

https://dspace.org/

A simulation study to explore the dangers of ignoring competing

Nephawe, Mafanedza

2026
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/50302
Nephawe, Mafanedza. (2026). A simulation study to explore the dangers of ignoring competing [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/50302

Downloaded from a DSpace repository. https://dspace.org/

A simulation study to explore the dangers of ignoring competing

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Mafanedza Nephawe
1442906

Supervised by Dr Justine Nasejje

January 18, 2026







Ethics Clearance Number: WCSAM-2025-01

A Research Paper submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the Master of Science e-science degree

Abstract

This study investigates the consequences of incorrectly handling competing risks in survival analysis, particularly applied to simulated customer churn and multimorbidity data. Traditional survival analysis methods often neglect competing risks, potentially leading to biased estimations and inaccurate conclusions. This research aims to evaluate the impact of neglecting competing risks using comprehensive simulation studies, addressing a significant gap in practical methodological validation.

To achieve this, synthetic datasets representing customer churn scenarios and multimorbidity contexts were generated, encompassing varying sample sizes of 1,000, 2,000, and 3,000.The Churn data was simulated to be linear and the multimorbidity data was simulated to be non-linear. Four competing risks survival models Cause Specific Cox, Random Survival Forest (RSF), DeepHit, and the Fine-Gray model were employed and rigorously compared. Each model was implemented to evaluate discriminative performance, calibration accuracy, and predictive consistency using key metrics: C-index, Brier score, and ROC AUC, with five-fold cross-validation applied to ensure robustness.

The results indicated significant variations in model performance and clearly demonstrated the implications of ignoring competing risks. In the customer churn scenarios, the Cox proportional hazards model consistently exhibited superior discriminative performance (C-index and ROC AUC values generally above 0.8) and superior calibration (Brier scores around 0.1) across all dataset sizes (N = 1000, 2000, 3000). This can be attributed to the churn data being simulated to be linear and the presence of competing risk being negligible. Random Survival Forest (RSF) and DeepHit models performed comparably well, though without substantial gains from increased sample sizes, whereas the Fine-Gray model consistently showed the weakest performance, with lower discriminative abilities (C-index and ROC AUC around 0.6) and poorer calibration (Brier scores around 0.2).

In multimorbidity scenarios, RSF and DeepHit emerged as the best performing models, achieving high discrimination (C-index and ROC AUC around or exceeding 0.8) and excellent calibration (Brier scores close to 0.1), particularly with larger dataset sizes. Traditional methods (Cox and Fine-Gray) displayed moderate to low performance, particularly for Disease B (C-index around 0.4-0.6 and higher Brier scores around 0.2-0.3), indicating limited effectiveness in complex health scenarios when compared to machine learning methods.

The study underscores the substantial risks of neglecting competing events in survival analysis, clearly showing that failure to properly account for competing risks can inflate event probabilities and distort predictive accuracy. This research reaffirms the need to carefully select analytical approaches aligned with the specific aims and available data in real-world applications. In conclusion, accurately modeling competing risks is crucial across domains such as business analytics and clinical research. This work provides empirical evidence to guide researchers and practitioners in adopting appropriate competing-risk methodologies, ultimately promoting more precise, reliable, and domain-specific survival analyses.

Declaration

I, Mafanedza Nephawe, affirm that the contents of this research proposal are entirely my own work. This Research Paper is submitted in pursuit of a Master of Science degree in e-science at the University of the Witwatersrand. It has not been submitted to any other institution or for any other degree.



Acknowledgements

I am deeply grateful to my supervisor, Dr. Justine Nasejje, for her generous sharing of knowledge, expertise, and invaluable patience and feedback. I also wish to acknowledge the financial support provided by the DSI-NICIS National e-Science Postgraduate Teaching and Training Platform (NEPTTP), which made this project possible. Their assistance has been crucial in my academic journey.

Research Output

The following manuscript, based on the simillar work presented in this thesis, is currently under review

- Mafanedza Nephawe, Pierre-Philippe Dechant, and Justine B. Nasejje. *DeepFrailty and machine learning modelling for predicting under-five survival in southern Africa*. *Annals of Epidemiology*, Manuscript Number: AEP-D-25-00309 (Under review).

Contents

Preface
Abstract i
Declaration ii
Acknowledgements iii
Table of Contents iv

1 Introduction **1**
1.1 Background 1
1.1.1 Problem Statement 2
1.1.2 Aim 2
1.1.3 Objectives 3
1.2 Overview of the Dissertation 3

2 Literature Review **4**
2.1 Nature of survival analysis 4
2.2 Competing risk models 6
2.2.1 Cause Specific Model 7
2.2.2 Fine-Gray Model 8
2.3 Machine learning in general 8
2.3.1 RSF 8
2.4 Deeplearning 9
2.4.1 DeepHit 9
2.5 Evaluation techniques 11
2.5.1 C-Index 11
2.5.2 Cumulative/Dynamic C-index: 12
2.5.3 Time-dependent Area Under the Receiver Operating Characteristic curve (ROC) 12
2.5.4 Brier Score: 13
2.5.5 Integrated Brier Score (IBS): 13

3 Methodology **14**
3.1 Data 14
3.1.1 Simulating Competing Risk Data for Customer Churn 14
3.1.2 Simulating Multimorbidity Situations in Health Science 15
3.2 Model Fitting 18

3.2.1 Customer Churn 19
3.2.2 Multimorbidity 20

4 Results and Discussion **21**
4.1 On Dangers of Ignoring Competing Risks 21
4.2 Customer Churn Results 23
4.2.1 Cross validation results 23
4.3 Multimobidity Results 26
4.3.1 Cross Validation Results 26
4.4 Discussion 30
4.4.1 Model Performance and Implications 30
4.4.2 Comparative Analysis: Customer Churn vs Multimorbidity 31
4.4.3 Alignment and Divergence with Existing Literature 31
4.4.4 Strengths and Limitations of the Study 32

5 Conclusion **33**
5.0.1 Future Research Directions 34

References **37**

Chapter 1

Introduction

This opening chapter describes the context and importance of competing risks in survival analysis. Introduces the research problem, the aim and the objectives, highlighting the shortcomings that arise when events are not properly modelled. The chapter also discusses how neglecting competing risks can bias outcomes, leading to over- or underestimation of event probabilities.

1.1 Background

Survival analysis is a method in the statistics field that looks on examining time-to-event data, primarily concerned with the amount of time till a particular occurrence occurs [Stanish *et al.* 1978]. However, in real-world scenarios, an individual may be exposed to various types of events that can prevent the happening of the event of interest these are referred to as competing risks [Jiang *et al.* 2018].

Competing risks can be defined by considering the probability of observing a specific event of interest when other events might preclude its occurrence. When competing risks are present, the overall probability of experiencing a particular event, say event 1, is not simply its marginal probability because the occurrence of another event, say event 2, might prevent event 1 from ever happening[Noordzij *et al.* 2013]. This leads to a situation where standard survival analysis methods, which often treat competing events as censoring, can overestimate the cumulative incidence of the event of interest.

Overlooking competing risks can result in biased estimates and inaccurate conclusions. Conventional survival analysis methods like the Cox proportional hazards model and the Kaplan-Meier estimator, fail to account for competing risks and may overestimate the occurrence rate of the event in question.

Several studies have highlighted the significance of considering competing risks in survival analysis. Fine and Gray [1999] created a subdistribution hazard model that directly estimates the cumulative incidence function when competing risks are present. Putter *et al.* [2007] discussed approaches for analyzing data with competing risks and emphasized the need for appropriate statistical techniques.

Competing risk has found successful applications in various domains, including credit. Blumenstock *et al.* [2022] explored innovative machine learning approaches in survival analysis, particularly focusing on competing risk within US mortgages. They assess the efficacy of DeepHit, which is competing risk deep-learning model, alongside random survival forests, thereby advancing survival analysis through the integration of deep learning techniques and addressing complex risk scenarios. Similarly, Pierri and Caroni [2020] delves into assessing bankruptcy risks for companies, utilizing survival analysis, competing risks, and multistate models to comprehend factors influencing firm bankruptcy, providing a comprehensive framework for predicting financial distress and guiding financial risk management decisions.

Fine and Gray [1999] Highlighted that, competing risk models facilitate a nuanced understanding of risk by simultaneously considering multiple outcomes, thus enhancing risk assessment accuracy compared to traditional methods. They offer flexibility in modeling various event types and dependencies, accommodating diverse applications and censoring mechanisms [Lu 2014].

While existing literature acknowledges the presence of competing risks, most research focuses primarily on theoretical aspects and basic applications of competing risks in survival analysis, therefore there is a significant gap in thoroughly exploring the impact through simulated scenarios, by building on the foundations of existing literature and applying and testing theoretical concepts in practical, simulated environments, this study fills this gap through providing empirical evidence on the consequences of neglecting competing risks.

This study uses a simulation study to effectively show and highlight the consequences of ignoring competing risks in survival analysis. A simulation study is an effective method to explore the impact of not paying close attention to competing risks. By creating a controlled environment where the true event probabilities are known, researchers can compare and assess how different statistical methods perform and assess the extent of bias introduced when competing risks are not properly accounted for, in our study we will incorporate both classical survival analysis models and the deep-learning models.

1.1.1 Problem Statement

Survival analysis has become increasingly complex across various fields, highlighting a critical need to address the potential consequences of disregarding competing risks. Neglecting competing risks in survival analysis methodologies can lead to significant inaccuracies. This research aims to investigate these hazards by exploring the ramifications of overlooking competing risks through simulated scenarios.

1.1.2 Aim

The aim of this study is to comprehensively examine the implications of ignoring competing risks in survival analysis through a simulation-based exploration. By simulating scenarios that reflect real-world complexities, this research seeks to clarify the potential pitfalls and consequences of overlooking competing risks, thereby contributing to

a more robust understanding of survival analysis methodologies and promoting the adoption of appropriate risk assessment strategies across various domains.

1.1.3 Objectives

1. To design and implement a series of simulation scenarios that accurately reflect real-world situations where competing risks are present in survival analysis and systematically analyze the impact of ignoring competing risks on the accuracy and reliability of survival analysis outcomes.
2. To compare results obtained with and without consideration of competing risks using both classical survival analysis models and deep-learning models such as Random survival forest, fine-gray model, Cause specific model and Deephit.
3. To provide empirical evidence and insights into the potential consequences of disregarding competing risks in survival analysis, thereby informing researchers and practitioners about the importance of incorporating appropriate methodologies for handling competing risk data.

1.2 Overview of the Dissertation

This dissertation is organised into five chapters, each building upon the findings and methods introduced in the preceding sections. Following this introductory chapter, which outlines the motivation, aim, and objectives of the study, the Literature Review examines relevant research on survival analysis, competing risks, and the specific machine-learning techniques deployed. The subsequent Methodology chapter presents the design and simulation of the data, describing how we systematically generated customer churn and multimorbidity scenarios and then fitted the four competing-risks models. The Results and Discussion chapter integrates numerical outcomes with interpretive analysis, linking each experiment's findings to existing scholarly work and underscoring the implications for both business and clinical domains. Finally, the Conclusion summarises the key contributions, addresses limitations, and offers suggestions for future research.

Chapter 2

Literature Review

This chapter surveys foundational concepts in survival analysis, placing emphasis on the nature of competing risks and their formal definitions. It contrasts cause-specific and subdistribution hazard modelling, then examines advanced approaches such as random survival forests and deep learning. Empirical studies and relevant theoretical frameworks are also explored, setting a foundation for our methodological design.

