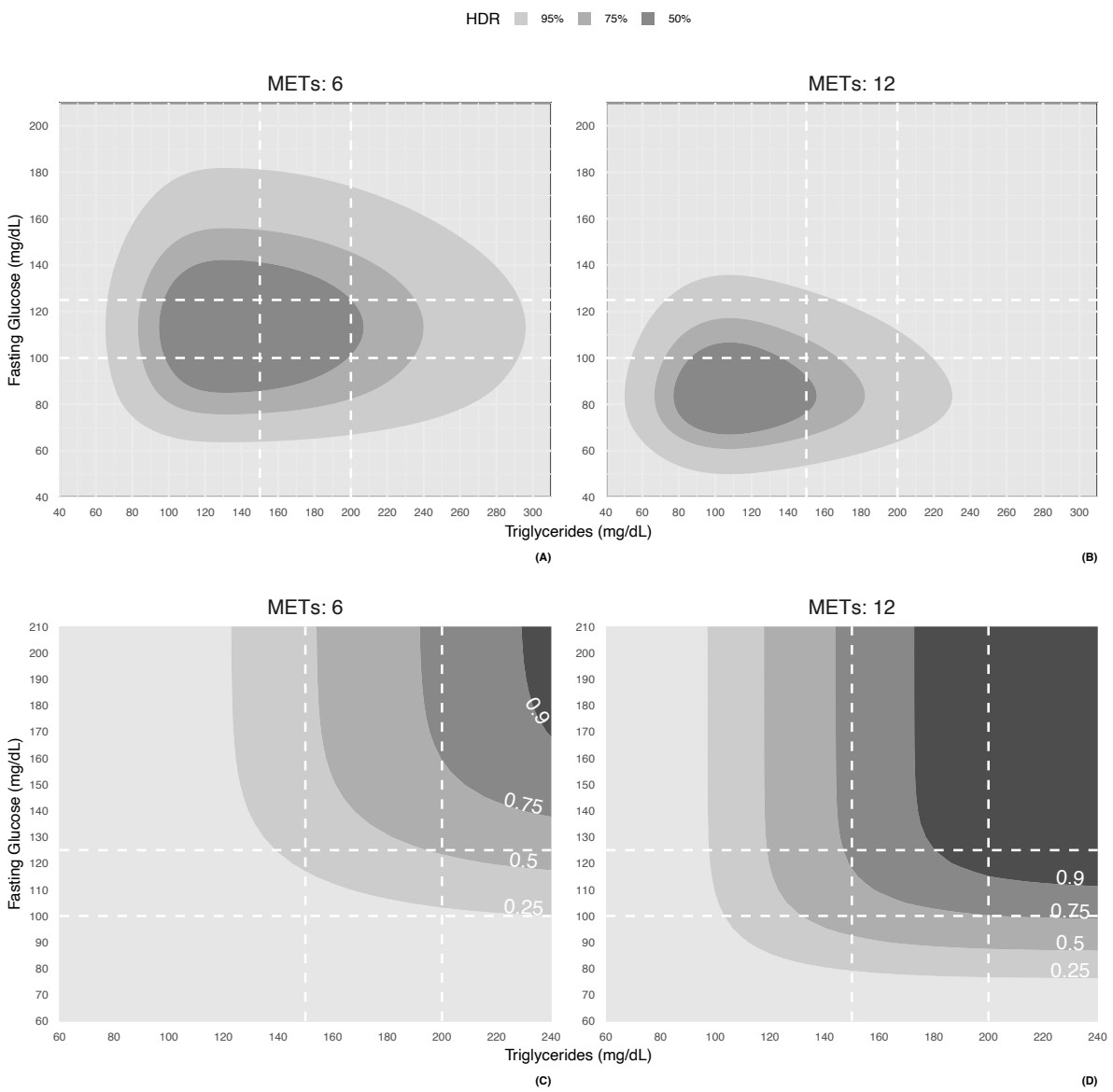



Figure 9.2: Results of the large N artificial scenario to show what the scenarios in the Montecarlo simulation procedure are. The 2 dimensional HDRs (A,B) and Upper Bound ETI (C,D) of glucose and triglycerides given values of Physical Activity are reported, along with HDR and Upper Bound ETI (C,D)

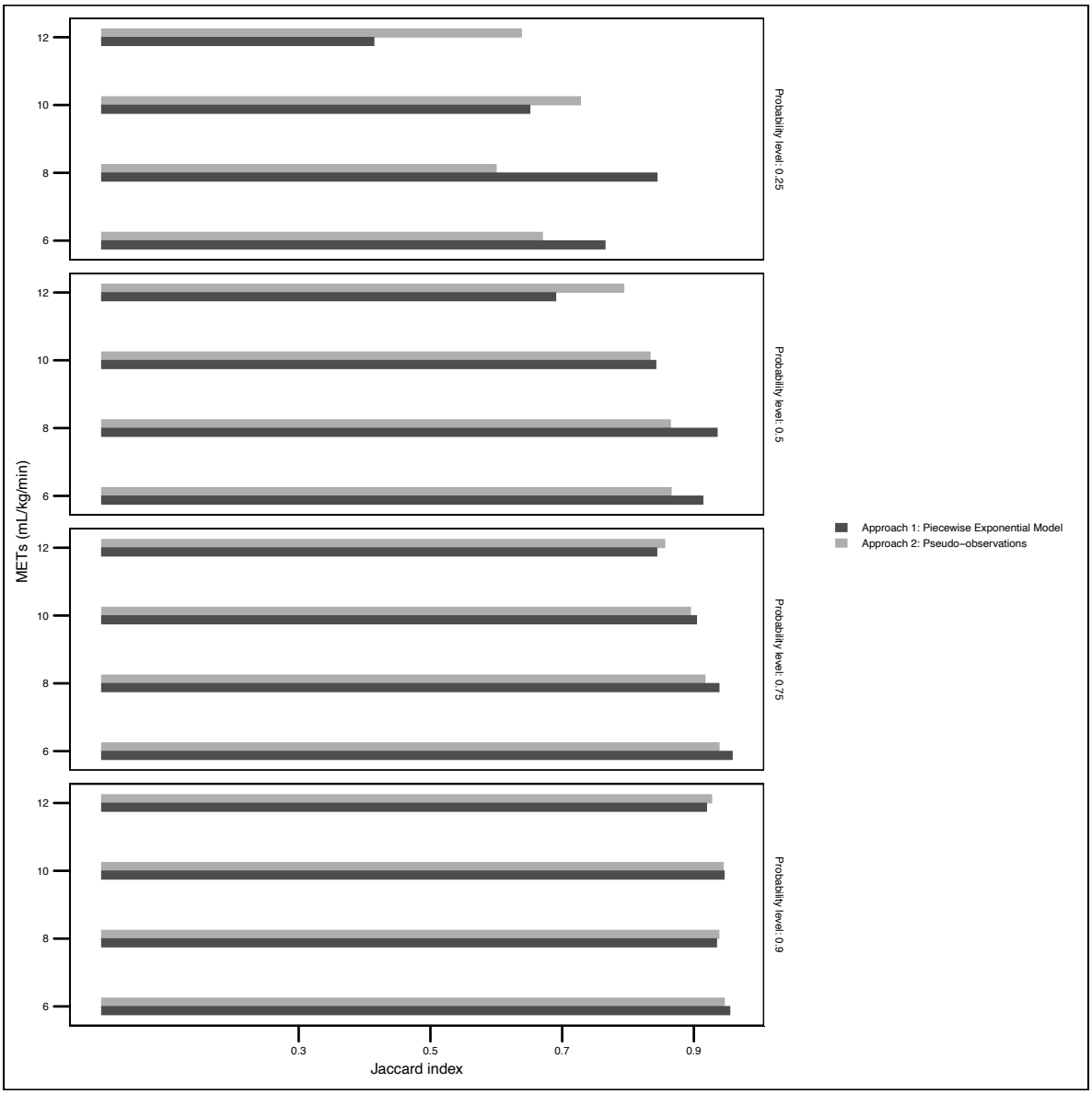



Figure 9.3: Results of the Montecarlo simulation. Finite sample performances of the PWE and the pseudo-observation model defined by the Jaccard indexes on the HDR intervals evaluated at 0.25, 0.50, 0.75 and 0.90 probability levels.

Table 9.1: Results of the Montecarlo simulation for the univariate distribution Plasma Glucose. Finite sample performances of the PWE and the pseudo-observation model defined by the Jaccard indexes on the HDR intervals evaluated at 0.25, 0.50, 0.75 and 0.90 probability levels.

