Evaluation of Predictive Performance for Non-Linear Survival Models

Alfensi Faruk

School of Mathematics

University of Leeds

A thesis submitted for the degree of

Doctor of Philosophy

July 2024

Acknowledgements

First of all, I would like to express my sincere appreciation to all who have helped me in any respect during the preparation of this thesis.

I am deeply grateful to my supervisors, Dr Georgios Aivaliotis and Dr Jan Palczewski, for their invaluable help, support, guidance throughout my PhD study. Their mentorship has been instrumental in my academic growth. I feel fortunate to have had the opportunity to work with them these past years and I sincerely thank them for everything they have done for me. This thesis would not have been possible without their kindly guidance. I thank Dr Lanpeng Ji and Dr Seppo Virtanen for their feedback in the annual reviews.

I want to thank Indonesia Endowment Fund for Education (LPDP) for the funding provided during this PhD. I thank to Research Computing at the University of Leeds for their High Performance Computing facilities (ARC3 and ARC4).

Finally, I dedicate this thesis to my wife, my daughter, my parents, my brother, and my sisters for their endless support and encouragement. Special thanks go to my beloved wife Nurul Fajriyah Widya Utami and my beautiful daughter Maryam Azzahra Faruk for all their patience, understanding, care, and love.

Abstract

This thesis explores several measures of predictive performance for non-linear survival models, which fall into discrimination and calibration categories. We show that the integrated Brier score is more convincing when used with the integrated Brier score for Kaplan Meier estimator as the reference. To enhance its interpretability, we propose the normalised Brier score, centered integrated Brier score, and normalised centered integrated Brier score. They are developed by incorporating the variabilities of the predicted survival curves into the integrated Brier score.

To cope with non-proportional hazard data, a discrimination measure, time-dependent Uno's C-index, is proposed. This proposal is the result of a thorough examination of Uno's C-index's pitfalls. Its convergence is demonstrated in detail, following Nolan and Pollard's results for U-statistics. This comprehensive approach instils confidence in our findings.

We introduce pair calibration, a mean squared error of the model's outcomes and their respective predicted probabilities, as a novel measure for the survival models. We discuss pitfalls of the pair calibration and propose some reference values to cope with such issues.

While the unweighted measures' bias can be downward or upward depending on the fitted models, our proposed measures are unbiased as the censoring rate increases. Through numerical examples, we found that discrimination and calibration may oppose each other depending on the fitted model and the data structure.

Abbreviations

T_i	Event time of individual i
D_i	Censoring time of individual i
X_i	Observed time of individual i, i.e. $\min(T_i, D_i)$
$\mathbb{Z}_i$	$(p \times 1)$ vector of covariates of individual i
$S(t; \mathbb{Z}_i)$	Predicted survival curve at period t of individual i given $\mathbb{Z}_i$
$h(t; \mathbb{Z}_i)$	Predicted hazard probability at period t of individual i given $\mathbb{Z}_i$
$G(t)$	Probability of i to be censored after period t, i.e. $\mathbb{P}(D_i > t)$
$\widehat{G}_n(t)$	Predicted $G(t)$ based on $\{D_i, \cdots, D_n\}$ in the test data
$\widehat{\mathrm{C}}_n^{\mathrm{har}}$	Harrell's C-index
$\widehat{\mathrm{C}}_n^{\mathrm{uno}}$	Uno's C-index
$\widehat{\mathrm{C}}_n$	Time-dependent concordance
$\widehat{\mathrm{BS}}_n(t)$	Brier score at period t
$\widehat{\mathrm{IBS}}_n$	Integrated Brier score
$\widehat{\mathrm{IBSKM}}_n$	$\widehat{\mathrm{IBS}}_n$ for Kaplan-Meier (KM) estimator
$\widehat{\mathrm{NBS}}_n(t)$	Normalised Brier score at period t
$\widehat{\mathrm{NIBS}}_n$	Normalised integrated Brier score
$\widehat{\mathrm{NIBS}}_n^{\varepsilon}$	$\widehat{\mathrm{NIBS}}_n$ truncated at $\varepsilon \in [0, 1]$
$\widehat{\mathrm{NIBSKM}}_n$	$\widehat{\mathrm{NIBS}}_n$ for KM estimator
$\widehat{\mathrm{CBS}}_n(t)$	Centered Brier score at period t
$\widehat{\mathrm{CIBS}}_n$	Centered integrated Brier score