2.1 Nature of survival analysis

Survival analysis, is an indispensable tool in various fields, it is characterized by its focus on time-to-event outcomes and its adept handling of censored data. This methodology finds its roots in seminal works like Cox [1972] where the Kaplan-Meier estimator was introduced. This non-parametric method remains fundamental in estimating survival probabilities from censored data, enabling researchers to account for incomplete information and derive unbiased estimates of event occurrence probabilities over time.

Censoring is a prevalent feature in survival data that occurs when the precise timing of the incident is undisclosed for some persons due to factors namely loss to follow-up or the conclusion of the study period. Klein *et al.* [2003] provides an in-depth exploration of censoring mechanisms and methodologies for appropriately handling censored data in survival analysis. By incorporating techniques like Kaplan-Meier estimation and Cox regression modeling, researchers can effectively utilize censored data and extract meaningful insights regarding event occurrence probabilities.

The definitions and descriptions of survival and hazard functions are central to survival analysis. These functions provide invaluable understanding of the timing and risk factors contributing to the events of interest. Cox [1972] laid the groundwork for regression models in survival analysis, as well as the Cox proportional hazards model. This influential work introduced a flexible framework for examining the effects of covariates on survival time, facilitating the identification of factors influencing event occurrence probabilities.

Furthermore, survival analysis encompasses a diverse array of statistical techniques, including parametric and non-parametric methods, as well as regression modeling and competing risks analysis. Therneau *et al.* [2000] delves into advanced topics in survival analysis, offering insights into extending the Cox model and accommodating complex survival data structures. By leveraging these techniques, researchers can gain a deeper comprehension of the timing and risk factors contributing to the events of interest, thereby informing decision-making processes in various domains.

In summary, survival analysis, rooted in influential works by Kaplan, Meier, Cox, and others, offers a powerful framework for analyzing time-to-event outcomes and handling censored data. Through the approximation and analysis of survival and hazard functions, along with the application of diverse statistical techniques, survival analysis enables researchers to derive meaningful insights into event occurrence probabilities and the factors influencing them.

The Basic Functions Of Survival Analysis

Survival analysis, a statistical method commonly applied in medical, social, and engineering sciences, examines the duration until a particular event takes place. It is especially useful for investigating the time to failure or the duration until events such as death, relapse, or mechanical system failure. The fundamental concepts of survival analysis as well as the survival function, hazard function, censoring, and the Kaplan-Meier estimator.

Let T be a non-negative random variable representing the time until the event of interest occurs. The distribution of T can be characterized by the following basic functions:

The probability density function, denoted by $f(t)$, is defined as the limit of the probability that an event occurs within a small interval $[t, t+\Delta t)$, per unit time:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t)}{\Delta t}. \tag{2.1}$$

The survival function, $S(t)$, represents the probability that an individual survives beyond time t. It is the complement of the cumulative distribution function $F(t) = P(T \leq t)$:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)\, du. \tag{2.2}$$

Properties of $S(t)$ include $S(0) = 1$, $S(\infty) = 0$, and that it is a non-increasing function of time.

The hazard function, $h(t)$, also known as the hazard rate or instantaneous failure rate, is the probability that the event occurs in the infinitesimal interval $[t, t+\Delta t)$, given that the individual has survived up to time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.3}$$

This function is crucial in survival analysis as it describes the instantaneous risk of the event at any specific time point.

The cumulative hazard function, $H(t)$, describes the accumulated risk up to time t and is defined as the integral of the hazard function:

$$H(t) = \int_0^t h(u)\, du. \tag{2.4}$$

A fundamental relationship in survival analysis links the survival function directly to the cumulative hazard function:

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t h(u)\, du\right). \tag{2.5}$$

This relationship allows for the estimation of the survival function once the hazard form is specified.

Censoring takes place when the precise timing of an event has not been recorded for some individuals within the study. Instead, only incomplete data is available, for instance, when the event has not happened by a specific point in time or when the individual becomes unavailable for further observation. This partial information needs to be appropriately handled in survival analysis. Klein *et al.* [2003] provides a comprehensive treatment of censoring and its effect on survival analysis, offering methodologies for dealing with different types of censoring mechanisms.

The Kaplan-Meier estimator is a technique that is not parametric employed to approximate the survival function from censored data. It offers a step-wise approximation of the survival function by considering the recorded survival times and the quantity of individuals at risk at each time point. Kaplan and Meier [1958] introduced this estimator, which has become a widely utilized tool in survival analysis by reason of its simplicity and effectiveness, particularly when the fundamental distribution of survival times is unknown or non-parametric. In conclusion, survival analysis offers powerful tools for studying time-to-event data. The concepts of survival function, hazard function, censor-ing, and the Kaplan-Meier estimator form the foundation of this methodology, enabling researchers to gain insights into the underlying processes governing the occurrence of events over time.

2.2 Competing risk models

Consider K different event types, where $Z \in \{0, \ldots, +K\}$ represents the type of event that was observed(with $Z = 0$ typically indicating censoring). Various approaches have been proposed for defining CR survival models.

Initially, these models used the function cause-specific (CS) hazard to measure rate of occurrence at the instant for the k-th event type among individuals who have not yet experienced any event. The cause specific hazard function is defined as

$$h_{CS_k}(t) = \lim_{\delta t \to 0} \frac{\Pr\left(t \le T \le t + \delta t,\ Z = k \mid T > t\right)}{\delta t}. \qquad (2)$$

This function is the sum of all CS hazards represented as $h(t) = \sum_{k=1}^{K} h_{CS_k}(t)$. Else, competing risks (CR) survival models can be described using the CIF:

$$CIF_k(t) = \Pr(T \leq t, Z = k). \tag{3}$$

This function represents the likelihood of viewing the k-th event type before time t (and earlier than alternative events). It can also be expressed using the sub-distribution hazard function, commonly known as the Fine-Gray hazard Gray [1988]:

$$\begin{aligned} h_{FG_k}(t) &= -\frac{d\log(1 - CIF_k(t))}{dt} \\ &= \lim_{\delta t \to 0} \frac{\Pr\left(t < T \leq t + \delta t,\, Z = k \mid T > t \cup (T < t \cap Z = k)\right)}{\delta t}. \end{aligned} \tag{4}$$

This function quantifies the instantaneous proportion of the k-th phenomenon for individuals who have not yet experienced that phenomenon by time t, including those who have underwent a competing event. The main distinction intermediary to equations (2) with the addition of (4) lies in the risk set utilized to detail the probability.

Conclusively, latent failure time CR models infer $T = \min\{T_1, \ldots, T_K\}$ and $Z = \text{argmin}_k\{T_k\}$, where T_k is an event-specific time that remains covert saving $Z = k$. These models are commonly characterized by way of the joint survival function,

$$S_{T_1,\ldots,T_K}(t_1, \ldots, t_K) = Pr(T_1 > t_1, \ldots, T_K > t_K).$$

Nevertheless, the negligible distributions of the quiescent times T_k are non-identifiable saving presupposition we cant test (e.g., independence intermediary to T_ks Cox [1972]; Tsiatis [1975] or dependency coming about because of a copula Zheng and Klein [1995]) are imposed.

2.2.1 Cause Specific Model

In the cause-specific model for competing risks, each event type k has its own hazard function, denoted as $h_{CS_k}(t)$. This function quantifies the instantaneous rate at which individuals who have not yet encountered any event will experience that specific event type k at time t. Mathematically, it is defined as:

$$h_{CS_k}(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t, D = k | T \geq t)}{\Delta t}. \tag{5}$$

To evaluate the effect of covariates (risk factors) on this hazard, the Cox proportional hazards regression model is commonly applied. Let $\mathbf{X}$ denote a vector of covariates and $\boldsymbol{\beta}_k$ denote the vector of regression coefficients specific to cause k. The cause-specific

hazard regression model is given by:

$$h_{CS_k}(t|\mathbf{X}) = h_{0k}(t) \exp(\boldsymbol{\beta}_k^\top \mathbf{X}), \tag{6}$$

where $h_{0k}(t)$ represents the baseline cause-specific hazard function for event k (the hazard when all covariates are zero). The linear predictor $\boldsymbol{\beta}_k^\top \mathbf{X}$ acts multiplicatively on the baseline hazard.

The overall hazard function $h(t)$ for competing risks is the sum of the cause-specific hazard functions for all K event types:

$$h(t|\mathbf{X}) = \sum_{k=1}^{K} h_{CS_k}(t|\mathbf{X}) = \sum_{k=1}^{K} h_{0k}(t) \exp(\boldsymbol{\beta}_k^\top \mathbf{X}). \tag{7}$$

This approach allows us to analyze the risk of each specific event type independently, treating other competing events as censored. It is particularly useful when the interest lies in understanding the biological or mechanistic etiology of each event type.

2.2.2 Fine-Gray Model

For the Fine-Gray competing risks model, the focus shifts to the Cumulative Incidence Function (CIF), denoted as $CIF_k(t)$ **?**. This function symbolizes the likelihood of the k-th event type occurring prior to time t. The model is based on the subdistribution hazard function, $h_{FG_k}(t)$, which quantifies the instantaneous risk of the k-th event for individuals who have not yet experienced event k (including those who may have already experienced a competing event). It is defined as:

$$h_{FG_k}(t) = -\frac{d \log(1 - CIF_k(t))}{dt}. \tag{8}$$

Similar to the cause-specific approach, the Fine-Gray model links the subdistribution hazard to covariates using a proportional hazards formulation. The regression equation is defined as:

$$h_{FG_k}(t|\mathbf{X}) = h_{0k}^{FG}(t) \exp(\boldsymbol{\gamma}_k^\top \mathbf{X}), \tag{9}$$

where $h_{0k}^{FG}(t)$ is the baseline subdistribution hazard and $\boldsymbol{\gamma}_k$ is the vector of regression coefficients for cause k.

The key difference lies in the interpretation of the linear predictor. In the cause-specific model, $\exp(\boldsymbol{\beta}_k)$ describes the effect of a covariate on the instantaneous rate of the event in subjects who are currently event-free. In the Fine-Gray model, $\exp(\boldsymbol{\gamma}_k)$ describes the effect of the covariate directly on the cumulative incidence function (probability of the event occurring over time), accounting for the presence of competing risks.

2.3 Machine learning in general

2.3.1 RSF

Ishwaran *et al.* [2014] introduced random survival forests (RSF) as an extension of the well-established random forest algorithm, specifically designed for competing risks scenarios. The key objective of RSF is to approximate the cumulative incidence function (CIF) for competing risks. This algorithm shares the fundamental principles of random

forest: it generates B bootstrap samples from the dataset and a competing risk tree is constructed for each sample.

Similar to random forest, Random Survival Forest implements the random subspace mechanism by randomly selecting a subset of candidate variables ($M \leq p$) when dividing a node Breiman [2001]. The variable used for the split and its corresponding threshold are determined using Gray's log-rank splitting rule for competing risks. The tree is grown iteratively until each of the terminal node has more than a specified number of instances ($n > 0$).

Each survival tree produces an estimate of $CIF_k^b(t|x)$, where b represents the bootstrap samples. These estimates are then averaged to obtain the final estimated CIF. RSF offers the advantage of not relying on parametric assumptions, enabling autonomous modeling of nonlinearities and interaction effects among covariates, unlike traditional survival models.

2.4 Deeplearning

2.4.1 DeepHit

DeepHit (DH) stands as a many-tiered neural network(nn) designed specifically for survival analysis involving competing risks. Its architecture is designed to calculate the total incidence function over all occurrences [Lee *et al.* 2018]. Drawing from the principles of multi-task learning, DHT integrates a core subnetwork shared among all risk types along with K specialized cause-specific subnetworks. Shown in Figure 2.1 is the architectural built of DH for 2 risks [Lee *et al.* 2018].