(a) Up-down approach

| | **Jaccard Index** | |
|---|---|---|
| Covariate (X) | Mean | SD |
| Probability level: 0.25 | | |
| 6 | 0.766 (0.671) | 0.147 (0.188) |
| 8 | 0.845 (0.601) | 0.124 (0.231) |
| 10 | 0.652 (0.729) | 0.120 (0.176) |
| 12 | 0.415 (0.639) | 0.120 (0.192) |
| Probability level: 0.5 | | |
| 6 | 0.914 (0.866) | 0.051 (0.088) |
| 8 | 0.936 (0.865) | 0.051 (0.082) |
| 10 | 0.843 (0.834) | 0.064 (0.093) |
| 12 | 0.691 (0.794) | 0.074 (0.102) |
| Probability level: 0.75 | | |
| 6 | 0.959 (0.939) | 0.023 (0.034) |
| 8 | 0.939 (0.918) | 0.030 (0.037) |
| 10 | 0.905 (0.896) | 0.033 (0.040) |
| 12 | 0.844 (0.857) | 0.040 (0.056) |
| Probability level: 0.9 | | |
| 6 | 0.955 (0.947) | 0.022 (0.027) |
| 8 | 0.935 (0.939) | 0.022 (0.027) |
| 10 | 0.947 (0.945) | 0.024 (0.029) |
| 12 | 0.920 (0.928) | 0.027 (0.034) |

(b) Right-left approach

| | **Jaccard Index** | |
|---|---|---|
| Covariate (X) | Mean | SD |
| Probability level: 0.25 | | |
| 6 | 0.718 (0.702) | 0.142 (0.167) |
| 8 | 0.861 (0.641) | 0.128 (0.226) |
| 10 | 0.685 (0.750) | 0.115 (0.166) |
| 12 | 0.447 (0.693) | 0.123 (0.191) |
| Probability level: 0.5 | | |
| 6 | 0.893 (0.898) | 0.059 (0.069) |
| 8 | 0.934 (0.871) | 0.044 (0.078) |
| 10 | 0.808 (0.816) | 0.068 (0.096) |
| 12 | 0.710 (0.822) | 0.074 (0.097) |
| Probability level: 0.75 | | |
| 6 | 0.952 (0.944) | 0.023 (0.030) |
| 8 | 0.950 (0.935) | 0.028 (0.035) |
| 10 | 0.918 (0.915) | 0.034 (0.038) |
| 12 | 0.873 (0.887) | 0.035 (0.049) |
| Probability level: 0.9 | | |
| 6 | 0.958 (0.951) | 0.021 (0.026) |
| 8 | 0.953 (0.952) | 0.021 (0.026) |
| 10 | 0.932 (0.941) | 0.026 (0.030) |
| 12 | 0.930 (0.936) | 0.027 (0.034) |

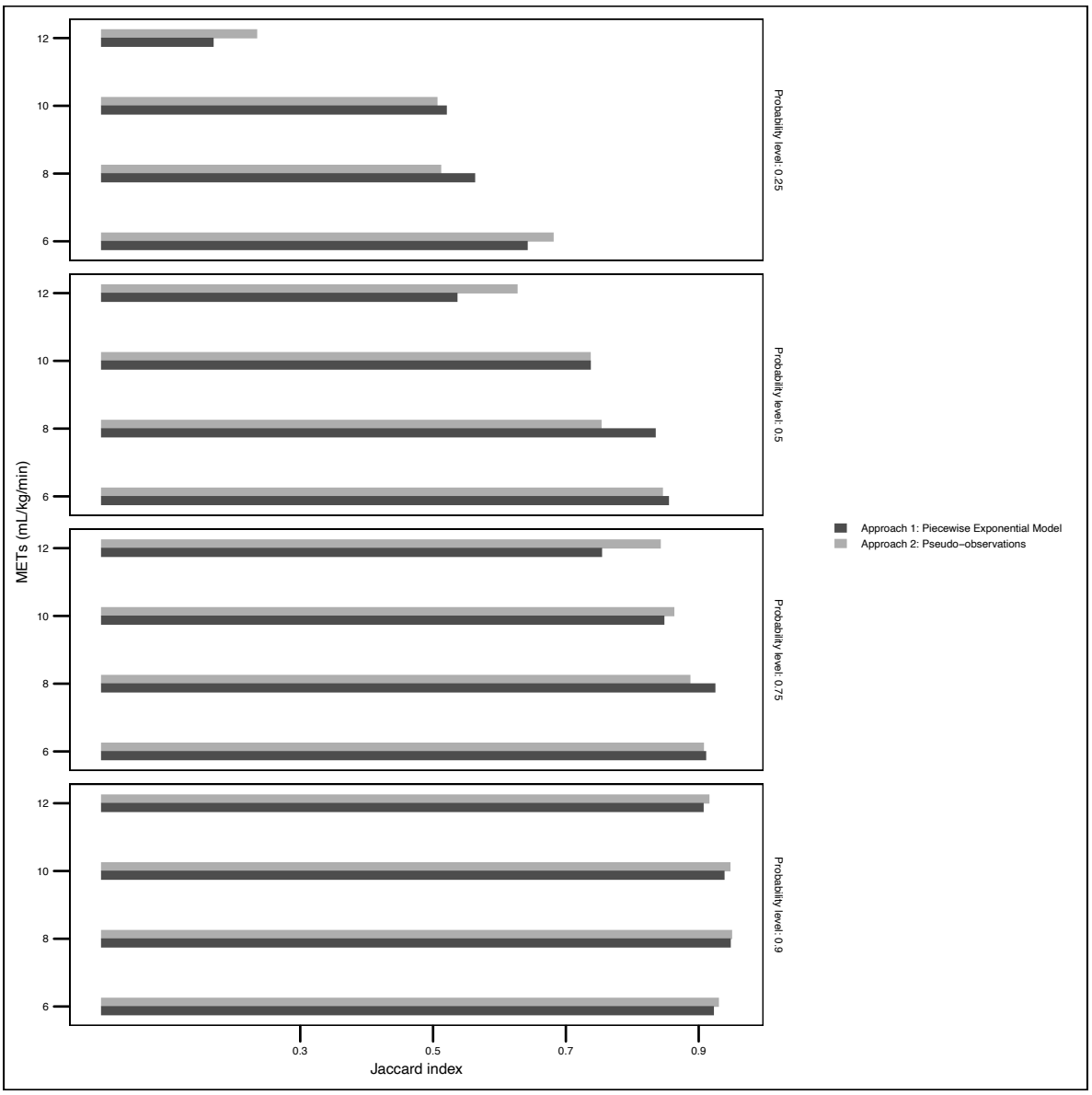



Figure 9.4: Results of the Montecarlo simulation for the univariare distribution Plasma Triglycerides. Finite sample performances of the PWE and the pseudo-observation model defined by the Jaccard indexes on the HDR intervals evaluated at 0.25, 0.50, 0.75 and 0.90 probability levels.

Table 9.2: Results of the Montecarlo simulation. Finite sample performances of the PWE and the pseudo-observation model defined by the Jaccard indexes on the HDR intervals evaluated at 0.25, 0.50, 0.75 and 0.90 probability levels.

(a) Up-down approach

| | **Jaccard Index** | |
|---|---|---|
| Covariate (X) | Mean | SD |
| Probability level: 0.25 | | |
| 6 | 0.643 (0.682) | 0.168 (0.188) |
| 8 | 0.563 (0.512) | 0.297 (0.192) |
| 10 | 0.521 (0.507) | 0.271 (0.182) |
| 12 | 0.170 (0.235) | 0.173 (0.109) |
| Probability level: 0.5 | | |
| 6 | 0.855 (0.846) | 0.065 (0.086) |
| 8 | 0.835 (0.754) | 0.130 (0.090) |
| 10 | 0.738 (0.737) | 0.134 (0.095) |
| 12 | 0.537 (0.627) | 0.132 (0.088) |
| Probability level: 0.75 | | |
| 6 | 0.911 (0.908) | 0.026 (0.036) |
| 8 | 0.925 (0.888) | 0.028 (0.038) |
| 10 | 0.848 (0.863) | 0.044 (0.049) |
| 12 | 0.755 (0.843) | 0.066 (0.051) |
| Probability level: 0.9 | | |
| 6 | 0.923 (0.930) | 0.021 (0.027) |
| 8 | 0.948 (0.950) | 0.025 (0.029) |
| 10 | 0.939 (0.948) | 0.027 (0.027) |
| 12 | 0.908 (0.916) | 0.029 (0.046) |

(b) Right left approach

| | **Jaccard Index** | |
|---|---|---|
| Covariate (X) | Mean | SD |
| Probability level: 0.25 | | |
| 6 | 0.651 (0.699) | 0.152 (0.176) |
| 8 | 0.579 (0.540) | 0.281 (0.195) |
| 10 | 0.539 (0.543) | 0.267 (0.185) |
| 12 | 0.180 (0.259) | 0.173 (0.108) |
| Probability level: 0.5 | | |
| 6 | 0.857 (0.869) | 0.058 (0.080) |
| 8 | 0.852 (0.771) | 0.085 (0.090) |
| 10 | 0.739 (0.743) | 0.135 (0.094) |
| 12 | 0.537 (0.634) | 0.131 (0.086) |
| Probability level: 0.75 | | |
| 6 | 0.907 (0.926) | 0.025 (0.032) |
| 8 | 0.924 (0.890) | 0.028 (0.037) |
| 10 | 0.845 (0.865) | 0.043 (0.048) |
| 12 | 0.778 (0.862) | 0.065 (0.047) |
| Probability level: 0.9 | | |
| 6 | 0.916 (0.937) | 0.026 (0.031) |
| 8 | 0.946 (0.950) | 0.026 (0.029) |
| 10 | 0.937 (0.948) | 0.026 (0.027) |
| 12 | 0.914 (0.916) | 0.029 (0.044) |