$\widehat{\text{NCBS}}_n(t)$	Normalised centered Brier score at period t
$\widehat{\text{NCIBS}}_n$	Normalised centered integrated Brier score
$\widehat{\text{NCIBS}}_n^{\varepsilon}$	$\widehat{\text{NCIBS}}_n$ truncated at $\varepsilon \in [0, 1]$
$\widehat{\text{NCIBSKM}}_n$	$\widehat{\text{NCIBS}}_n$ for KM estimator
$\widehat{\text{C}}_n^w$	Time-dependent Uno's C-index
$\widehat{\text{PC}}_n^1$	The first estimator of pair calibration
$\widehat{\text{PC}}_n^2$	The second estimator of pair calibration
$\widehat{\text{PC}}_n^{\tau_1,\tau_2}$	Estimator of pair calibration truncated at τ_1 and τ_2
$\widehat{\text{PC}}_n^{1,T_{max}}$	Estimator of pair calibration truncated at 1 and T_{max}
ref_1	Reference based on proportion of outcomes for $\widehat{\text{PC}}_n^1$
ref_2	Reference based on proportion of outcomes for $\widehat{\text{PC}}_n^2$
ref_3	Reference based on proportion of outcomes for $\widehat{\text{PC}}_n^{1,T_{max}}$

Contents

List of Figures

Chapter 1

Introduction

Advances in machine learning (ML) and artificial intelligence (AI) have enhanced accuracy in prediction for risks. The latter could relate either to classification problems, or to survival problems, which will be the focus of this thesis. Indicatively, we mention random survival forests (Ishwaran *et al.*, 2008) and neural networks (Gensheimer & Narasimhan, 2019).

Survival analysis, a branch of statistics that analyses the time until the event of interest occurs, has seen significant advancements in recent decades. Although survival analysis is a common term in medicine, it is also known as event history analysis, reliability analysis, and duration time analysis in many other fields of science. One unique feature of survival analysis is the presence of censoring in the data. When we do not know the exact event times of some individuals over the follow-up, the individuals are then called censored individuals. Using standard statistical methods to handle the censored individuals may not be appropriate and lead to bias. Hence, developing statistical methodologies for survival analysis has become one of the most popular research fields in statistics. Survival models, a celebrated area of success in statistics since the appearance of Cox proportional hazard (CPH) model (Cox, 1972), is of immense importance to a number of application areas, including finance and medicine. Another classical survival model is Kaplan-Meier (KM) estimator (Kaplan & Meier, 1958), a simple survival model ignoring the individuals' characteristics. Along with the CPH model, the KM estimator is still widely used as the benchmark for more sophisticated survival models. The many advances in statistical models for survival analysis (Kleinbaum

& Klein, 2010; Yang & Zou, 2013) have allowed for more complicated relationships between risk and covariates affecting the risk, however they remain focus on some form of (generalised) linear relationship. The introduction of AI and ML approaches rendered the users free of any form of linearity assumptions, allowed for complicated, highly non-linear relationships between risk and covariates to be captured and enhanced the potential accuracy of predictions. All this is at the expense of a lower ability to interpret the predictions and additional challenges in applying traditional measures for assessing the prediction accuracy.

Assessing the quality of risk predictions is crucial as it gives users confidence in their models and decisions. There is an array of measures of risk prediction accuracy in the literature. These measures fall broadly into two categories: calibration and discrimination (see D'Agostino & Nam (2003) for a review). Calibration assesses how closely the number of predicted events matches the number of observed events over the follow-up time. For example, Hosmer-Lemeshow goodness-of-fit test (Hosmer & Lemesbow, 1980) or 1-calibration tests whether the number of predicted events equals the number of observed events at a single prediction time. Whereas D-calibration (Haider *et al.*, 2020) employs the entire survival curve to test the calibration of survival distribution over follow-up time. The D-calibration error can be incorporated into the objective function during the training process (Goldstein *et al.*, 2020). Brier score (Brier & Allen, 1951), which has been adapted to cope with censored individuals (Graf *et al.*, 1999), is commonly considered as a calibration measure (DeGroot & Fienberg, 1983; Haider *et al.*, 2020). At each time point, the Brier score calibrates the actual individuals' status with the predicted survival probability while its respective integrated Brier score summarises the Brier scores over the follow-up.