The initial entry point for covariate values is a shared subnetwork comprising L_s fully connected layers. The vector $f_s(x)$ produced by this subnetwork captures a latent representation of the data that is shared by the K competing risks. Each cause-specific subnetwork then takes this latent representation $f_s(x)$ as input, alongside the original covariates passed through residual connections. This arrangement allows the cause-specific subnetworks to learn relationships particular to each risk among the covariates, while still utilizing the common latent representation of the covariates.

The DH output layer comes from stacking the cause-specific subnetwork output layers. Employing the softmax function, this output layer transforms the scores from the hidden layers into probabilities. In mathematical terms, the DeepHit output y can be expressed as,

$$y = [y_1, 1, \ldots, y_1, t_{\max}, \ldots, y_K, 1, \ldots, y_K, t_{\max}] \tag{10}$$

The approximated joint probability distribution for the K occurrences given x is matched by this output Lee *et al.* [2018],

$$
\begin{aligned}
&P(k = 1, t = 1 \mid x), \ldots, P(k = 1, t = t_{\max} \mid x), \\
&\ldots, \\
&P(k = K, t = 1 \mid x), \ldots, P(k = K, t = t_{\max} \mid x)
\end{aligned}
\tag{11}
$$

We can thus get, the CIF from using this estimated joint probability distribution Gabriel Blumenstock and Seow [2022],

$$
\begin{aligned}
\mathrm{CIF}_k(t \mid x) &= P(T \leq t, K = k \mid x) \\
&= \sum_{t'=1}^{t} P(T = t', K = k \mid x) \\
&= \sum_{t'=1}^{t} y_{k,t'}, \qquad k = 1, \ldots, K
\end{aligned}
\tag{12}
$$

To optimise connection weights, DeepHit minimises a loss function adapted to estimate the cumulative incidence function using right-censored data. This loss function comprises two components along with a weighting term β,

$$
L = L_1 + \beta L_2 \tag{13}
$$

where

$$
L_1 = \sum_{i=1}^{N} \mathbb{1}_{k_i=0} \log y_{i,k_i} + \mathbb{1}_{k_i=0} \log \left(1 - \sum_{k=1}^{K} F_k(t_i \mid x_i)\right) \tag{14}
$$

$$
L_2 = \sum_{k=1}^{K} \sum_{i \neq j} A_{k,i,j} \exp\left(-\beta \left[F_k(t_i \mid x_i) - F_k(t_i \mid x_j)\right]\right) \tag{15}
$$

In L_1, the first term is a log-likelihood loss, penalising deviations of the predicted probabilities for censored instances from the ideal estimate. The second term in L_1 penalizes deviations of the predicted cumulative survival probabilities for uncensored instances.

L_2 is a ranking loss function that relies on the notion of concordance, assessing the difference between estimated accumulated probabilities in various time instances for comparable pairs of instances, as noted by Gabriel Blumenstock and Seow [2022].

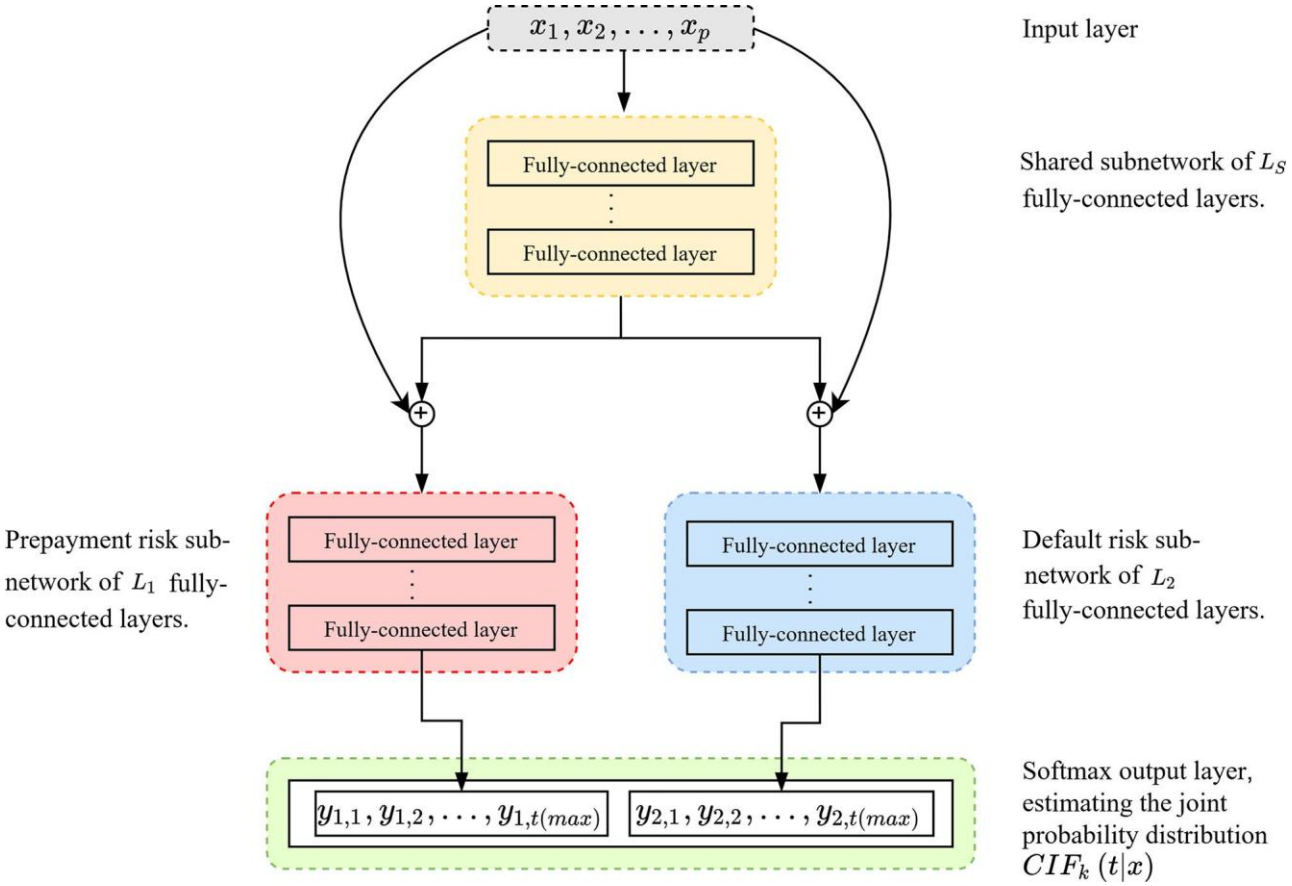


Figure 2.1: The architecture of deep hit with 2 risks as adapted from Gabriel Blumenstock and Seow [2022]

2.5 Evaluation techniques

2.5.1 C-Index

The Concordance Index (C-index) is commonly used standard of measurement for evaluating the discriminatory power of survival models. It measures consistency between predicted risks and observed outcomes, indicating how well a model ranks people according to their survival times Harrell Jr *et al.* [1982]. A value of 1 on the C-index denotes perfect concordance, meaning that all pairings of individuals are correctly ranked by the model based on their survival times. The C-index goes from 0 to 1, a value of 0.5 suggests random performance (i.e., no discrimination).

Mathematically, the C-index can be defined as follows: Let T_i be the true survival time of individual i, and let $\hat{T}_i$ be the predicted survival time. Additionally, let δ_i denote the event indicator, where $\delta_i = 1$ if an event occurred for individual i and $\delta_i = 0$ otherwise Uno *et al.* [2011]; Harrell Jr *et al.* [1982]. The C-index is calculated as the proportion of all pairs of individuals i and j where the forecasted survival time of individual i is smaller than the forecasted survival time of person j, given that both individuals have different true survival times,

$$C = \frac{\sum_{i<j} \mathrm{I}\ \hat{T}_i < \hat{T}_j\ \mathrm{I}(T_i < T_j) + \frac{1}{2} \sum_{i=j} \mathrm{I}\ \hat{T}_i = \hat{T}_j\ \mathrm{I}(T_i \leq T_j)}{\sum_{i<j} \mathrm{I}(T_i < T_j)} \tag{16}$$

where $I(\cdot)$ is the indicator function.

This equation accounts for censored data and ensures that only pairs with different true survival times contribute to the calculation.

2.5.2 Cumulative/Dynamic C-index:

This is a time-dependent variation of the C-index, often called the Cumulative/Dynamic (C/D) concordance index. It evaluates the model's ability to discriminate risk at a specific time horizon t by assessing whether subjects who experience the event before time t have higher predicted risk at t than those who do not. This is crucial for models where predictive performance at specific clinical decision points is of interest. Mathematically, for a given time t, it can be defined as:

$$C(t) = \frac{\sum_{i=j} \Delta_i \cdot I(T_i \leq t, T_j > t) \cdot I(\hat{\pi}_i(t) > \hat{\pi}_j(t))}{\sum_{i=j} \Delta_i \cdot I(T_i \leq t, T_j > t)},$$

where T_i and T_j are the observed survival times of individuals i and j, Δ_i is the event indicator (1 if event observed, 0 if censored), $\hat{\pi}_i(t)$ and $\hat{\pi}_j(t)$ are the predicted risk scores (or 1 minus survival probability) at time t, and $I(\cdot)$ is the indicator function. The denominator counts all usable comparable pairs at time t, while the numerator counts those pairs where the model's risk predictions are correctly ordered.

2.5.3 Time-dependent Area Under the Receiver Operating Characteristic curve (ROC)

The Time-dependent Area Under the Receiver Operating Characteristic curve (AUC-ROC) is a metric used to assess the discriminatory ability of survival models over time. While traditional AUC-ROC evaluates the model's discrimination ability at a single time point, the time-dependent AUC-ROC evaluates discrimination across various time points, providing insight into how the model's discrimination changes over time.

Mathematically, the time-dependent AUC-ROC is computed using the area beneath the ROC curve at different time points Heagerty *et al.* [2000]. If we let the time of interest be T, and The anticipated survival function is $\hat{S}(t)$. Plotted on the ROC curve are the sensitivity against the false positive rate (1 - specificity) at different threshold settings. The integral under this ROC curve is then used to get the AUC-ROC.

$$\text{AUC–ROC}(t) = \int_0^1 \text{Sensitivity}_{(t)}\, d(1 - \text{Specificity}(t)) \tag{17}$$

where Sensitivity(t) and Specificity(t) represent the sensitivity and specificity of the model at time t, respectively. The time-dependent AUC-ROC thus provides a comprehensive assessment of the model's discrimination ability over time, allowing it to measure how the model performs at different stages of follow-up.

2.5.4 Brier Score:

The precision of probabilistic forecasts is gauged by the Brier score. The mean squared difference between the actual results and the expected probability is used to calculate it. For survival models, a time-dependent version of the Brier score can be used to evaluate performance at different time points, reflecting the model's overall calibration and accuracy. The accuracy of probabilistic predictions is defined as,

$$BS = \frac{1}{N} \sum_{t=1}^{N} (f_t - o_t)^2,$$

where f_t is the predicted probability of an event occurring at time t, o_t is the actual outcome (1 if the event occurs, 0 otherwise) and N is the total number of predictions. A lower Brier Score indicates better overall predictive performance, reflecting a combination of calibration and prediction refinement.

2.5.5 Integrated Brier Score (IBS):

The IBS is an aggregate measure of the Brier score over a range of time points. It provides a single summary statistic that reflects the total effectiveness of the survival model across the entire follow-up period. This metric is helpful for models like the Fine-Gray model, where long-term prediction accuracy is essential.