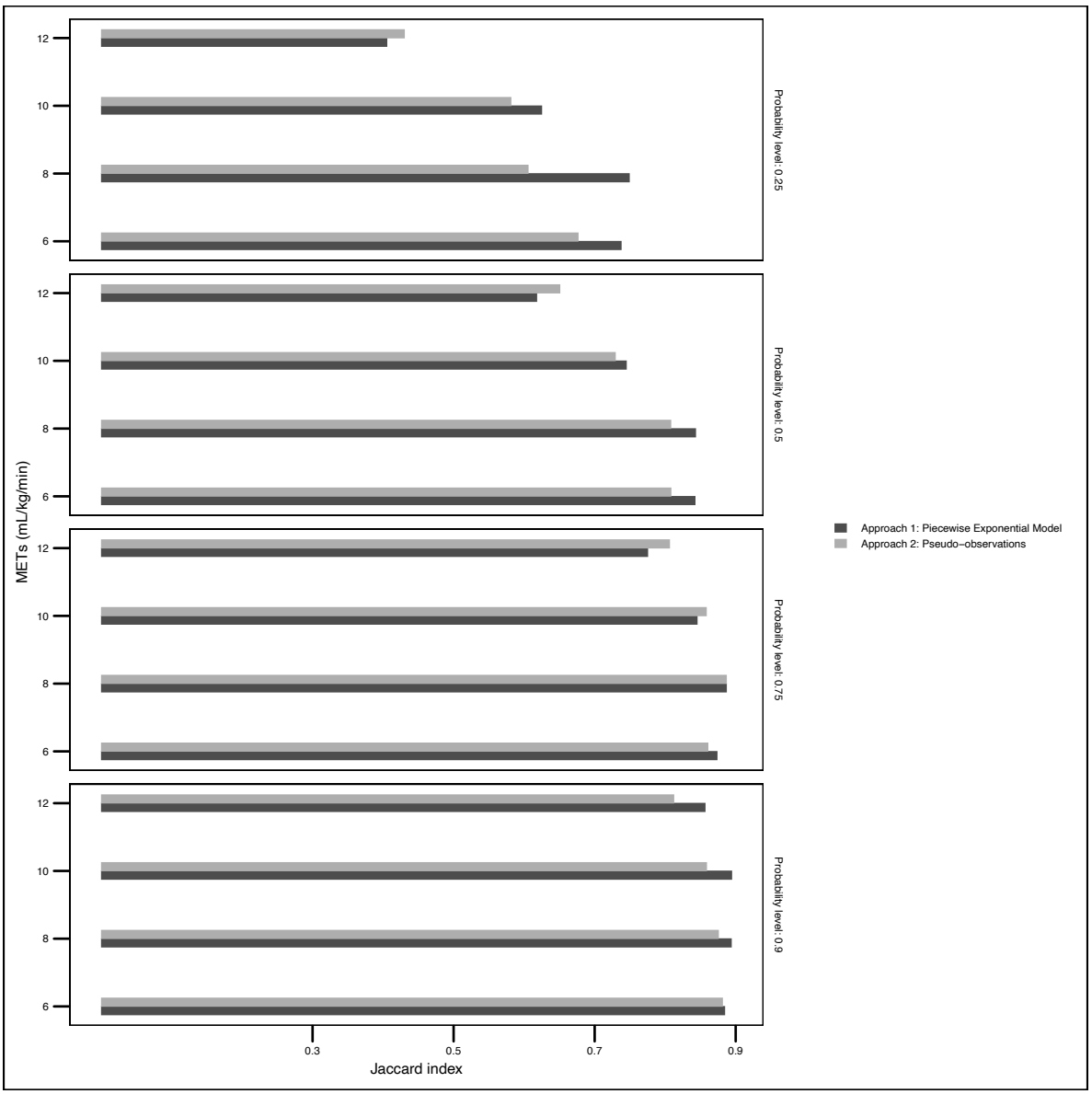



Figure 9.5: Results of the Montecarlo simulation for the bivariate distribution Plasma Glucose and Plasma Triglycerides. Finite sample performances of the PWE and the pseudo-observation model defined by the Jaccard indexes on the HDR intervals evaluated at 0.25, 0.50, 0.75 and 0.90 probability levels.

Table 9.3: Results of the Montecarlo simulation for the bivariate distribution Plasma Glucose and Plasma Triglycerides. Finite sample performances of the PWE and the pseudo-observation model, inside the brackets, defined by the Jaccard indexes on the HDR intervals evaluated at 0.25, 0.50, 0.75 and 0.90 probability levels.

**Comparison of Model Performance**

*Results for PWE Model (Pseudo-observation Model)*

| | **Jaccard Index** | |
|---|---|---|
| Covariate (X) | Mean | SD |
| Probability level: 0.25 | | |
| 6 | 0.738 (0.677) | 0.078 (0.093) |
| 8 | 0.750 (0.606) | 0.112 (0.111) |
| 10 | 0.626 (0.582) | 0.116 (0.100) |
| 12 | 0.406 (0.431) | 0.106 (0.094) |
| Probability level: 0.5 | | |
| 6 | 0.843 (0.809) | 0.036 (0.044) |
| 8 | 0.844 (0.809) | 0.042 (0.039) |
| 10 | 0.746 (0.730) | 0.051 (0.045) |
| 12 | 0.618 (0.651) | 0.069 (0.053) |
| Probability level: 0.75 | | |
| 6 | 0.874 (0.861) | 0.026 (0.030) |
| 8 | 0.887 (0.887) | 0.025 (0.025) |
| 10 | 0.846 (0.859) | 0.027 (0.027) |
| 12 | 0.776 (0.807) | 0.032 (0.031) |
| Probability level: 0.9 | | |
| 6 | 0.885 (0.882) | 0.025 (0.028) |
| 8 | 0.894 (0.876) | 0.027 (0.033) |
| 10 | 0.895 (0.859) | 0.022 (0.033) |
| 12 | 0.857 (0.813) | 0.024 (0.037) |

In particular we modeled possible a linear effect of TSF on the hazard function and we relaxed the proportionality assumption of the hazard by specifying the interaction terms $\text{TSF}\delta_k \times B_k(\text{glu})$ and $\text{TSF}\delta_k \times B_k(\text{bp})$ in the models. We used B-splines to specify the baseline hazard function associated to reference value of the covariate TSF. The number of the spline bases was selected considering the lowest value of the BIC. We transformed the hazard predictions into conditional PDFs and CDFs showed in figure 3 and we applied the bootstrap quantile method to obtain 95% simultaneous confidence bands around the predicted functions.

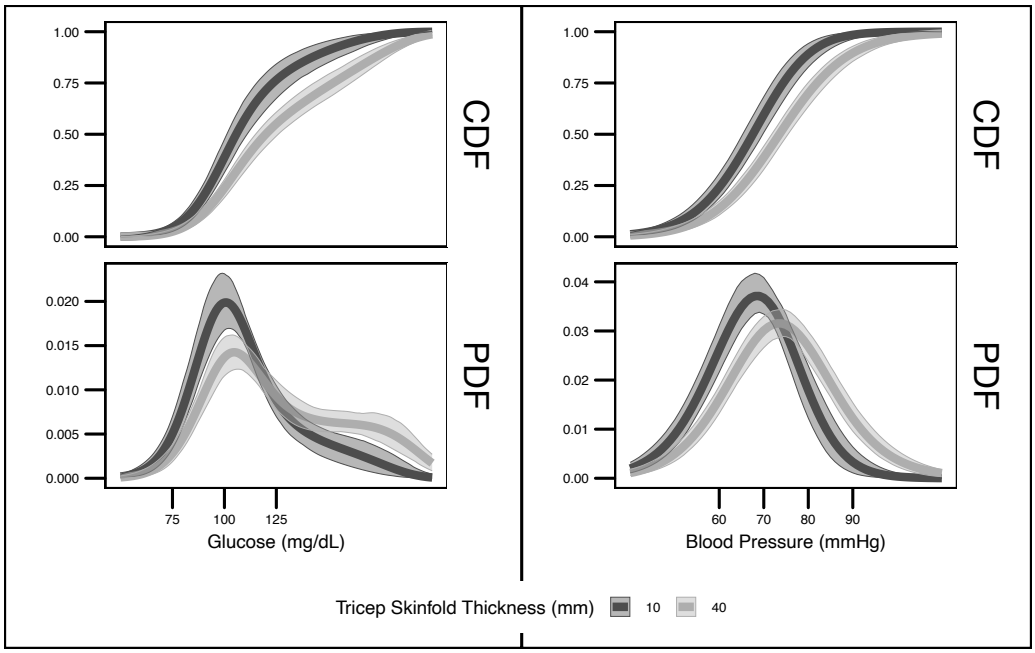



Figure 9.6: conditional PDF and CDF obtained with a PWE model targeting the hazard function, along with their 95% simultaneous confidence bands (shadowed regions) obtained with the bootstrap quantile method. The functions are displayed conditional on value of TSF of 10 and 40 mm.

Figure 9.6 illustrates the conditional cumulative distribution function and the conditional probability density estimates for FG and BP for two different levels (10 and 40 mm) of adiposity. It is clear that there is association between the TSF thickness and both plasma glucose as the cumulative distribution function estimated is shifted towards the right for higher TSF. However, for plasma glucose, the effect of adiposity is realized on the shape of its distribution more than on the location. Notably, we observe a progressive flattening of the distribution's peak, accompanied by a slightly shift in the mode and probability mass toward higher FG values. The highest density regions (HDRs) in Figure 9.7 further highlight this effect.

In contrast, BP levels are also associated with adiposity, but the relationship primarily manifests as a shift of the location and the overall distribution toward higher BP values rather than a significant change in the shape of the distribution.