Although the integrated Brier score is one of the most applied calibration measures in assessing the predictive performance of survival models, it may not properly demonstrate the models' predictive performance because the data structure predominantly affects its value. For example, Aivaliotis *et al.* (2021) used the integrated Brier score to evaluate the fitted model performance in breast cancer data. The data contains around 95.5% individuals who did not get breast cancer by the end of the study. The data structure has caused the predicted survival curves to be close to one regardless of the fitted models, while most outcomes are

also one. As a result, the integrated Brier score will be low and close to zero. One potential solution to cope with such a problem is to use integrated Brier score for KM estimator. Since the KM estimator ignores each individual's characteristics, it provides predicted survival curves that are common for all individuals. In other words, the KM estimator has poor prediction capability when we predict individual-specific survival curves. Therefore, using the integrated Brier score as the reference for the integrated Brier score of the fitted survival model is reasonable. The paper by Schumacher *et al.* (2003) is an example of work employing the Brier score for the KM estimator as the benchmark of other Brier scores obtained from the other fitted models. Another drawback of the integrated Brier score is its limited interpretability, where we can only say that the closer the integrated Brier score to zero, the better the model performance. The closer the integrated Brier score to one is, the worse the model's performance is. This problem is common for performance measures based on mean squared error. We can explicitly incorporate the "skill" into the Brier score (see Stephenson (2000) for more detail). In particular, we include a new reference to the Brier score so that its value is relative to the reference.

Discrimination on the other hand assesses how well a model can discriminate two different groups of the model's outcomes (D'Agostino & Nam, 2003). For instance, suppose we have a sample of cancer patients categorised into two groups, namely patients with and without some treatments after surgery. A survival study shows that the given treatments have significantly improved the life expectancy of the patients several years later. If our model has excellent discrimination capability, it can correctly distinguish the two groups of patients and identify the life expectancy improvement in the group with the treatments. For a review of survival model discrimination measures, see the paper by Rahman *et al.* (2017). A type of standard discrimination measure is the concordance index (C-index), trying to estimate the probability of "agreement" between the risk prediction and the actual outcome using pairwise comparisons. Harrell's C-index (Harrell *et al.*, 1982) is the most established one. It is appropriate to assess the prediction performance of any survival models whose outputs are time-independent predicted risk scores, such as the CPH model and parametric survival regressions. Another

example is Gönen's C-index (Gönen & Heller, 2005), the same as Harrell's C-index regarding the evaluated predictive risk scores. However, it is intractable since its use is only limited to the CPH model. Although the CPH model is the most commonly applied survival model, it assumes that proportional hazard (PH) holds in the data. The PH assumption states that the risks of experiencing the event of interest, e.g. the chance of getting heart failure in the next ten years, of two individuals are always proportional to each other as time goes by over the follow-up. This assumption is very strong since, in high-dimensional data, especially when some covariates are categorical, the assumption is prone to be violated as too many individuals' characteristics must be kept proportional. Moreover, in long follow-up, it is unlikely always to maintain the proportional risks, especially when dealing with data relating to living individuals. Consequently, some standard discrimination measures, such as Harrell's C-index and Gönen's C-index, are not proper for assessing the predictive performance of non-PH survival models.

One of the challenges arising from the introduction of highly non-linear survival models (such as ML and AI models but also statistical models) is the violation of the PH assumption. The violation allows for the possibility of a reversal of the relative risk relationship between two individuals, thus making the calculation of discrimination measures such as C-index precarious. Antolini's "time-dependent" version of Harrell's C-index (Antolini *et al.*, 2005) lends itself to address this challenge as it uses the risk predicted by the model at the time of the first event between a pair of individuals. However, no proof of convergence of the estimator to the probability of concordance is provided in Antolini *et al.* (2005). Time-dependent area under curve (AUC) (Heagerty & Zheng, 2005; Pepe *et al.*, 2008) is the area under receiver operating characteristics (ROC) curve drawn by plotting false-positive rate and true-positive rate in x-axis and y-axis, respectively. It is not an overall measure like Antolini's measure and only reports model performance at each time within the follow-up evaluating baseline or time-dependent risk scores.

Another well-documented problem of risk prediction evaluation measures, both calibration and discrimination, is the potential "bias" of these measures due to censored observations (Gerds *et al.*, 2013; Kvamme & Borgan, 2023). Uno's

C-index (Uno *et al.*, 2011) was developed to overcome the bias of the original C-index by applying an inverse probability of censoring weighted (IPCW) approach, where the probability of censoring does not depend on the covariates. The authors in Uno *et al.* (2011) also show that the estimator for Uno's C-index converges to the population probability of interest. Several IPCW-based C-indices for standard Harrell's C-index, where the censoring survival probability is conditional on covariates, are discussed in Gerds *et al.* (2013).

This thesis focuses on discrimination and calibration measures for potentially non-linear models and right-censored data for a single event of interest. In general, the main goals of this thesis are: (1) to propose several measures, such as modified integrated Brier scores, time-dependent Uno's C-index, and pair calibration, in evaluating the predictive performance of the non-linear survival models, (2) to investigate the behaviour of the proposed measures through numerical experiments, and (3) to prove the convergences of some of the proposed measures, e.g. the convergence of time-dependent Uno's C-index, which are not trivial.