These evaluation metrics help in comprehensively assessing the performance of different survival models, considering both their ability to rank individual survival times correctly and their accuracy in predicting event probabilities over time. By using a combination of these metrics, researchers and clinicians can better understand the strengths and limitations of each model in the context of their specific application. Mathematically, it is defined as,

$$IBS = \int_{t_1}^{t_{max}} BS_c(t)\, dw(t),$$

where $BS_c(t)$ is the time-dependent Brier Score at time t, $w(t)$ is a weighting function, often defined as $w(t) = t/t_{max}$, t_1 and t_{max} represent the lower and upper bounds of the time interval. The integral is typically estimated using numerical methods such as the trapezoidal rule. This metric is particularly useful for models like the Fine-Gray model, where long-term prediction accuracy is essential.

Chapter 3

Methodology

Here, the methodological approach is presented. The chapter describes how customer churn and multimorbidity data are simulated, specifying covariates, event types, and censoring mechanisms. It clarifies the architectures, hyperparameters, and estimation procedures for the four models: Cause-Specific Cox, Random Survival Forest, DeepHit, and the Fine–Gray model. Steps for model validation and performance metrics are included.

3.1 Data

This study uses simulated data sets. While constant hazard models are the easiest and therefore the best understood, they will be unrealistic in many situations of applications, thus this study will propose application-dependent hazard models for events simulations. We will simulate a customer churn situation and a multimorbidity situation, to highlight the dangers of ignoring competing risks in diverse fields.

3.1.1 Simulating Competing Risk Data for Customer Churn

Customer churn refers to the percentage of customers who stop using a company's product or service during a specific time frame. It's a crucial metric for businesses to evaluate their customer retention. To calculate churn rate, divide the number of lost customers during a period (e.g., a quarter) by the initial number of customers. Ideally, companies aim for a churn rate close to 0. Reducing churn can significantly impact profitability, as returning customers tend to spend more on products and services.

To simulate competing risk data for customer churn, we will employ a simulation approach that incorporates multiple risk factors that influence churn behavior. Instead of using a constant hazard model, We will utilize hazard functions to model the time-varying risk of customer churn and identify critical drop-off points in the customer lifecycle. The simulation will include variables such as age, salary, marital status, number of dependents, and additional covariates to show the nature of churn behavior.

Simulation Setup

We simulate a cohort of customers and track their churn events over a specified period. Each customer is characterised by their demographic and socioeconomic attributes, as well as other covariates that influence churn behaviour.

Demographic and socioeconomic attributes are generated as follows. Age (A) is simulated from a normal distribution with a mean and standard deviation reflecting the target population. Salary (S) is drawn from a distribution that aligns with the income distribution in South Africa. Marital status (M) is generated as a binary variable, where 0 represents unmarried and 1 represents married, using a binomial distribution. The number of dependents (D) is simulated from a distribution reflecting typical family sizes in the region.

To simulate churn events, we define two competing risks: voluntary churn (such as switching to a competitor) and involuntary churn (such as leaving due to financial difficulties). Hazard rates are assigned to each competing risk factor, allowing them to vary over time. We use a time-dependent Cox proportional hazards model to simulate these hazard rates for each risk factor. Specifically, the hazard rate at time t for each competing risk j is given by:

$$h_j(t|X) = h_{0j}(t) \exp(\beta_j^T X)$$

where $h_j(t|X)$ is the hazard rate for risk j at time t given covariates X, $h_{0j}(t)$ is the baseline hazard for risk j at time t, and β_j is the vector of coefficients for covariates X for risk j.

We model the competing risks in two opposing scenarios. In the first scenario, a weight at infinity is assigned to one of the competing risks, indicating an infinitely high hazard rate for that risk. This represents a situation where certain types of churn events are highly probable compared to others, expressed as $h_1(t|X) = \infty$. In the second scenario, we simulate the other competing risk without a weight at infinity, reflecting a more balanced hazard rate with the equation $h_2(t|X) = h_{02}(t) \exp(\beta_2^T X)$.

3.1.2 Simulating Multimorbidity Situations in Health Science

Unlike constant hazard models, which are widely used, our simulation incorporates time-dependent and non-linear cause-specific hazards to capture the changing nature of morbidity patterns. The simulation includes variables such as age, marital status, area of residence (urban or rural), number of dependents, and behavioural covariates such as smoking and alcohol use. These variables are included to account for the various factors influencing morbidity.

For demographic and socioeconomic attributes, age (A) is simulated from a distribution reflecting the age structure of the target population. Marital status (M) is represented as a binary variable, with 0 for unmarried and 1 for married, based on population statistics. Area of residence (R) is categorised as urban or rural, aligned with population distribution data. The number of dependents (D) is generated to reflect typical family sizes. Additional covariates such as smoking status, BMI, socioeconomic status,

and access to healthcare services are also simulated, using their respective distributions and known effects on morbidity.

Data Simulation for Competing Risks of Customer Churn

We outline how we simulate a dataset of $N = 1000$ observations for a competing risks problem with two types of customer churn: voluntary churn (such as switching to another provider) and involuntary churn (such as financial difficulties). We label these risks as Risk 1 (voluntary) and Risk 2 (involuntary). The process includes the generation of demographic and socioeconomic variables, simulation of event times, and handling of censoring.

Generating Demographic and Socioeconomic Attributes

We first simulate a set of customer-level variables that may influence churn hazards. Age (A) is assumed to follow a Normal distribution with mean $\mu_A = 35$ and standard deviation $\sigma_A = 10$, i.e., $A \sim \mathsf{N}(\mu_A, \sigma_A^2)$. Salary ($S$) follows a Lognormal distribution, $S \sim \text{Lognormal}(\mu_{\ln S}, \sigma_{\ln S}^2)$, where $\mu_{\ln S} = 10$ and $\sigma_{\ln S} = 0.5$ to provide a realistic skewed income distribution. Marital status (M) is a binary variable (0 for unmarried, 1 for married) simulated from a Binomial distribution with $p_M = 0.5$, giving a 50% chance of being married. The number of dependents (D) is modelled as a count variable from a Poisson distribution with $\lambda_D = 1.2$, capturing typical small family sizes.

We also include two additional covariates. Region (R) is a categorical variable with three categories, $\{A, B, C\}$, assigned probabilities $\{0.4, 0.4, 0.2\}$ respectively. Tenure (T_e), defined as the number of months the customer has stayed with the company, is simulated from a Normal distribution with mean 36 and standard deviation 12, with negative values censored at zero.

Competing Risks Framework

There are two churn risks: voluntary churn (Risk 1) and involuntary churn (Risk 2). The sample is split into two scenarios to demonstrate different hazard behaviours. In Scenario 1, a weight at infinity is assigned for Risk 1, making its hazard extremely large and causing it to occur almost instantly for most customers. In Scenario 2, both risks are assigned normal hazard rates, allowing either to occur in a more balanced fashion. Half the observations are randomly assigned to Scenario 1 and the other half to Scenario 2.

Event Time Simulation

For each risk $j \in \{1, 2\}$, we assume an exponential baseline hazard with a log-linear covariate effect. The hazard function at time t for risk j is

$$h_j(t \mid \mathbf{X}) = \lambda_j(t) \exp\left(\boldsymbol{\beta}_j^\top \mathbf{X}\right),$$

where $\lambda_j(t)$ is the baseline rate, and the linear predictor is given explicitly as:

$$\boldsymbol{\beta}_j^\top \mathbf{X} = \beta_{j0} + \beta_{j1} A + \beta_{j2} S_m + \beta_{j3} S_f + \beta_{j4} V + \beta_{j5} E + \beta_{j6} M,$$

with A denoting age, S_m and S_f smoking indicators (male/female), V vaccination status, E environmental exposure level, and M biomarker measurement, in this example.

Assuming a constant baseline hazard, the survival time T_j for risk j follows an exponential distribution with rate parameter $\theta_j = \lambda_j \exp(\boldsymbol{\beta}_j^\top \mathbf{X})$, i.e., $T_j \sim \text{Exponential}(\theta_j)$. In Scenario 1, the "weight at infinity" for Risk 1 corresponds to θ_1 being extremely large, making T_1 effectively zero; this is approximated by sampling from an exponential distribution with mean 10^{-6}.

Observed Time and Event Type

For each individual, T_1 and T_2 are generated from their respective exponential distributions, using the scenario-dependent rate for T_1 if assigned to Scenario 1. A censoring time C is generated uniformly between 0 and 100 months. The observed time for each individual is $T = \min(T_1, T_2, C)$. The event type E is determined by comparing these times: if T_1 is the earliest and less than C, $E = 1$; if T_2 is the earliest and less than C, $E = 2$; otherwise, the individual is censored ($E = 0$).

Simulation of Multimorbidity Data

We describe the procedure for simulating multimorbidity data with two competing causes: cardiovascular disease (CVD) and diabetes. The simulation incorporates demographic, socioeconomic, and behavioural factors with a time-dependent hazard structure, using a piecewise exponential approach to approximate a flexible parametric survival model.

Demographic and Socioeconomic Covariates

A total of $N = 1000$ individuals are generated, each with a set of attributes. Age (A) is drawn from a Normal distribution with mean 50 and standard deviation 15, censored below zero if negative. Marital status (M) is binary (0 = unmarried, 1 = married) and follows a Bernoulli distribution with $p_M = 0.6$. Area of residence (R) is binary (0 = rural, 1 = urban) with $p_R = 0.55$. Number of dependents (D) is drawn from a Poisson distribution with $\lambda_D = 1.5$. Smoking status (S_m) and drinking status (S_d) are each binary, with probabilities $p_s = 0.3$ and $p_d = 0.4$ respectively. Body Mass Index (BMI) is drawn from a Normal distribution with mean 27 and standard deviation 4, truncated below 15. Socioeconomic status (SES) is ordinal, taking values from 1 to 5 with probabilities 0.1, 0.2, 0.4, 0.2, and 0.1, respectively. Healthcare access (H) is binary, with probability 0.7 for good access.

Competing Morbidities

We focus on two morbidity causes: cardiovascular disease (CVD) and diabetes. For each cause j, the time-to-event T_j is simulated from a piecewise exponential model.

Piecewise Time-Dependent Hazard

To approximate a flexible parametric model, time is divided into intervals $[0, 5)$, $[5, 10)$, and $[10, \infty)$. Each cause j has interval-specific baseline hazards $(\lambda_{j1}(t), \lambda_{j2}(t), \lambda_{j3}(t))$, and the log-hazard at time t is given by $\log(\lambda_{jk}(t)) + \beta_j^T X$, where k denotes the interval. The hazard function in interval k is $h_j(t \mid X) = \lambda_{jk}(t) \exp(\beta_j^T X)$. For each interval, a waiting time is drawn from an exponential distribution with the given rate. If the waiting time is less than the interval length, this is the event time; otherwise, the process repeats in the next interval, subtracting the elapsed interval length from the remaining follow-up time.

Event Times and Censoring

For each individual and each cause, a linear predictor $\eta_{ij} = \beta_{j0} + \beta_{j1} A_i + \beta_{j2} S_{m,i} + \cdots$ is calculated, and T_{ij} is sampled using the piecewise exponential method. The maximum follow-up time is set to 20 years. The observed event time for an individual is $T = \min(T_{i1}, T_{i2}, C)$, where C is a censoring time drawn from a uniform distribution between 0 and 20 years. The event type E is assigned according to which event time is earliest, or is set to 0 if censored.