Figure 9.8 displays the HDRs computed from the joint probability density estimates of fasting

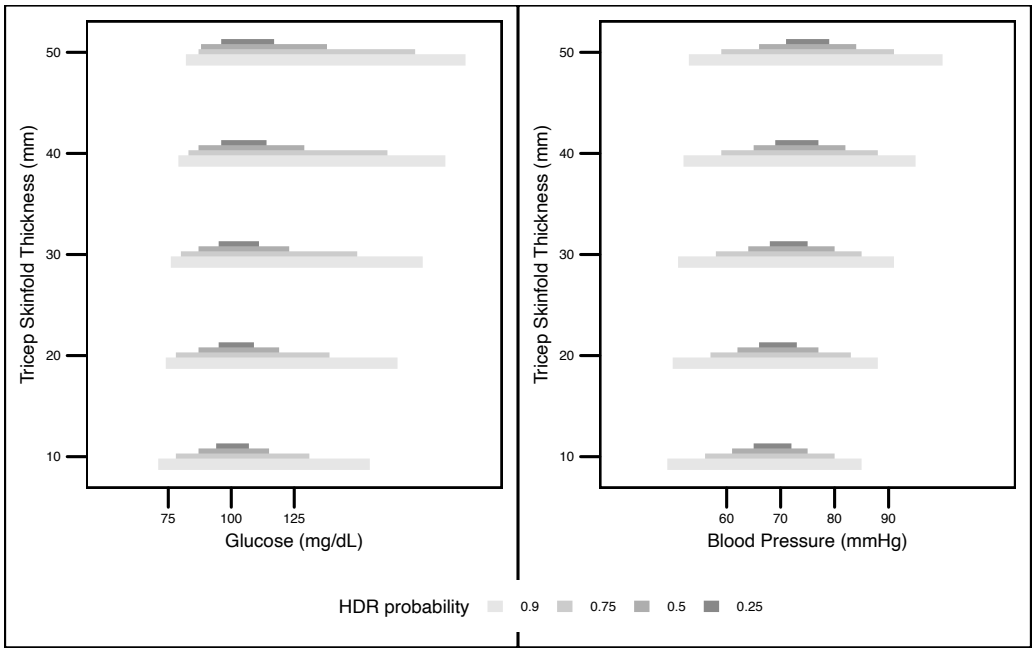



Figure 9.7: 25%, 50%, 75%, 90% HDR for plasma glucose and blood pressure in the PIMA Indians Study, obtained adopting the Up-Down algorithm.

glucose (FG) and blood pressure (BP) from lower (10 mm, panel A) to higher (50 mm, panel B) values of triceps skinfold thickness (TSF). Conditional on TSF, FG and BP appear to be independent, as the density contours are symmetric with respect to their principal axes aligned with the X and Y axes. As TSF increases, the probability mass shifts toward higher values of both FG and BP. For example, at a TSF of 10 mm, the 50% highest density region (HDR) is confined to values below 125 mg/dL for FG and below 80 mmHg for BP, with a significant portion of the density lying below 100 mg/dL and 80 mmHg, respectively. In contrast, at a TSF of 45 mm, the 50% HDR shifts to values well above 125 mg/dL for FG and 80 mmHg for BP. Notably, the effect of adiposity is not limited to a simple shift in the distribution toward higher values of the metabolic components; it also involves a significant change in the shape of the joint probability density.

## 9.3 Discussion

This study introduces a flexible, semiparametric framework for modeling the entire conditional distribution of continuous outcomes, representing a significant and welcome advancement over traditional regression techniques that focus solely on the conditional mean. The central aspect is the adaptation of well-established tools from survival analysis—specifically, the modeling of the hazard function, $h(y \mid X)$, using splines within a generalized linear model structure—to the context of fully observed, non-censored data. This approach directly addresses a fundamental limitation of mean-based regression: the inability to capture complex dose-response relationships where an exposure may alter not just the location (e.g., mean or median) but

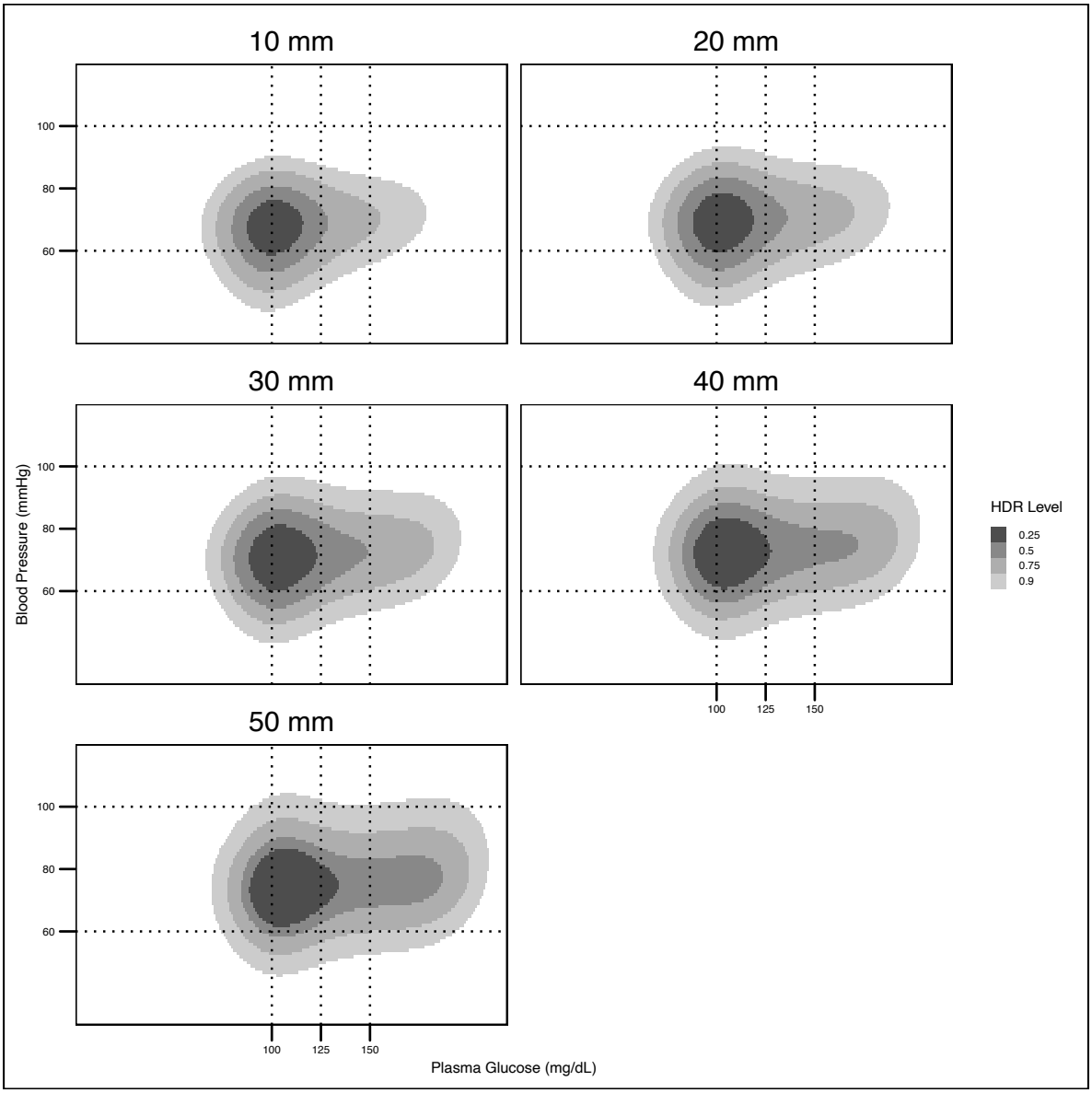



Figure 9.8: 25%, 50%, 75%, 90% HDRs computed from the joint probability density estimates of fasting glucose (FG) and blood pressure (BP).

also the scale (variance) and shape (skewness, modality) of an outcome's distribution, as highlighted in Figure 1 of the paper.

By leveraging the mathematical machinery developed for time-to-event data, the proposed method provides a robust and interpretable way to describe how the entire probability landscape of a health outcome shifts in response to varying levels of an exposure. The core of the method lies in estimating a smooth, conditional hazard or cumulative distribution function and then using the fundamental relationships between the hazard, cumulative distribution, and density functions to derive a complete picture of the outcome's behavior. This repurposing of survival methodology, for a different class of problems offers a new lens through which to examine dose-response phenomena in epidemiology and public health.

The Monte Carlo simulation study provided valuable initial evidence for the proposed method's performance and finite-sample properties under controlled conditions. The consistently high Jaccard index values (mean $> 0.7$ for most scenarios) for the overlap between estimated and true Highest Density Regions are encouraging and suggest that the method is well-calibrated in moderately dense data regions. However, the performance degradation observed in two specific scenarios warrants a deeper and more critical consideration. The reduced accuracy at the extremes of the covariate distribution (e.g., METs $> 12$) and the difficulty in estimating low-probability HDRs (e.g., 25%) could be due to data sparsity and the inherent variability at the mode of a distribution, respectively. However, spline-based models are well-known to exhibit higher variance at the boundaries of the data support, where there are fewer data points to anchor the curve. In the present work, we did not employ any form of penalization or regularization for the spline terms, a standard practice in the generalized additive model (GAM) literature specifically designed to control over fitting and improve stability, particularly in sparse data regions. Thus, the observed performance drop should therefore be viewed not merely as an unavoidable consequence of sparsity, but as a test of the model's numerical stability that reveals an area for potential improvement. Future work should rigorously investigate the integration of penalized splines (e.g., thin plate regression splines with smoothness selection via cross-validation or REML) to enhance the model's robustness and predictive performance at the edges of the covariate distribution.