Final Dataset

The final simulated dataset contains all individual covariates, the observed time to event, and the event type (0 = censored, 1 = CVD, 2 = Diabetes), the disease will be referred to as disease A and disease B respectively. Each individual is followed up for a maximum of 20 years, ensuring a complete and realistic competing risks dataset for subsequent analysis.

3.2 Model Fitting

In this section, we describe our methodology for simulating data and fitting four separate models for competing-risks survival analysis: the Cause-Specific Cox model, Random Survival Forest (RSF), DeepHit, and the Fine–Gray model. We present two parallel case studies; Customer Churn and Multimorbidity, each with distinct data structures and simulation processes. We then provide details of how each model was specified and fit, noting which models were implemented in Python and which were fit using R.

3.2.1 Customer Churn

Data Structures

To explore the phenomenon of customer churn with two competing risks (voluntary and involuntary), we generated synthetic datasets of varying sizes (1000, 2000, and 3000 observations). Each simulated record reflects an individual customer's profile and outcomes, including:

Covariates such as Age, Salary, Region, Tenure, and ProductUsage (or similar features, depending on scenario). Competing-risk outcomes, capturing either Voluntary Churn (Risk 1) or Involuntary Churn (Risk 2).Censoring times ensuring that not all customers churn during the observation window. A fraction of instances remain censored at the end of simulation to avoid degenerate survival functions.

We used Python to produce these synthetic data. For each observation, two potential event times (corresponding to the two churn risks) were sampled from exponential or related distributions, influenced by linear predictors in the covariates. The observed time for each individual was set to the minimum of these two risk times and an independent censoring time. The event status was labeled *0* (censored), *1* (voluntary churn), or *2* (involuntary churn).

To investigate the influence of data size, we repeated the simulation for datasets containing 1000, 2000, and 3000 observations, keeping the same parameterization but drawing new random seeds.

Models

Three of the four survival models were fit in Python. The first was the Cause-Specific Cox Model (CSC), where separate cause-specific hazards were estimated for voluntary and involuntary churn. For each model, individuals who experienced the alternative event or were censored were treated as non-events (i.e., censored) for that particular cause. Model fitting for CSC was performed using the sksurv library.

The second model was the Random Survival Forest (RSF), also implemented in Python using the RandomSurvivalForest functionality from sksurv. RSF is a nonparametric ensemble approach where each tree is trained on a bootstrap sample of the data. Nodes are split using a log-rank splitting rule adapted for competing risks, and the resulting model produces cumulative incidence estimates for each event type.

DeepHit, the third model, was implemented using the pycox library. DeepHit is a deep neural network designed for survival analysis with competing risks. In our setup, the network typically comprised two or more hidden layers with 32 units per layer. The architecture was trained to learn a discrete-time representation of the cumulative hazard, using a loss function that combines partial log-likelihood and a ranking component suitable for right-censored and competing-risk data.

The fourth model, the Fine-Gray subdistribution hazards model, was fit in R using the cmprsk package. This approach directly models the cumulative incidence function of

each risk, accounting for the fact that the occurrence of one event type precludes the occurrence of others.

For each dataset size, all models were evaluated using five-fold cross-validation. Each model was refit on the training folds and evaluated on held-out test folds. The main performance metrics assessed were the weighted Brier score, time-dependent area under the curve (AUC), and the C-index.

3.2.2 Multimorbidity

Data Structures

For the second case study, we simulated a multimorbidity scenario in which individuals face two mutually exclusive disease diagnoses (Risk 1 and Risk 2). Again, the sample sizes of 1000, 2000, and 3000 allowed us to observe how model performance changes with more data. Each synthetic individual is characterized by clinical or demographic factors such as Age, BloodPressure, Cholesterol, and Smoking Status. We assigned event times for disease A and B using exponential (or gamma) distributions whose rates depended on these covariates.

We introduced random censoring times, ensuring a proportion of participants remained event-free and censored through the observation period. As with the churn setting, the presence of two risks means that if one disease occurs first, the other cannot occur, yielding a standard competing-risks structure.

Models

The same four survival approaches were employed for the multimorbidity simulation. We used Python for the Cause-Specific Cox model, Random Survival Forest, and DeepHit. In R, we fit the Fine-Gray subdistribution hazards model. The model specifications mirror those used for the churn scenario, except that the features (blood pressure and cholesterol) and hazard parameters were adjusted to reflect typical clinical contexts (e.g., older age correlating with higher baseline hazard rates). We again used cross-validation to compare model performance via Brier scores, C-index, and AUC at one or more evaluation times.

All model hyperparameters, such as the tree count in RSF, layer sizes in DeepHit, or the penalty structure in Fine-Gray, remained consistent across data sizes for fairness. Changes in sample size thus reveal the robustness of each model to smaller or larger training sets. Further, by investigating two distinct domains (customer churn vs. multimorbidity), we evaluate how these methods generalize across fields with different covariates and hazard relationships.

The code used in this research can be obtained in this link https://github.com/mafanedza/msc_escience_research_code

Chapter 4

Results and Discussion

This chapter presents empirical findings from the simulation studies. Performance metrics (C-Index, Brier scores, and ROC AUC) are shown across different dataset sizes and event types. The discussion interprets patterns in calibration and discrimination, comparing results for the churn versus multimorbidity settings. Literature links are provided to confirm or challenge previous insights on competing-risks modelling and machine-learning capabilities.

4.1 On Dangers of Ignoring Competing Risks

In the presence of multiple potential events, such as additional causes of failure or mortality, treating all non-primary outcomes simply as censored observations can distort the true probability of experiencing the event of interest. Figures 4.1 and 4.2 illustrate this issue by comparing Kaplan-Meier (KM) estimates (solid lines) with Cumulative Incidence Function (CIF) estimates (dashed lines) for two scenarios: customer churn and multimorbidity. The KM approach implicitly assumes that individuals who undergo a competing event remain at risk for the primary outcome, thereby inflating the apparent risk over time.

In the customer churn scenario (Figure 4.1), the KM curves lie above the CIF curves for each dataset size (1000, 2000, and 3000) however there is a negligible presence of competing risk in this instance. The survival curves show that ignoring the competing event leads to an overestimation of the churn incidence. If a competing event such as switching to an entirely different service or experiencing a disqualifying event occurs, then those individuals should no longer contribute to the risk set for churn. By not accounting for such events, the KM method effectively treats them as though they could still churn in the future, thereby biasing the results upwards.

A similar pattern emerges in the multimorbidity scenario (Figure 4.2), where individuals may experience a second, more severe comorbid condition that precludes the primary outcome of interest. The KM estimates again overstate the probability of experiencing that primary event because they disregard the removal of participants who encounter a competing event. In a clinical setting, this overestimation can misinform

health practitioners, leading to erroneous conclusions about disease incidence and potentially skewing resource allocation or intervention strategies.

By accurately accounting for the fact that individuals who experience one event are no longer at risk for another, analysts and practitioners obtain more reliable estimates of incidence. This distinction is particularly crucial in multimorbidity research, where the presence of multiple chronic conditions often leads to competing clinical outcomes, and in domains like customer churn, where diverse events can remove individuals from the risk set in ways that naive KM analysis fails to capture.

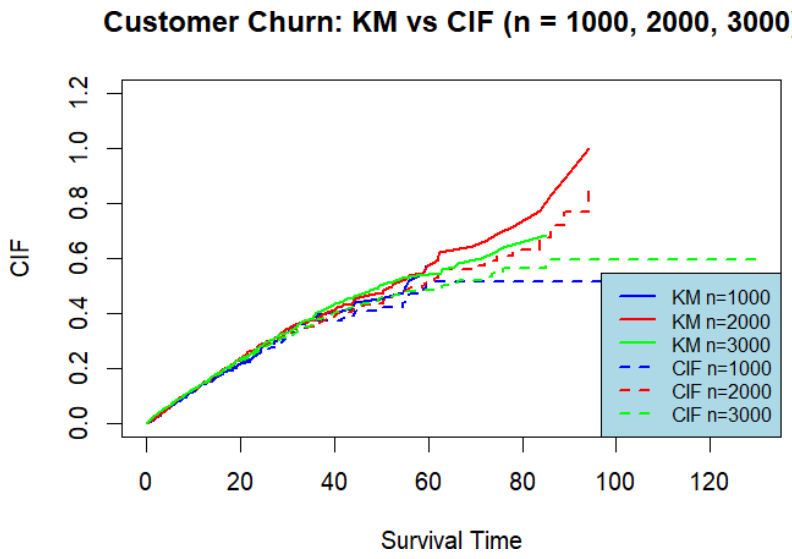


Figure 4.1: Customer Churn: KM vs CIF for dataset sizes 1000, 2000, and 3000.

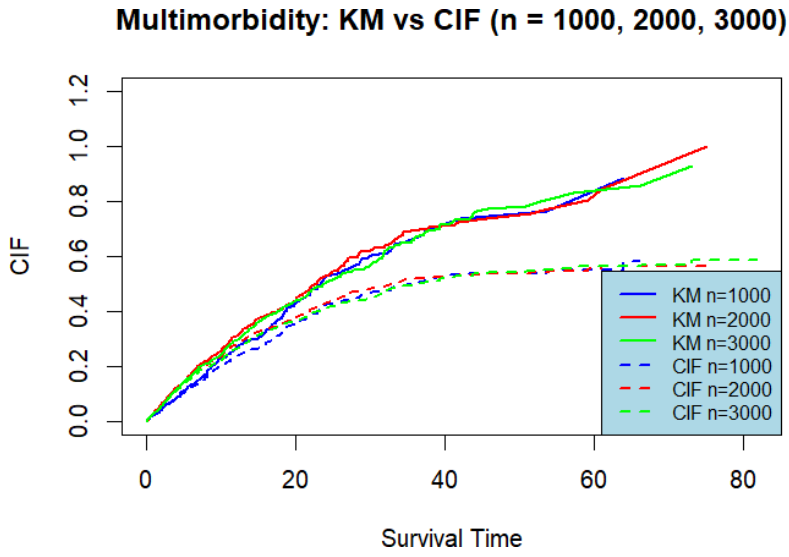


Figure 4.2: Multimorbidity: KM vs CIF for dataset sizes 1000, 2000, and 3000.

4.2 Customer Churn Results

In this section we present results from the customer churn simulation of the 3 datasets and four models.

4.2.1 Cross validation results

This section will show the results from the training and testing process using 5 fold cross validation on the customer churn datasets.