Interestingly, our proposed modeling approaches are in direct dialogue with the Cumulative Probability Models (CPMs) (Liu et al. 2017) . The Proportional Odds (PO) logistic regression model is the most prominent example, as a uniquely robust, flexible, and general framework for analyzing both ordinal and continuous outcomes. The core philosophy underpinning this approach is to model a monotonic transformation of the cumulative probability, $g(P(Y \leq y \mid X))$, using a semiparametric structure. In this structure, a series of intercept terms nonparametrically estimates the baseline CDF, while a set of regression coefficients estimates the effects of covariates on this distribution. The transformation model on pseudo-observations adopted in the previous paper is essentially a cumulative probability model when, with fully observed outcomes, pseudo-observations are just values 0 and 1. However, instead of fitting discrete intercepts terms, we adopted natural splines to specify the functional shape of the cumulative distribution function. Moreover, we relaxed the proportionality of the odds or the hazard assumption including interaction terms between the baseline function and the covariates.

The Piece-wise exponential model models the continuous hazard function of the outcome. Interestingly, also the Proportional Continuation-Ratio, a form of a cumulative probability model, models a discrete version of the hazard function of an outcome variable. The PCR model can be seen as a discrete counterpart to the PWE model. Just as the PWE model offers a flexible way to approximate a continuous hazard function without imposing a strict parametric form, the PCR model provides a flexible framework for understanding the "hazard" of progressing through an ordered set of outcomes. This fundamental similarity underscores a unified conceptual approach to modeling continuous outcomes. The essential difference lies in the likelihood adopted by the models. While the PCR adopts a multinomial likelihood, the Piece-wise exponential model adopts a poisson likelihood.

When compared to alternative frameworks, the proposed method occupies a strategic middle ground. It avoids the rigid parametric assumptions and potential convergence issues of Generalized Additive Models for Location, Scale, and Shape (GAMLSS), offering greater robustness at the cost of some flexibility. In contrast to quantile regression, which provides a "discretized" view by modeling specific quantiles independently, this unified approach yields a complete and internally consistent picture of the distribution, sidestepping the issue of crossing quantile functions.

In conclusion, this work makes a contribution by presenting a robust, flexible, and interpretable tool for dose-response analysis. It successfully bridges concepts from survival analysis and ordinal regression to offer a compelling alternative for modeling continuous outcomes. While opportunities for refinement exist—particularly in model diagnostics, computational efficiency, and extensions to dependent and multivariate data—the framework provides researchers with a powerful method to move beyond the mean and uncover a more complete understanding of how exposures shape the entire probability landscape of health outcomes.

# 10 Conclusions

Statistical models have always been fascinating because they can construct a powerful and rigorous narrative of reality. Their power lies in the flexibility of learning from data, while their rigorousness lies in providing robust generalizations of that narrative through inference and probability.

Nowadays, flexibility is a must. We often encounter biological processes and phenomena that rarely hold the assumptions of classic parametric models. Moreover, considering the data and computational power revolution we have witnessed, we can—and must—move beyond parsimony to embrace complexity, with the aim of providing personalized and meaningful insights to policymakers, stakeholders, and patients.

At present, we have achieved remarkable performance in flexibly modeling complex biomedical phenomena using sophisticated statistical, machine learning, and even 'deep learning' models. Still, the most common models adopted in the clinical and epidemiological literature often favor parsimony. A perfect example is the continued reliance on the linear model for continuous non-censored health outcomes and the Cox model (and its extensions to competing risks) for censored health outcomes.

While the appeal of a parsimonious model as 'being simple to compute' no longer applies—we don't need to manually use matrix algebra, and computational power has fueled the development of numerous optimization algorithms—I argue that the appeal of 'being easy to communicate' remains prevalent, especially in the public health community. A $\beta$ coefficient in a linear model, accompanied by a 'statistically significant' p-value, is a straightforward way to communicate an association. Similarly, a $\beta$ coefficient and its p-value in a Cox proportional hazards routine offer a deceptively simple interpretation for a censored outcome. The limitations of p-values have been extensively critiqued in the statistical literature. Briefly, this single, often misinterpreted number has come to dominate statistical practice, leading to poor scientific decision-making. This thesis is in line with a more thoughtful and nuanced approach to statistical inference.

This is not to suggest a lack of statistical sophistication among public health colleagues—I had the pleasure of collaborating with many excellent professionals during my PhD. Rather, it highlights the inherent challenge of comprehending a time-varying hazard ratio, a difficulty I frequently encountered when communicating these findings in clinical and epidemiological papers.

Statistical models within the Generalized Linear Model framework offer a good trade-off between flexibility and interpretability. Moreover, the causal inference machinery (e.g., IPW,

propensity score matching, g-formula) is easily implemented. As described in this thesis, flexibility is obtained by specifying the functional form of the baseline hazard/risk and adding interaction terms with covariates to model time-varying effects. Worrying about non-proportional hazards is not merely a statistical refinement: if one were to model a survival scenario whose data-generating mechanism comes from a normal model (the simplest case for non-censored outcomes), one would have to specify either non-proportional hazards or non-proportional risks for the covariates.

The interpretability of a statistical model refers to the degree to which a human can understand the cause and effect of its internal workings. In the context of the flexible parametric models used in this thesis, we must specify not only the prognostic relationship between the covariates but also the degree of smoothness of the baseline distributional functions and their interactions with other covariates.

This duality presents both an opportunity and a challenge for scientists. An opportunity, because contrary to advanced ML models, scientists have a full grasp of why a model is outputting certain results. A challenge, because it is the scientist's responsibility to specify the correct model structure for a fair prediction and communication of risk.

The thesis proposed a framework for the responsible application and communication of flexible parametric models in public health and addressed three principal bottlenecks for the analysis of time-to-event data for researchers:

The Data Bottleneck: Researchers may ask, "Do I have enough data to reliably fit these complex models?

The Methodological Bottleneck: Researchers may then ask, "How do I correctly build, specify, and validate these models?"

The Communication Bottleneck: Finally, researchers must ask, "How do I explain the complex, time-varying results to a non-statistical audience of clinicians, policymakers, and patients?"

In the first part of the thesis, I addressed the questions of how complex (flexible) the model should be and how many events are needed to obtain robust estimates, dealing with the first two bottlenecks.

In the study on the number of events per parameter (EPP) required for accurate causal measures of the average treatment effect, I found that 20 EPP are recommended, and that parameters associated with splines and time-varying effects should be included in the parameter count for EPP calculations. This was shown using both parametric models and more complex scenarios resembling real clinical data, confirming these findings. This suggests that while flexible parametric models are an invaluable tool, they require a greater number of events to achieve the required flexibility. This might be less of a problem when analyzing registry data but may be problematic when analyzing or reanalyzing small clinical trials.

Having at least 20 EPP is just the first step. A second aspect of the thesis was the formulation of a workflow for flexible model specification to estimate cumulative probabilities. I emphasized the need to go beyond statistical tests for assessing interaction or time-varying effects, favoring instead perturbation procedures with the bootstrap, accompanied by measures of discrimination and indexes like the Net Reclassification Improvement (all corrected for optimism)

evaluated on the cumulative quantity itself. Focusing on predictive measures of cumulative quantities is highly suggested when the average treatment effect is the estimand of interest, but it is also valuable for communicating the prognostic relationship of covariates. Indeed, evaluating how a time-varying effect practically improves the reclassification of a prognostic model—treating that term as if it were a new biomarker—allows us to decide, in agreement with clinicians, whether the improvement is relevant enough to retain that component in the algorithm.

The present work does not underestimate the relevance of Schoenfeld residuals for detecting violations of proportional hazards. However, the standard Schoenfeld test implicitly checks for a linear relationship between the residuals and time. If the true departure from proportionality is non-linear, the test may lack power to detect this crucial model misspecification. This limitation highlights the urgent need for more assumption-free and clinically relevant tools to guide model specification.

For this precise reason, the workflow developed in this thesis moves beyond a reliance on such tests. Instead, it favors robust perturbation procedures with the bootstrap, accompanied by measures of discrimination and reclassification evaluated on the cumulative quantity of interest. This approach allows for the detection of meaningful departures from proportionality, regardless of their functional form, thereby providing a more reliable basis for model selection.

While this workflow provides a rigorous methodology for model construction, the challenge of interpreting and communicating the complex outputs of such models—particularly time-varying effects—remains. A statistically sound model is of limited practical value if its findings cannot be made accessible to clinicians, policymakers, and patients. In the last part of the thesis, I addressed the third bottleneck for researchers, formalizing a new way of communicating treatment effects in time-to-event analysis, especially useful when time-varying effects are modeled.

In the end, time-to-event analysis is about time, a continuous outcome. As with any continuous outcome, it is about communicating the uncertainty around a most likely value, conditional on prognostic information. Whether the health outcome is disease-free survival time or plasma glucose, the focus is on the redistribution of probability or risk associated with (or caused by) an exposure.

Explaining uncertainty to a scientifically diverse audience presents a significant challenge. While showing adjusted survival or cumulative distribution functions might seem a straightforward approach for communicating a probability distribution, it is known to be difficult to interpret for patients and policymakers (Fuller, Dudley, and Blacktop 2004; De La Maza et al. 2018).