N= 1000

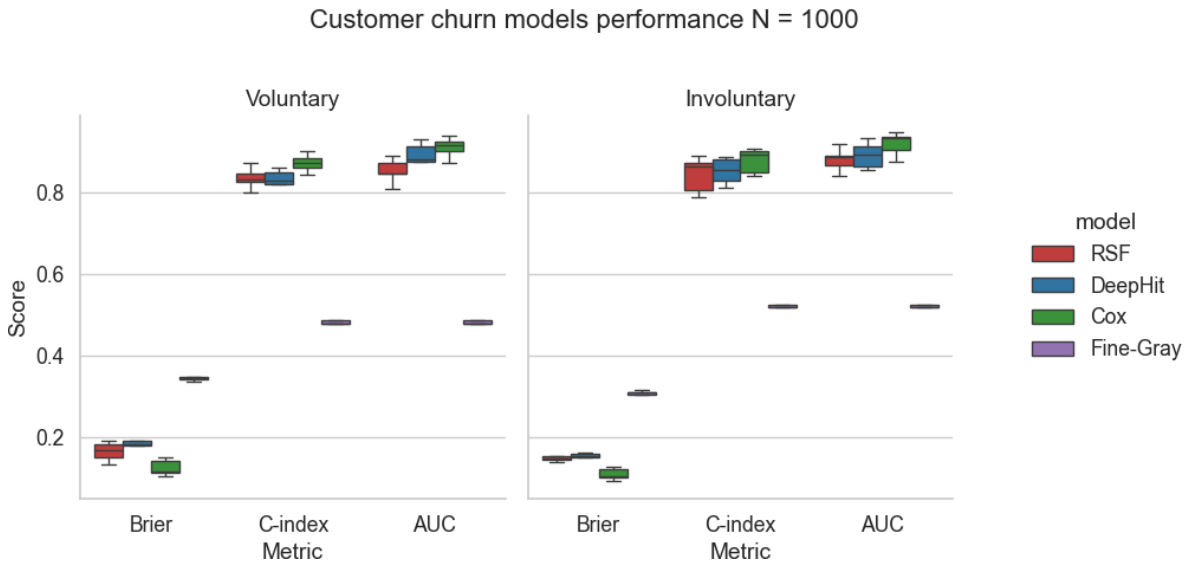


Figure 4.3: Model performance comparison for customer churn simulation ($N = 1000$), evaluating the Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray models on Voluntary and Involuntary churn using Brier score, C-index, and ROC AUC.

Figure 4.3 presents the comparative performance of four competing-risk survival models (Random Survival Forest, DeepHit, Cox, and Fine–Gray) on simulated customer churn data of size $N = 1000$. The evaluation metrics include the Brier score (where lower values indicate better calibration), C-index, and ROC AUC (with higher values indicating superior discrimination).

For voluntary churn, the Cox model exhibits notably strong performance, achieving the lowest Brier scores (approximately 0.1) and the highest values in terms of both C-index and ROC AUC (above 0.8), demonstrating superior calibration and discriminative capability relative to the other models. DeepHit and RSF follow closely behind Cox, displaying similar discriminative performance (C-index and ROC AUC close to 0.8), but slightly higher Brier scores, suggesting somewhat less accurate calibration. The Fine-Gray model demonstrates moderate performance, with significantly lower C-index and ROC AUC values around 0.6, indicating a weaker discriminative ability compared to the

other models. Its calibration, as evidenced by higher Brier scores (around 0.2), is also less precise.

For involuntary churn, a similar performance hierarchy is observed. Cox remains the most robust model, with consistently high discrimination metrics (C-index and ROC AUC around or slightly above 0.8), along with the lowest Brier scores among the models (around 0.1). RSF and DeepHit models yield comparable performances, with C-index and ROC AUC values just below Cox's metrics, and moderate Brier scores. The Fine-Gray model again lags behind, indicating limited discriminative capacity (C-index and ROC AUC values around 0.6) and poorer calibration, evident from the elevated Brier scores.

N = 2000

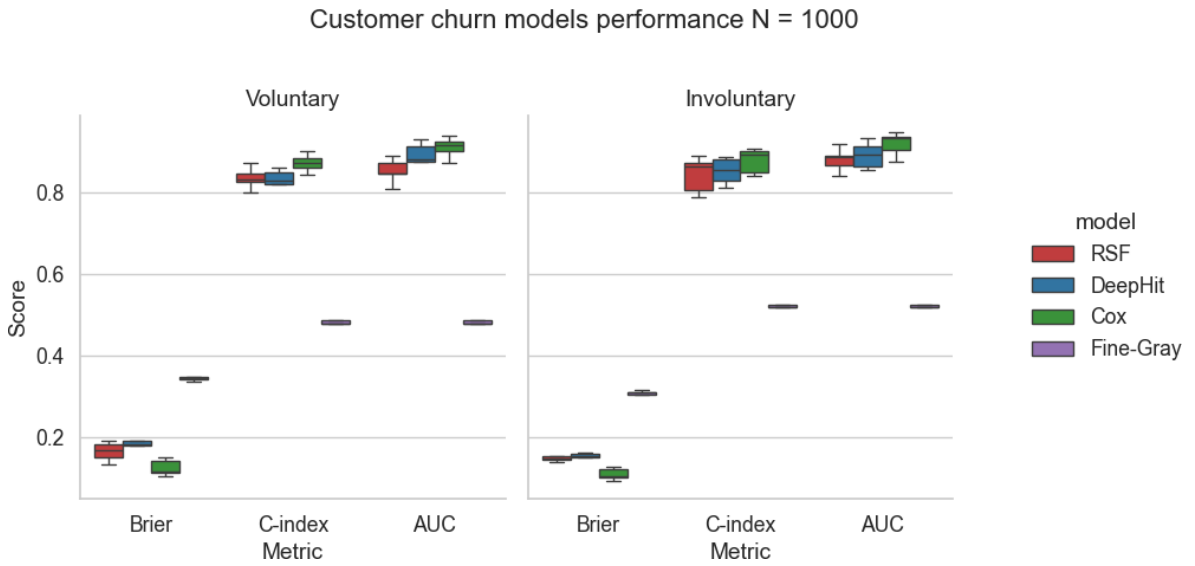


Figure 4.4: Model performance comparison for customer churn simulation ($N = 2000$), evaluating the Random Survival Forest (RSF), DeepHit, Cox, and Fine–Gray models on Voluntary and Involuntary churn using Brier score, C-index, and ROC AUC.

Figure 4.4 illustrates the performance of four competing-risks survival models Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray on customer churn simulation data with a sample size of $N = 2000$.

In the case of voluntary churn, the Cox model consistently outperforms the other methods, achieving the highest C-index and ROC AUC scores (both above 0.8), along with the lowest Brier scores (around 0.1), indicating superior predictive precision and calibration. The RSF and DeepHit models show slightly lower but comparable discrimination performance (with C-index and ROC AUC close to 0.8). However, their calibration performance is modestly weaker, as evidenced by their marginally higher Brier scores compared to Cox. The Fine-Gray model clearly trails behind the other models, demonstrating a lower discriminative ability with C-index and ROC AUC values around 0.6, and considerably higher Brier scores (around 0.2), reflecting poorer calibration.

For involuntary churn, the Cox model maintains its position as the top performer, exhibiting consistently strong discriminative capabilities (C-index and ROC AUC above 0.8) and superior calibration (Brier scores around 0.1). RSF and DeepHit display competitive performance with respect to discrimination, closely following the Cox model, but still perform slightly worse in calibration as indicated by marginally higher Brier scores. Fine–Gray again shows the weakest overall performance, with significantly lower discrimination (C-index and ROC AUC around 0.6) and noticeably poorer calibration reflected in elevated Brier scores.

These results reaffirm that, for a dataset size of $N = 2000$, the Cox proportional hazards model remains the most reliable and effective method for modelling both voluntary and involuntary churn. RSF and DeepHit provide viable alternatives, though without significant improvement over Cox. The Fine-Gray approach continues to show limitations under these synthetic churn conditions, suggesting careful consideration when deploying it in similar practical scenarios.

N = 3000

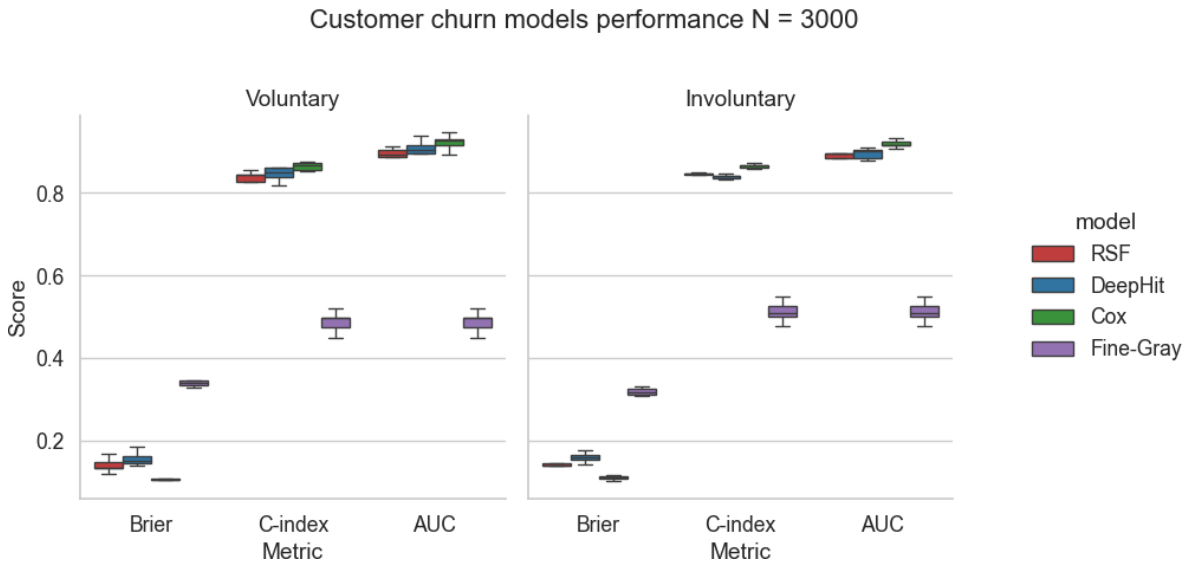


Figure 4.5: Model performance comparison for customer churn simulation ($N = 3000$), evaluating Random Survival Forest (RSF), DeepHit, Cox, and Fine–Gray models on Voluntary and Involuntary churn using Brier score, C-index, and ROC AUC.

Figure 4.5 summarises the comparative performance of four competing-risk survival models Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray applied to customer churn data simulated with $N = 3000$.

For voluntary churn, the Cox proportional hazards model continues to demonstrate superior performance, consistently achieving the highest C-index and ROC AUC scores (both exceeding 0.8), alongside significantly lower Brier scores (around 0.1). This result indicates excellent predictive accuracy, calibration, and discriminative capability.

The RSF and DeepHit models perform comparably and close to Cox, displaying similar discrimination capabilities (C-index and ROC AUC values around 0.8), though with slightly less accurate calibration as reflected in moderately higher Brier scores. Again, the Fine-Gray model notably underperforms in comparison, exhibiting much lower discriminative ability with C-index and ROC AUC scores approximately at 0.6, and higher calibration error (Brier scores around 0.2).

About involuntary churn, the Cox model clearly maintains its superior standing, achieving consistently higher discriminative metrics (C-index and ROC AUC around 0.8) and superior calibration with the lowest Brier scores (approximately 0.1). RSF and DeepHit models again display competitive discriminative performance, closely aligned with Cox's results, but with slightly weaker calibration indicated by marginally higher Brier scores. Fine-Gray remains the weakest model overall, with significantly lower discrimination power (C-index and ROC AUC near 0.6) and poorer calibration, evidenced by elevated Brier scores.

These results reinforce earlier findings and clearly identify the Cox proportional hazards model as the most robust and effective method for modelling competing churn risks, even at larger sample sizes ($N = 3000$). The RSF and DeepHit models offer viable alternatives without providing significant performance improvement over Cox. The Fine-Gray model consistently underperforms, highlighting its limitations under the considered simulation scenario, particularly in contexts that demand high discriminative performance and calibration accuracy.

4.3 Multimobidity Results

In this section we present results from the multimorbidity simulation of the 3 datasets and four models.

4.3.1 Cross Validation Results

This section will show the results from the various folds during the training and testing process using 5 fold cross validation on the multimobidity datasets.