A key contribution of the second part of the thesis was the formalization of novel (causal) measures of the treatment effect in time-to-event analysis—the restricted Highest Risk Density Region (HRDR) and the Highest Net Risk Difference (HNRDR) region. These measures provide an intuitive, visual representation of formal causal estimands, such as the difference in potential survival functions ($S^{(1)}(t) - S^{(0)}(t)$), thereby making the magnitude, timing, and uncertainty of a treatment's effect more accessible to a clinical audience.

In particular, I evaluated flexible frameworks for obtaining and displaying distributional effects efficiently. I adopted two modeling frameworks to obtain and display the probability density functions (PDFs) along with the cumulative distribution functions (CDFs) of the outcome. This is known to be an effective way to communicate uncertainty (Ibrekk and Morgan 1987). In this sense, the HRDR and HNRDR, restricted to the end of follow-up, improve the interpretation of these functions and their differences.

The concept of displaying risk redistribution with HRDR and HNRDR is also applicable within the classical Cox model. However, I argue that the adoption of flexible models is favored for the resulting smoothness of the PDF and CDF estimates, which is also preferred by the algorithms used to construct the intervals. In this framework, the problem of communicating a time-varying hazard ratio becomes less central, as the focus shifts to the effect on the whole distribution—though its correct modeling remains critical.

The use of flexible models to derive a PDF or CDF from a hazard function comes naturally from the basic relationships between these distributional functions. In the context of non-censored data, PDF computation from a discrete hazard function has been adopted using a 'pooled logistic regression' framework to derive IP weights for continuous exposures in marginal structural models (Díaz Muñoz and Laan 2011). This 'pooled logistic regression' is the classic proportional continuation-ratio model, discrete-time hazard model, which has been extensively assessed in this thesis and is a well-discussed ordinal model (McCullagh 1985) and (Liu et al. 2017) .

Although we might consider a single summary interval, a drawback of any probability interval, HRDR and HNRDR included, is that they can 'hide' the true shape of the distribution, so that observers may perceive all values within the interval as equally probable (Spiegelhalter, Pearson, and Short 2011) . Thus, I consider it critical to display the whole PDF along with the intervals, or at least several intervals at different probability levels, to better summarize risk redistribution.

The estimation performance of these intervals has been promising. Some future perspectives are extending the HRDR and HNRDR framework to handle more complex scenarios, such as competing risks. Developing user-friendly software packages or plugins (e.g., in R or Stata) to facilitate the broader adoption of these novel measures by the applied research community. Conducting empirical studies with clinicians and patients to formally evaluate the effectiveness of these new visual aids in improving risk comprehension and shared decision-making.

The public health community continues to rely on parsimonious statistical models, such as the Cox proportional hazards model, due to their perceived communicability. This reliance persists despite the frequent violation of their underlying assumptions and the availability of more powerful, flexible alternatives capable of capturing complex biomedical phenomena. This thesis has addressed the critical challenge of how to robustly specify, validate, and communicate the findings from these more sophisticated models, with the aim of bridging the gap between statistical innovation and its practical application in clinical and epidemiological research.

# References


"A Report on the Natural Duration of Cancer." 1927. *JAMA: The Journal of the American Medical Association* 88 (7): 507. https://doi.org/10.1001/jama.1927.02680330059037.

Aitkin, Murray, Brain Francis, and John Hinde. 2005. *Statistical Modelling in GLIM 4*. Oxford University PressOxford. https://doi.org/10.1093/oso/9780198524137.001.0001.

Akaike, Hirotogu. 1992. "Information Theory and an Extension of the Maximum Likelihood Principle." In, 610–24. Springer New York. https://doi.org/10.1007/978-1-4612-0919-5_38.

Alba, Ana Carolina, Thomas Agoritsas, Michael Walsh, Steven Hanna, Alfonso Iorio, P. J. Devereaux, Thomas McGinn, and Gordon Guyatt. 2017. "Discrimination and Calibration of Clinical Prediction Models." *JAMA* 318 (14): 1377. https://doi.org/10.1001/jama.2017.12126.

Altman, Douglas G, and Patrick Royston. 2006. "The Cost of Dichotomising Continuous Variables." *BMJ* 332 (7549): 1080.1. https://doi.org/10.1136/bmj.332.7549.1080.

Ambrogi, Federico, Elia Biganzoli, and Patrizia Boracchi. 2008. "Estimates of Clinically Useful Measures in Competing Risks Survival Analysis." *Statistics in Medicine* 27 (30): 6407–25. https://doi.org/10.1002/sim.3455.

———. 2009. "Estimating Crude Cumulative Incidences Through Multinomial Logit Regression on Discrete Cause-Specific Hazards." *Computational Statistics & Data Analysis* 53 (7): 2767–79. https://doi.org/10.1016/j.csda.2009.01.001.

Antolini, Laura, Patrizia Boracchi, and Elia Biganzoli. 2005. "A Time-Dependent Discrimination Index for Survival Data." *Statistics in Medicine* 24 (24): 3927–44. https://doi.org/10.1002/sim.2427.

Barnett, Ofra, and Ayala Cohen. 2000. "The Histogram and Boxplot for the Display of Lifetime Data." *Journal of Computational and Graphical Statistics* 9 (4): 759–78. https://doi.org/10.1080/10618600.2000.10474912.

Beckstead, Jason W., and Theresa M. Beckie. 2010. "How Much Information Can Metabolic Syndrome Provide?" *Medical Decision Making* 31 (1): 79–92. https://doi.org/10.1177/0272989x10373401.

Belot, Aurélien, Michal Abrahamowicz, Laurent Remontet, and Roch Giorgi. 2010. "Flexible Modeling of Competing Risks in Survival Analysis." *Statistics in Medicine* 29 (23): 2453–68. https://doi.org/10.1002/sim.4005.

Bender, Andreas, Fabian Scheipl, Johannes Piller, and Philipp Kopper. 2018. "Pammtools: Piece-Wise Exponential Additive Mixed Modeling Tools for Survival Analysis." The R Foundation. https://doi.org/10.32614/cran.package.pammtools.

Berger, Moritz, Matthias Schmid, Thomas Welchowski, Steffen Schmitz-Valckenberg, and Jan Beyersmann. 2018. "Subdistribution Hazard Models for Competing Risks in Discrete Time." *Biostatistics* 21 (3): 449–66. https://doi.org/10.1093/biostatistics/kxy069.

Beyersmann, Jan, Aurélien Latouche, Anika Buchholz, and Martin Schumacher. 2009. "Sim-

ulating Competing Risks Data in Survival Analysis." *Statistics in Medicine* 28 (6): 956–71. https://doi.org/10.1002/sim.3516.

Biganzoli, Elia M., Patrizia Boracchi, Federico Ambrogi, and Ettore Marubini. 2006. "Artificial Neural Network for the Joint Modelling of Discrete Cause-Specific Hazards." *Artificial Intelligence in Medicine* 37 (2): 119–30. https://doi.org/10.1016/j.artmed.2006.01.004.

Biganzoli, Giacomo, Giuseppe Marano, and Patrizia Boracchi. 2025. "Modeling Strategies for a Flexible Estimation of the Crude Cumulative Incidence in the Context of Long Follow-Ups: Model Choice and Predictive Ability Evaluation." *BMC Medical Research Methodology* 25 (1). https://doi.org/10.1186/s12874-025-02650-x.

Bohl, Alex A., David K. Blough, Paul A. Fishman, Jeffery R. Harris, and Elizabeth A. Phelan. 2012. "Are Generalized Additive Models for Location, Scale, and Shape an Improvement on Existing Models for Estimating Skewed and Heteroskedastic Cost Data?" *Health Services and Outcomes Research Methodology* 13 (1): 18–38. https://doi.org/10.1007/s10742-012-0086-x.

Boracchi, Patrizia, Elia Biganzoli, and Ettore Marubini. 2003. "Joint Modelling of Cause-Specific Hazard Functions with Cubic Splines: An Application to a Large Series of Breast Cancer Patients." *Computational Statistics & Data Analysis* 42 (1-2): 243–62. https://doi.org/10.1016/s0167-9473(02)00122-6.

Burnham, Kenneth P., and David R. Anderson, eds. 2004. *Model Selection and Multimodel Inference*. Springer New York. https://doi.org/10.1007/b97636.

Canale, Antonio, Daniele Durante, and David B. Dunson. 2018. "Convex Mixture Regression for Quantitative Risk Assessment." *Biometrics* 74 (4): 1331–40. https://doi.org/10.1111/biom.12917.

Cheng, S. C., Jason P. Fine, and L. J. Wei. 1998. "Prediction of Cumulative Incidence Function Under the Proportional Hazards Model." *Biometrics* 54 (1): 219. https://doi.org/10.2307/2534009.

Cox, D. R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 34 (2): 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Crowther, Michael J., and Paul C. Lambert. 2013. "Simulating Biologically Plausible Complex Survival Data." *Statistics in Medicine* 32 (23): 4118–34. https://doi.org/10.1002/sim.5823.

Cui, James. 2007. "QIC Program and Model Selection in GEE Analyses." *The Stata Journal: Promoting Communications on Statistics and Stata* 7 (2): 209–20. https://doi.org/10.1177/1536867x0700700205.

Dawson, Neal V., and Robert Weiss. 2012. "Dichotomizing Continuous Variables in Statistical Analysis." *Medical Decision Making* 32 (2): 225–26. https://doi.org/10.1177/0272989x12437605.

De La Maza, Cristóbal, Alex Davis, Cleotilde Gonzalez, and Inês Azevedo. 2018. "Understanding Cumulative Risk Perception from Judgments and Choices: An Application to Flood Risks." *Risk Analysis* 39 (2): 488–504. https://doi.org/10.1111/risa.13206.