N = 1000

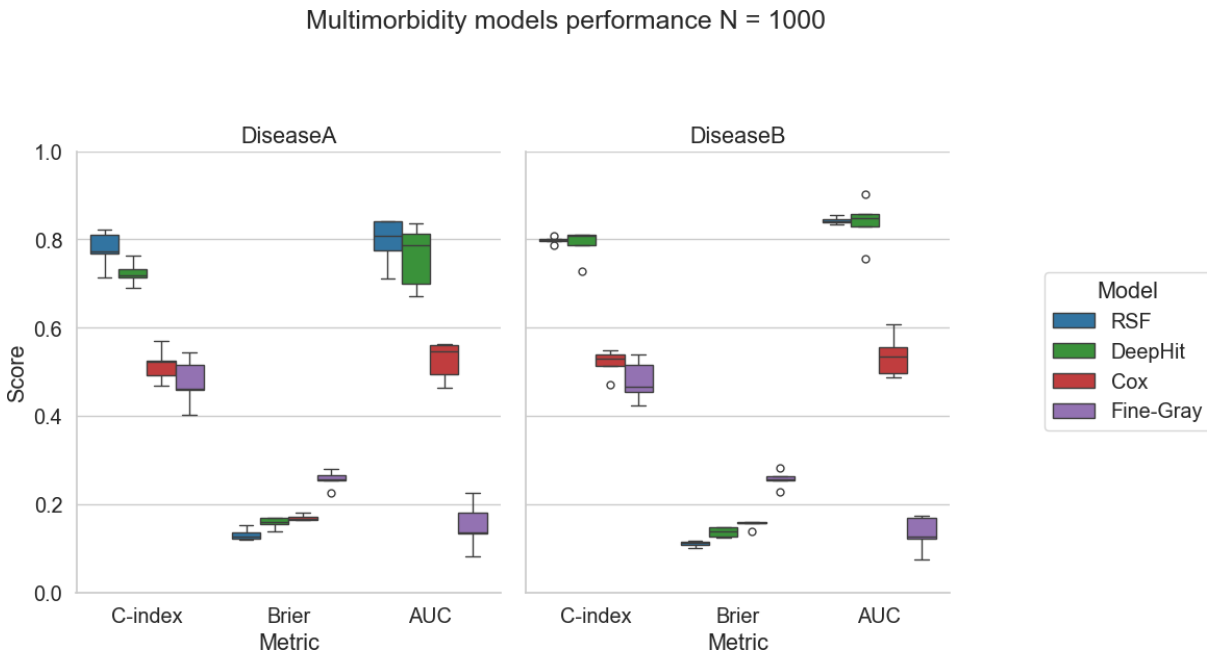


Figure 4.6: Model performance comparison for multimorbidity simulation ($N = 1000$), evaluating Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray models on Disease A and Disease B using C-index, Brier score, and ROC AUC.

Figure 4.6 depicts the performance of four competing-risks survival models Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray in the multimorbidity simulation scenario with a dataset size of $N = 1000$.

For Disease A, RSF and DeepHit display the strongest overall discriminative capabilities, with both achieving high C-index and ROC AUC values around or above 0.8. The Cox model exhibits moderate performance with somewhat lower discriminative ability (C-index and ROC AUC approximately 0.6). In terms of calibration, RSF and DeepHit again outperform Cox, showing lower Brier scores (approximately 0.1), indicative of better predictive calibration. The Fine-Gray model performs modestly, displaying intermediate discrimination (C-index and ROC AUC close to Cox's performance, around 0.6), and notably higher Brier scores (around 0.2), reflecting weaker calibration.

For Disease B, RSF again demonstrates superior performance, obtaining the highest discrimination scores (C-index and ROC AUC around or exceeding 0.8). DeepHit closely follows RSF in discriminative performance, while Cox and Fine-Gray notably lag behind, achieving C-index and ROC AUC values close to 0.4 and 0.6, respectively, highlighting limited predictive capability for Disease B. Calibration performance remains consistently strong for RSF and DeepHit, as indicated by their low Brier scores (around 0.1). Cox and Fine-Gray, however, show less accurate calibration, reflected by slightly higher Brier scores.

These results for the multimorbidity scenario suggest that, with $N = 1000$, RSF and DeepHit provide superior predictive performance for both diseases. RSF consistently excels in discrimination and calibration, followed closely by DeepHit. Cox and Fine–Gray perform adequately for Disease A but struggle notably in the context of Disease B, underscoring the complexity and variability in modelling competing morbidities accurately at smaller sample sizes.

N = 2000

Figure 4.7 illustrates the comparative performance of four competing-risk survival mod-els Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray in multimorbidity simulations with dataset size $N = 2000$.

For Disease A, the RSF and DeepHit models demonstrate notably strong and comparable discriminative performance, both attaining high C-index and ROC AUC values, generally above 0.8. Cox and Fine-Gray models show similar discriminative results, but at significantly lower levels, with C-index and ROC AUC values around 0.6, indicating limited ability to distinguish between events and non-events accurately. Calibration-wise, RSF and DeepHit outperform other models substantially, achieving low Brier scores close to 0.1. In contrast, Cox and Fine–Gray models exhibit weaker calibration accuracy, reflected by higher Brier scores, approximately 0.2.

In the context of Disease B, RSF and DeepHit again lead in discriminative power, consistently achieving C-index and ROC AUC scores at or above 0.8. Cox and Fine–Gray remain significantly behind, with lower discriminative capability (C-index and ROC AUC values near 0.4 and 0.6, respectively). The calibration accuracy follows a similar trend, with RSF and DeepHit showing superior performance, indicated by very low Brier scores (around 0.1), whereas Cox and Fine-Gray present moderately higher Brier scores, implying relatively poorer calibration.

For $N = 2000$, RSF and DeepHit clearly outperform the classical models (Cox and Fine-Gray) in predicting competing risks in multimorbidity contexts, both in terms of discrimination and calibration. The classical methods, while simpler to interpret, display substantial limitations, particularly in accurately discriminating Disease B, suggesting the advantages of adopting flexible, data-driven machine learning methods when handling complex competing-risk data.

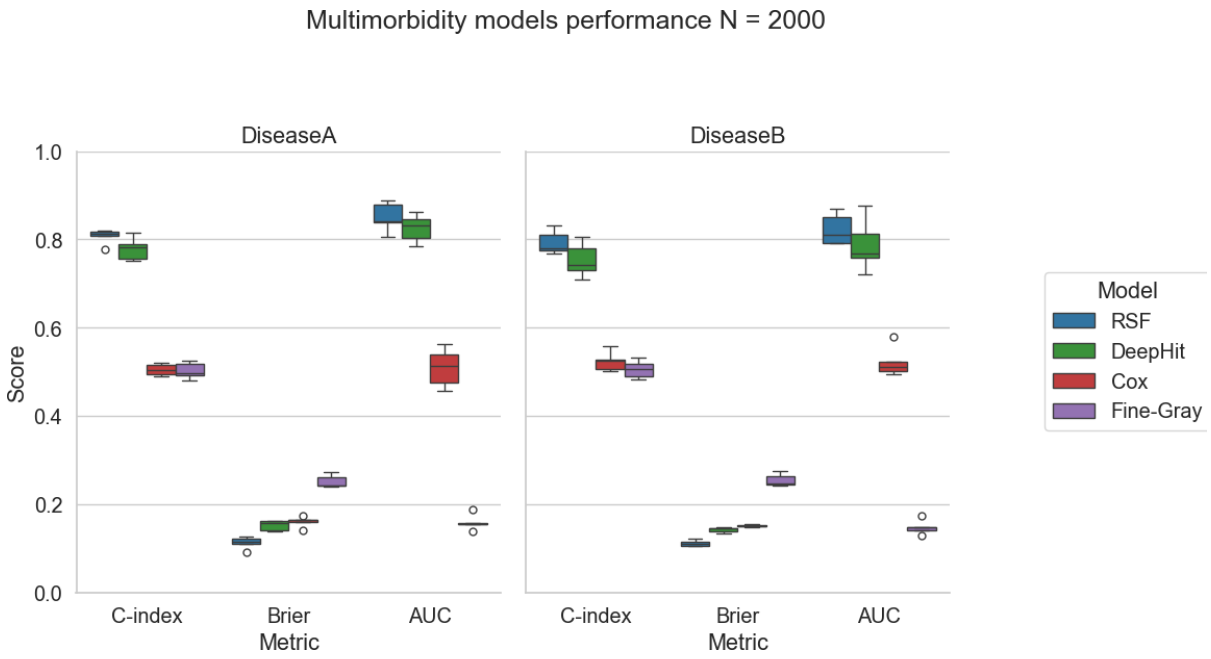


Figure 4.7: Model performance comparison for multimorbidity simulation ($N = 2000$), evaluating Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray models on Disease A and Disease B using C-index, Brier score, and ROC AUC.

N = 3000

Figure 4.8 presents the comparative performance of four competing-risks survival mod-els; Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray in the multimorbidity simulation context with dataset size $N = 3000$. Models are evaluated using the Brier score (lower scores indicate better calibration), as well as the C-index and ROC AUC (higher scores reflect greater discriminative capability).

For Disease A, RSF and DeepHit models continue to outperform traditional models (Cox and Fine-Gray), consistently achieving higher discrimination, as indicated by superior C-index and ROC AUC scores exceeding 0.8. Among these, RSF shows slightly better stability and marginally stronger discriminative performance compared to DeepHit. Cox and Fine-Gray exhibit significantly lower discriminative performance, with C-index and ROC AUC values clustering around 0.5, highlighting their limited capability in accu-rately distinguishing Disease A events from non-events. Regarding calibration accuracy, RSF and DeepHit again achieve very low Brier scores (approximately 0.1), indicative of excellent calibration. Conversely, Cox and Fine-Gray show relatively poorer calibration, reflected by their higher Brier scores (around 0.2 to 0.3).

Similarly, for Disease B, RSF and DeepHit maintain their superior discriminative performance, obtaining high C-index and ROC AUC scores, close to or surpassing 0.8. RSF shows slightly higher discriminative capability and stability compared to DeepHit. Cox and Fine-Gray remain significantly behind, with their discriminative performance remaining moderate (C-index and ROC AUC values approximately 0.6), underscoring their limited predictive accuracy. Calibration performance remains strong for RSF and

DeepHit, demonstrated by notably low Brier scores around 0.1, while Cox and Fine-Gray show weaker calibration, evidenced by higher Brier scores (approximately 0.2 to 0.3).

These findings reinforce the conclusion that, at larger sample sizes ($N = 3000$), advanced machine learning methods such as RSF and DeepHit clearly outperform traditional methods (Cox and Fine-Gray) in predicting competing risks in multimorbidity scenarios. This is evident both in terms of discrimination and calibration. Thus, these results highlight the importance of employing flexible data driven modelling approaches in the context of multimorbidity and competing risk analysis when adequate data is available.

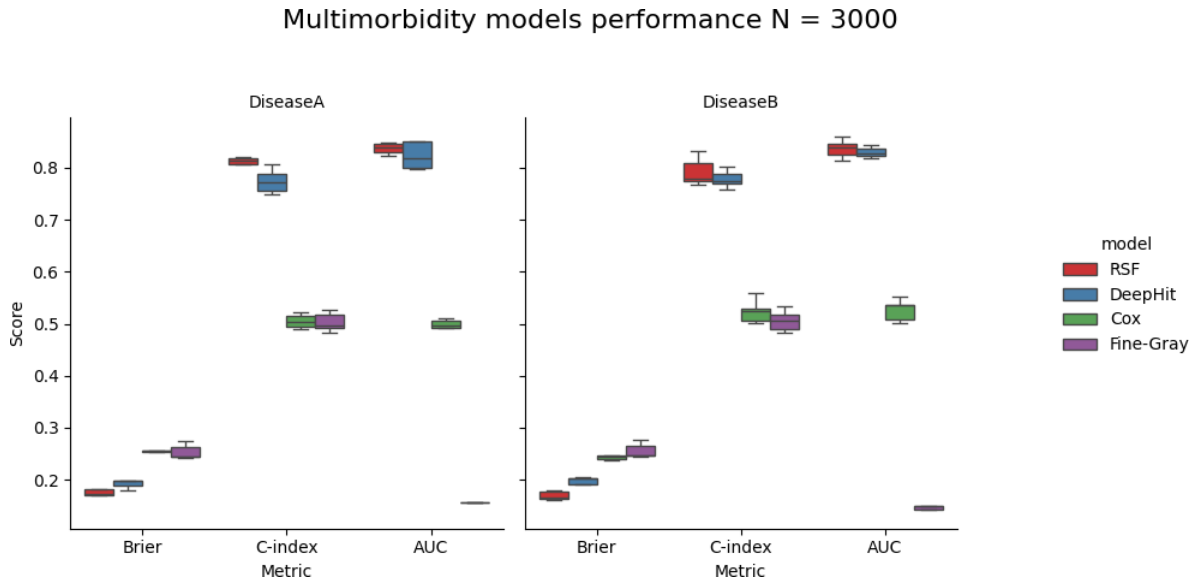


Figure 4.8: Model performance comparison for multimorbidity simulation ($N = 3000$), evaluating Random Survival Forest (RSF), DeepHit, Cox, and Fine-Gray models on Disease A and Disease B using Brier score, C-index, and ROC AUC.

4.4 Discussion

This study systematically evaluated the impact of correctly accounting for competing risks using survival analysis methods in two diverse scenarios, customer churn and multimorbidity. Four survival models, namely Cause-Specific Cox, Random Survival Forest (RSF), DeepHit, and Fine–Gray models, were rigorously compared through simulated data of sizes $N = 1000$, 2000, and 3000. The goal was to highlight potential risks associated with neglecting competing events, as ignoring these risks can lead to biased predictions, thereby misinforming decision-making processes.

4.4.1 Model Performance and Implications

In the customer churn scenarios, the Cox model consistently demonstrated superior discriminative performance and calibration across all dataset sizes, particularly excelling

in voluntary churn predictions (Figures 4.3 to 4.5). The RSF and DeepHit models closely followed Cox, showing competitive yet slightly weaker calibration. Fine–Gray, despite providing direct estimates of cumulative incidence functions, consistently exhibited lower discriminative ability and higher Brier scores, reflecting weaker calibration. These findings align with Blumenstock *et al.* [2022], who observed that machine learning approaches generally improve with larger datasets, yet classical models like Cox remain robust, particularly in datasets with clearly defined covariates.

For multimorbidity scenarios, RSF and DeepHit emerged as highly effective models, demonstrating superior discriminative capabilities and strong calibration (Figures 4.6 to 4.8). Notably, the Cox and Fine–Gray models performed poorly in distinguishing between diseases at smaller dataset sizes, showing substantial improvement only marginally at larger sizes. The observed limited performance of traditional models (Cox, Fine–Gray) for complex multimorbidity aligns with observations made by Lee *et al.* [2018], highlighting that deep learning methods better capture nonlinear relationships when sufficient data are available.

4.4.2 Comparative Analysis: Customer Churn vs Multimorbidity

Comparing the two scenarios, this study highlighted critical differences in how competing risks influence predictive accuracy. In customer churn, the differentiation between voluntary and involuntary churn was vital for accurate predictions, reinforcing the necessity of employing cause-specific or subdistribution hazard models to prevent biased estimations, as supported by Fine and Gray [1999]. Conversely, in multimorbidity scenarios, RSF and DeepHit significantly outperformed classical models, particularly at larger sample sizes, suggesting that more flexible, non-linear modeling methods are beneficial in complex health-related contexts where multiple chronic diseases interact dynamically.

These findings underscore the inherent risks associated with ignoring competing risks distorting the actual incidence probabilities and misguiding stakeholders, either in resource allocation in healthcare or retention strategies in business. Specifically, relying on standard survival estimators (such as Kaplan-Meier) in the presence of competing events often leads to a substantial overestimation of the absolute risk of the primary outcome, potentially triggering unnecessary or misdirected interventions Austin *et al.* [2016].

4.4.3 Alignment and Divergence with Existing Literature

The superior performance of Cox models in simpler scenarios corroborates existing literature suggesting traditional statistical methods remain reliable when dealing with moderately sized and clearly defined datasets Putter *et al.* [2007]; Klein and Moeschberger [2003]. In contrast, the significant advantage observed with RSF and DeepHit in complex scenarios supports recent literature advocating for machine learning methods in capturing intricate relationships within larger and more complex datasets Lee *et al.* [2018]; Blumenstock *et al.* [2022]. However, the comparatively weaker performance

of Fine–Gray across all simulations raises caution regarding its direct application without thorough data examination, reinforcing points made by Putter *et al.* [2007].

4.4.4 Strengths and Limitations of the Study

A key strength of this study was its systematic comparison across two distinctly different contexts—business analytics and health sciences which allowed for broader applicability and understanding of competing risk methodologies. Additionally, the use of rigorous metrics (C-index, Brier Score, ROC AUC) and robust cross-validation enhanced the reliability of findings.

However, several limitations are notable. First, simulated data, despite carefully designed distributions and relationships, inherently simplify real-world complexities. Additionally, the use of mostly exponential or piecewise exponential hazards in simulations may not accurately represent the complexity of real-life hazard dynamics, as noted by Beyersmann *et al.* [2011]. Real-world data typically contains time-dependent covari-ates, missing values, and confounding factors, none of which were explicitly modeled in this study.

Chapter 5

Conclusion

This dissertation investigated the importance of correctly modelling competing risks in survival analysis, specifically examining scenarios related to customer churn and multimorbidity. Through simulation-based analyses, we systematically compared the predictive performance of four competing-risk survival models: the Cause-Specific Cox, Random Survival Forest (RSF), DeepHit, and the Fine–Gray models, across dataset sizes of $N = 1000$, 2000, and 3000.

In customer churn scenarios, the results demonstrated the consistent superiority of the Cox proportional hazards model, which achieved high discriminative performance and excellent calibration across all evaluated dataset sizes. The RSF and DeepHit models showed competitive performance, but with minor differences in calibration and discrimination. The Fine–Gray model consistently underperformed, showing significant limitations in both discriminative power and calibration accuracy.

In contrast, for multimorbidity scenarios, RSF and DeepHit consistently outperformed classical models, particularly at larger dataset sizes. These advanced machine learning approaches exhibited robust discriminative capabilities and strong calibration, highlighting their suitability for handling complex and nonlinear relationships in competing-risks data. Conversely, traditional methods (Cox and Fine–Gray) showed comparatively weaker performance, especially for the more complex disease scenarios, revealing their limitations when faced with intricate risk structures.

These results underscore the critical implications of ignoring competing risks. Failing to account for these risks appropriately can lead to significant biases and misestimations in predicting event occurrences, thereby compromising decision-making processes in both business retention strategies and healthcare resource allocation.

Despite the strengths of this study, including rigorous simulation procedures, comprehensive model comparison, and clear relevance in distinct application domains, there remain certain limitations. The simulated nature of the datasets inherently limits direct generalizability to real-world scenarios, where complexities such as time-varying covariates, missing data, and unobserved confounding factors exist. Future research should aim to extend these analyses to more complex real-world datasets and explore further improvements in model calibration and hyperparameter tuning, particularly

for machine learning models such as DeepHit and RSF. Incorporating causal inference methodologies within competing risks frameworks also offers promising avenues for future investigation.

The study firmly establishes that accurately modelling competing risks significantly enhances predictive performance. The choice of modelling approach should carefully consider the complexity of the underlying data structure, sample size, and required interpretability. By selecting appropriate competing-risk models tailored to specific contexts, analysts and decision-makers can achieve more accurate, reliable, and impactful insights in both business analytics and public health domains.

5.0.1 Future Research Directions

Future studies could extend this work by incorporating more complex hazard structures, time-dependent covariates, and real-world datasets. Furthermore, exploring advanced hyperparameter tuning for machine learning models like DeepHit and RSF, and comparing performance in extensive observational studies or clinical trials would validate the practical applicability of these findings. Finally, integrating causal inference methods within competing risks frameworks could offer deeper insights into the causal relationships between covariates and multiple event outcomes.

References

[Austin *et al.* 2016] Peter C Austin, Douglas S Lee, and Jason P Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016.

[Beyersmann *et al.* 2011] Jan Beyersmann, Arthur Allignol, and Martin Schumacher. *Competing risks and multistate models with R*. Springer Science & Business Media, 2011.

[Blumenstock *et al.* 2022] Gabriel Blumenstock, Stefan Lessmann, and Hsin-Vonn Seow. Deep learning for survival and competing risk modelling. *Journal of the Operational Research Society*, 73(1):26–38, 2022.

[Breiman 2001] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[Cox 1972] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[Fine and Gray 1999] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.

[Gabriel Blumenstock and Seow 2022] Stefan Lessmann Gabriel Blumenstock and Hsin-Vonn Seow. Deep learning for survival and competing risk modelling. *Journal of the Operational Research Society*, 73(1):26–38, 2022.

[Gray 1988] Robert J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16(3):1141–1154, 1988.

[Harrell Jr *et al.* 1982] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

[Heagerty *et al.* 2000] Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[Ishwaran *et al.* 2014] Hemant Ishwaran, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange, and Bryan M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 04 2014.

[Jiang *et al.* 2018] Yue Jiang, Jason P Fine, and Amy K Mottl. Competing risk of death with end-stage renal disease in diabetic kidney disease. *Advances in Chronic Kidney Disease*, 25(2):133–140, 2018.

[Kaplan and Meier 1958] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[Klein and Moeschberger 2003] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2003.

[Klein *et al.* 2003] John P Klein, Melvin L Moeschberger, et al. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, 2003.

[Kleinbaum and Klein 1996] David G Kleinbaum and Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.

[Lee *et al.* 2018] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with compet-ing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Lu 2014] Yao Lu. *Statistical Methods for Competing Risks Model*. PhD thesis, Johns Hopkins University, 2014.

[Nelson 1972] Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

[Noordzij *et al.* 2013] Marlies Noordzij, Karen Leffondré, Karlijn J. van Stralen, Carmine Zoccali, Friedo W. Dekker, and Kitty J. Jager. When do we need competing risks methods for survival analysis in nephrology? *Nephrology Dialysis Transplantation*, 28(11):2670–2677, 08 2013.

[Pierri and Caroni 2020] Francesca Pierri and Chrys Caroni. Analysing the risk of bankruptcy of firms: survival analysis, competing risks and multistate models. *Demography of Population Health, Aging and Health Expenditures*, pages 385–394, 2020.

[Putter *et al.* 2007] Hein Putter, Marta Fiocco, and Ronald B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.

[Stanish *et al.* 1978] William M. Stanish, G.Y.H. Chi, William D. Johnson, Gary G. Koch, J.Richard Landis, and Shelley Liu-Chi. A computer program for the generalized chi-square analysis of competing risks grouped survival data (criscat). *Computer Programs in Biomedicine*, 8(3):208–223, 1978.

[Therneau *et al.* 2000] Terry M Therneau, Patricia M Grambsch, Terry M Therneau, and Patricia M Grambsch. Multiple events per subject. *Modeling survival data: extending the Cox model*, pages 169–229, 2000.

[Tsiatis 1975] A. Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):20–22, 1975.

[Uno *et al.* 2011] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[Zheng and Klein 1995] Ming Zheng and John P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.