Díaz Muñoz, Iván, and Mark J. van der Laan. 2011. "Super Learner Based Conditional Density Estimation with Application to Marginal Structural Models." *The International Journal of Biostatistics* 7 (1): 1–20. https://doi.org/10.2202/1557-4679.1356.

Doll, R., and A. B. Hill. 1956. "Lung Cancer and Other Causes of Death in Relation to Smoking." *BMJ* 2 (5001): 1071–81. https://doi.org/10.1136/bmj.2.5001.1071.

Donoghoe, Mark W., and Val Gebski. 2017. "The Importance of Censoring in Competing Risks Analysis of the Subdistribution Hazard." *BMC Medical Research Methodology* 17 (1). https://doi.org/10.1186/s12874-017-0327-3.

Efron, Bradley, and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429246593.

Fan, J. 2004. "A Crossvalidation Method for Estimating Conditional Densities." *Biometrika* 91 (4): 819–34. https://doi.org/10.1093/biomet/91.4.819.

Fine, Jason P., and Robert J. Gray. 1999. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association* 94 (446): 496–509. https://doi.org/10.1080/01621459.1999.10474144.

Fornili, Marco, Patrizia Boracchi, Federico Ambrogi, and Elia Biganzoli. 2018. "Modeling the Covariates Effects on the Hazard Function by Piecewise Exponential Artificial Neural Networks: An Application to a Controlled Clinical Trial on Renal Carcinoma." *BMC Bioinformatics* 19 (S7). https://doi.org/10.1186/s12859-018-2179-1.

Fuller, R, N Dudley, and J Blacktop. 2004. "Older People's Understanding of Cumulative Risks When Provided with Annual Stroke Risk Information." *Postgraduate Medical Journal* 80 (949): 677–78. https://doi.org/10.1136/pgmj.2004.019489.

Gehan, Edmund A, and Emil J Freireich. 2011. "The 6-MP Versus Placebo Clinical Trial in Acute Leukemia." *Clinical Trials* 8 (3): 288–97. https://doi.org/10.1177/1740774511407358.

Hall, Peter, Rodney C. L. Wolff, and Qiwei Yao. 1999. "Methods for Estimating a Conditional Distribution Function." *Journal of the American Statistical Association* 94 (445): 154–63. https://doi.org/10.1080/01621459.1999.10473832.

Harrell , Frank E. 2015. *Regression Modeling Strategies*. Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7.

HARRELL, FRANK E., KERRY L. LEE, and DANIEL B. MARK. 1996. "MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS." *Statistics in Medicine* 15 (4): 361–87. https://doi.org/10.1002/(sici)1097-0258(19960229)15:4%3C361::aid-sim168%3E3.0.co;2-4.

Hernán, Miguel A. 2010. "The Hazards of Hazard Ratios." *Epidemiology* 21 (1): 13–15. https://doi.org/10.1097/ede.0b013e3181c1ea43.

Hilden, Jørgen, and Thomas A. Gerds. 2013. "A Note on the Evaluation of Novel Biomarkers: Do Not Rely on Integrated Discrimination Improvement and Net Reclassification Index." *Statistics in Medicine* 33 (19): 3405–14. https://doi.org/10.1002/sim.5804.

Hyndman, Rob J. 1995. "Highest-Density Forecast Regions for Nonlinear and Non-Normal Time Series Models." *Journal of Forecasting* 14 (5): 431–41. https://doi.org/10.1002/for.3980140503.

Hyndman, Rob J., and Qiwei Yao. 2002. "Nonparametric Estimation and Symmetry Tests for Conditional Density Functions." *Journal of Nonparametric Statistics* 14 (3): 259–78. https://doi.org/10.1080/10485250212374.

Ibrekk, Harald, and M. Granger Morgan. 1987. "Graphical Communication of Uncertain Quantities to Nontechnical People." *Risk Analysis* 7 (4): 519–29. https://doi.org/10.1111/j.1539-6924.1987.tb00488.x.

Infante, Gabriele, Rosalba Miceli, and Federico Ambrogi. 2023. "Sample Size and Predictive Performance of Machine Learning Methods with Survival Data: A Simulation Study."

*Statistics in Medicine* 42 (30): 5657–75. https://doi.org/10.1002/sim.9931.

Ishwaran, Hemant, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange, and Bryan M. Lau. 2014. "Random Survival Forests for Competing Risks." *Biostatistics* 15 (4): 757–73. https://doi.org/10.1093/biostatistics/kxu010.

Kamarudin, Adina Najwa, Trevor Cox, and Ruwanthi Kolamunnage-Dona. 2017. "Time-Dependent ROC Curve Analysis in Medical Research: Current Methods and Applications." *BMC Medical Research Methodology* 17 (1). https://doi.org/10.1186/s12874-017-0332-6.

Kantidakis, Georgios, Hein Putter, Saskia Litière, and Marta Fiocco. 2023. "Statistical Models Versus Machine Learning for Competing Risks: Development and Validation of Prognostic Models." *BMC Medical Research Methodology* 23 (1). https://doi.org/10.1186/s12874-023-01866-z.

Kaplan, E. L., and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282): 457–81. https://doi.org/10.1080/01621459.1958.10501452.

Khadka, Aayush, Jillian Hebert, M. Maria Glymour, Fei Jiang, Amanda Irish, Kate Duchowny, and Anusha M. Vable. 2023. "Quantile Regressions as a Tool to Evaluate How an Exposure Shifts and Reshapes the Outcome Distribution: A Primer for Epidemiologists." http://dx.doi.org/10.1101/2023.05.02.23289415.

Klein, John P., and Per Kragh Andersen. 2005. "Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function." *Biometrics* 61 (1): 223–29. https://doi.org/10.1111/j.0006-341x.2005.031209.x.

Koenker, Roger, and Olga Geling. 2001. "Reappraising Medfly Longevity." *Journal of the American Statistical Association* 96 (454): 458–68. https://doi.org/10.1198/016214501753168172.

Koenker, Roger, and Kevin F Hallock. 2001. "Quantile Regression." *Journal of Economic Perspectives* 15 (4): 143–56. https://doi.org/10.1257/jep.15.4.143.

Lazarus-Barlow, W. S., and J. H. Leeming. 1924. "THE NATURAL DURATION OF CANCER." *BMJ* 2 (3320): 266–67. https://doi.org/10.1136/bmj.2.3320.266.

Leening, Maarten J. G., Ewout W. Steyerberg, Ben Van Calster, Ralph B. D'Agostino, and Michael J. Pencina. 2014. "Net Reclassification Improvement and Integrated Discrimination Improvement Require Calibrated Models: Relevance from a Marker and Model Perspective." *Statistics in Medicine* 33 (19): 3415–18. https://doi.org/10.1002/sim.6133.

Liu, Qi, Bryan E. Shepherd, Chun Li, and Frank E. Harrell. 2017. "Modeling Continuous Response Variables Using Ordinal Regression." *Statistics in Medicine* 36 (27): 4316–35. https://doi.org/10.1002/sim.7433.

Marano, Giuseppe, Giacomo Biganzoli, Ester Luconi, Elia Mario Biganzoli, and Patrizia Boracchi. 2025. "Can Smoothing Methods Recognize the Patterns of the Hazard Function in Complex Clinical Scenarios?" In, 43–57. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-90714-2_4.

Marra, Giampiero, and Simon N. Wood. 2011. "Practical Variable Selection for Generalized Additive Models." *Computational Statistics & Data Analysis* 55 (7): 2372–87. https://doi.org/10.1016/j.csda.2011.02.004.

May, Susanne, and Carol Bigelow. 2005. "Modeling Nonlinear Dose-Response Relationships in Epidemiologic Studies: Statistical Approaches and Practical Challenges." *Dose-Response* 3 (4). https://doi.org/10.2203/dose-response.003.04.004.

McCullagh, Peter. 1985. "Analysis of Ordinal Categorical Data." *Technometrics* 27 (3): 317–18. https://doi.org/10.1080/00401706.1985.10488059.

Montiel Olea, José Luis, and Mikkel Plagborg-Møller. 2018. "Simultaneous Confidence Bands: Theory, Implementation, and an Application to SVARs." *Journal of Applied Econometrics* 34 (1): 1–17. https://doi.org/10.1002/jae.2656.

Nadaraya, E. A. 1964. "On Estimating Regression." *Theory of Probability & Its Applications* 9 (1): 141–42. https://doi.org/10.1137/1109020.

Nelson, Robert G., James E. Everhart, William. C. Knowler, and Peter H. Bennett. 1988. "Incidence, Prevalence and Risk Factors for Non-Insulin-Dependent Diabetes Mellitus." *Primary Care: Clinics in Office Practice* 15 (2): 227–50. https://doi.org/10.1016/s0095-4543(21)01074-5.

Nelson, Wayne. 1969. "Hazard Plotting for Incomplete Failure Data." *Journal of Quality Technology* 1 (1): 27–52. https://doi.org/10.1080/00224065.1969.11980344.

O'Neill, Ben. 2021. "Computing Highest Density Regions for Continuous Univariate Distributions with Known Probability Functions." *Computational Statistics* 37 (2): 613–49. https://doi.org/10.1007/s00180-021-01133-z.

Peduzzi, Peter, John Concato, Alvan R. Feinstein, and Theodore R. Holford. 1995. "Importance of Events Per Independent Variable in Proportional Hazards Regression Analysis II. Accuracy and Precision of Regression Estimates." *Journal of Clinical Epidemiology* 48 (12): 1503–10. https://doi.org/10.1016/0895-4356(95)00048-8.

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. 1996. "A Simulation Study of the Number of Events Per Variable in Logistic Regression Analysis." *Journal of Clinical Epidemiology* 49 (12): 1373–79. https://doi.org/10.1016/s0895-4356(96)00236-3.

Pencina, Michael J., Ralph B. D'Agostino, and Ewout W. Steyerberg. 2010. "Extensions of Net Reclassification Improvement Calculations to Measure Usefulness of New Biomarkers." *Statistics in Medicine* 30 (1): 11–21. https://doi.org/10.1002/sim.4085.

Peng, Limin. 2021. "Quantile Regression for Survival Data." *Annual Review of Statistics and Its Application* 8 (1): 413–37. https://doi.org/10.1146/annurev-statistics-042720-020233.

Péron, Julien, Pascal Roy, Brice Ozenne, Laurent Roche, and Marc Buyse. 2016. "The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials." *JAMA Oncology* 2 (7): 901. https://doi.org/10.1001/jamaoncol.2015.6359.

Piccart, M. J. 2000. "Randomized Intergroup Trial of Cisplatin-Paclitaxel Versus Cisplatin-Cyclophosphamide in Women with Advanced Epithelial Ovarian Cancer: Three-Year Results." *Journal of the National Cancer Institute* 92 (9): 699–708. https://doi.org/10.1093/jnci/92.9.699.

Putter, Hein, Martin Schumacher, and Hans C. van Houwelingen. 2020. "On the Relation Between the Cause-Specific Hazard and the Subdistribution Rate for Competing Risks Data: The Fine–Gray Model Revisited." *Biometrical Journal* 62 (3): 790–807. https://doi.org/10.1002/bimj.201800274.

Ragland, David R. 1992. "Dichotomizing Continuous Outcome Variables: Dependence of the Magnitude of Association and Statistical Power on the Cutpoint." *Epidemiology* 3 (5): 434–40. https://doi.org/10.1097/00001648-199209000-00009.

Ramsay, J. O. 1988. "Monotone Regression Splines in Action." *Statistical Science* 3 (4).

https://doi.org/10.1214/ss/1177012761.

Redelmeier, Donald A., and Jonathan S. Zipursky. 2023. "A Dose of Reality About Dose–Response Relationships." *Journal of General Internal Medicine* 38 (16): 3604–9. https://doi.org/10.1007/s11606-023-08395-x.

Retsky, Michael, and Romano Demicheli. 2014. "Multimodal Hazard Rate for Relapse in Breast Cancer: Quality of Data and Calibration of Computer Simulation." *Cancers* 6 (4): 2343–55. https://doi.org/10.3390/cancers6042343.

Riley, Richard D, Kym IE Snell, Joie Ensor, Danielle L Burke, Frank E Harrell Jr, Karel GM Moons, and Gary S Collins. 2018. "Minimum Sample Size for Developing a Multivariable Prediction Model: PART II - Binary and Time-to-Event Outcomes." *Statistics in Medicine* 38 (7): 1276–96. https://doi.org/10.1002/sim.7992.

Ripley, Brian D. 1996. "Pattern Recognition and Neural Networks," January. https://doi.org/10.1017/cbo9780511812651.

Rodríguez-Girondo, Mar, Thomas Kneib, Carmen Cadarso-Suárez, and Emad Abu-Assi. 2013. "Model Building in Nonproportional Hazard Regression." *Statistics in Medicine* 32 (30): 5301–14. https://doi.org/10.1002/sim.5961.

Rosenberg, Philip S. 1995. "Hazard Function Estimation Using b-Splines." *Biometrics* 51 (3): 874. https://doi.org/10.2307/2532989.

Royston, Patrick, Douglas G. Altman, and Willi Sauerbrei. 2005. "Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea." *Statistics in Medicine* 25 (1): 127–41. https://doi.org/10.1002/sim.2331.

Ruan, Ping K., and Robert J. Gray. 2008. "Analyses of Cumulative Incidence Functions via Non-Parametric Multiple Imputation." *Statistics in Medicine* 27 (27): 5709–24. https://doi.org/10.1002/sim.3402.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701. https://doi.org/10.1037/h0037350.

Rudolph, Jacqueline E., Catherine R. Lesko, and Ashley I. Naimi. 2020. "Causal Inference in the Face of Competing Events." *Current Epidemiology Reports* 7 (3): 125–31. https://doi.org/10.1007/s40471-020-00240-7.

Spiegelhalter, David, Mike Pearson, and Ian Short. 2011. "Visualizing Uncertainty About the Future." *Science* 333 (6048): 1393–1400. https://doi.org/10.1126/science.1191181.

Stasinopoulos, D. Mikis, and Robert A. Rigby. 2007. "Generalized Additive Models for Location Scale and Shape (GAMLSS) in*R*." *Journal of Statistical Software* 23 (7). https://doi.org/10.18637/jss.v023.i07.

Stensrud, Mats J, John M Aalen, Odd O Aalen, and Morten Valberg. 2018. "Limitations of Hazard Ratios in Clinical Trials." *European Heart Journal* 40 (17): 1378–83. https://doi.org/10.1093/eurheartj/ehy770.

Strohmaier, Susanne, Kjetil Røysland, Rune Hoff, Ørnulf Borgan, Terje R. Pedersen, and Odd O. Aalen. 2015. "Dynamic Path Analysis – a Useful Tool to Investigate Mediation Processes in Clinical Survival Trials." *Statistics in Medicine* 34 (29): 3866–87. https://doi.org/10.1002/sim.6598.

Sutradhar, Rinku, and Peter C. Austin. 2018. "Relative Rates Not Relative Risks: Addressing a Widespread Misinterpretation of Hazard Ratios." *Annals of Epidemiology* 28 (1): 54–57. https://doi.org/10.1016/j.annepidem.2017.10.014.

Tutz, Gerhard, and Matthias Schmid. 2016. *Modeling Discrete Time-to-Event Data*. Springer

International Publishing. https://doi.org/10.1007/978-3-319-28158-2.

Uno, Hajime, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, Toshio Miyata, Deborah Schrag, et al. 2014. "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis." *Journal of Clinical Oncology* 32 (22): 2380–85. https://doi.org/10.1200/jco.2014.55.2208.

Venables, W. N., and B. D. Ripley. 2002. "Modern Applied Statistics with s." https://www.stats.ox.ac.uk/pub/MASS4/.

Verbeeck, Johan, and Everardo D Saad. 2024. "Rethinking Survival Analysis: Advancing Beyond the Hazard Ratio?" *European Heart Journal: Acute Cardiovascular Care* 13 (3): 313–15. https://doi.org/10.1093/ehjacc/zuae017.

Veronesi, U., E. Marubini, L. Mariani, V. Galimberti, A. Luini, P. Veronesi, B. Salvadori, and R. Zucali. 2001. "Radiotherapy After Breast-Conserving Surgery in Small Breast Carcinoma: Long-Term Results of a Randomized Trial." *Annals of Oncology* 12 (7): 997–1003. https://doi.org/10.1023/a:1011136326943.

Veronesi, Umberto, Roberto Saccozzi, Marcella Del Vecchio, Alberto Banfi, Claudio Clemente, Mario De Lena, Giuseppe Gallus, et al. 1981. "Comparing Radical Mastectomy with Quadrantectomy, Axillary Dissection, and Radiotherapy in Patients with Small Cancers of the Breast." *New England Journal of Medicine* 305 (1): 6–11. https://doi.org/10.1056/nejm198107023050102.

Wang, Duolao, Con Ariti, Tim Collier, and Stuart Pocock. 2011. "The Win Ratio: A New Approach to the Analysis of Composite Endpoints in Clinical Trials Based on Clinical Priorities." *Trials* 12 (S1). https://doi.org/10.1186/1745-6215-12-s1-a69.

Weir, I. R., G. D. Marshall, J. I. Schneider, J. A. Sherer, E. M. Lord, B. Gyawali, M. K. Paasche-Orlow, E. J. Benjamin, and L. Trinquart. 2019. "Interpretation of Time-to-Event Outcomes in Randomized Trials: An Online Randomized Experiment." *Annals of Oncology* 30 (1): 96–102. https://doi.org/10.1093/annonc/mdy462.

Wolbers, M., P. Blanche, M. T. Koller, J. C. M. Witteman, and T. A. Gerds. 2014. "Concordance for Prognostic Models with Competing Risks." *Biostatistics* 15 (3): 526–39. https://doi.org/10.1093/biostatistics/kxt059.

Zavorsky, Gerald S. 2025. "Debunking the GAMLSS Myth: Simplicity Reigns in Pulmonary Function Diagnostics." *Respiratory Medicine* 236 (January): 107836. https://doi.org/10.1016/j.rmed.2024.107836.

Zeng, Peng, Cheng Jiang, Anbang Liu, Xinyuan Yang, Feng Lin, and Lingli Cheng. 2024. "Association of Systemic Immunity-Inflammation Index with Metabolic Syndrome in U.S. Adult: A Cross-Sectional Study." *BMC Geriatrics* 24 (1). https://doi.org/10.1186/s12877-023-04635-1.

Zhou, Bingqing, Aurelien Latouche, Vanderson Rocha, and Jason Fine. 2010. "Competing Risks Regression for Stratified Data." *Biometrics* 67 (2): 661–70. https://doi.org/10.1111/j.1541-0420.2010.01493.x.