The modified integrated Brier scores comprise three different Brier score-based measures, such as normalised integrated Brier score, centered integrated Brier score, and normalised centered integrated Brier score. These measures are developed by modifying the squared error term within the Brier score. The primary motivation behind the modified integrated Brier scores is to develop new measures based on the integrated Brier score that are relative to the variability of the predicted survival curves.

Time-dependent Uno's C-index is the IPCW version of time-dependent concordance introduced by Antolini *et al.* (2005). As we show, it is an unbiased estimator for the population probability of interest. By applying the IPCW approach, each uncensored observation is penalised to compensate for censoring bias. We prove the convergence of the proposed estimator to the population probability. Our proof complements the proof of Uno's C-index in that we show that the assumptions of Theorem 4.3.1 for almost-sure convergence of U-statistics hold.

The third proposed measure is pair calibration, which calibrates the order of event times for pairs of two individuals with their respective predicted probability. It can be considered a calibration measure for the discrimination of risks

(ranking) predicted by the model. Thus, it stands somewhere between calibration and discrimination. In particular, pair calibration provides how good is information about the discrimination through calibration. We also show the convergence of our estimators. In the current literature, Brier score is a prominent example of this type of measure, where its discrimination part can be obtained by decomposition (Murphy, 1973; Yates, 1982).

This thesis comprises seven chapters. The introduction is presented in Chapter 1. Chapter 2 provides the basic notations, assumptions, and principles as the main foundations throughout this thesis. We start with basic notations and principles and then discuss the standard performance measures categorised into calibration and discrimination.

Chapter 3 first discusses the pitfalls of the integrated Brier score, particularly the effect of the distribution of survival times in the data on the value of the integrated Brier score. Then, we use the reference value based on the integrated Brier score for KM estimator to cope with such drawbacks. After that, we introduce some alternative versions of the integrated Brier score, namely the normalised integrated Brier score, centered integrated Brier score, and normalised centered integrated Brier score. In particular, we show how the integrated Brier score and its modified versions behave in the good and bad models. Two comprehensive simulation scenarios and implementations to real datasets, the cancer genomic atlas (TCGA) mutations data and breast cancer data, are then conducted to demonstrate the pitfalls. These thorough numerical experiments will become our primary motivations to use the integrated Brier score for KM estimator as a complement measure, instilling confidence in our proposed solution.

Chapter 4 develops time-dependent Uno's C-index to alleviate the possible bias occurring in time-dependent concordance. We show that existing applications of Uno's C-index are inappropriate for non-linear models with non-PH data. The numerical experiments motivate us to propose time-dependent Uno's C-index as the proper measure in such a situation. The convergence of time-dependent Uno's C-index estimator is shown using almost sure convergence of the U-statistics theorem. We also show that the time-dependent Uno's C-index estimator is less biased than the time-dependent concordance when the censoring rate is large via simulation studies and real-world examples.

In Chapter 5, we propose several estimators of pair calibration and demonstrate their convergence. Through simulation studies and real-world examples, we show the practical pitfalls of these estimators. This practical insight has led us to propose two measures that can be used as the reference values for the pair calibration. We also introduce the truncated pair calibration developed for anyone interested in the model performance at a specific sub-interval of the follow-up.

In Chapter 6, we conduct more comprehensive numerical experiments of the proposed measures via simulation studies. The chapter begins with two simulation studies: (1) varying censoring rates in test data, and (2) varying L^2-regularisation in the fitted models with fixed censoring rates. Next, we present two example cases where discrimination and calibration are not in line with each other: (1) good discrimination but poor calibration, and (2) good calibration but poor discrimination. Since throughout this thesis, we mainly apply the proposed measures to evaluate the model performance of a type of non-linear survival model, namely Nnet-survival, in the last section of this chapter, we also apply the proposed measure to evaluate the model performance of random survival forests. Finally, in the last chapter (Chapter 7), we summarise the conclusions, future work, and possible extensions.

Chapter 2

Preliminaries

This chapter provides basic concepts, types of survival models, and standard performance measures as the main foundations of this thesis. We start with basic notations and principles and then discuss standard performance measures categorised into calibration and discrimination at the end of this chapter.

2.1 Basic Notations and Principles

We denote by $\mathcal{T} = \{1, \cdots, T_{max}\}$ the discrete set of possible event times, where the elements of $\mathcal{T}$ are referred to as "periods". $\mathcal{T}$ may be dictated by the data or obtained by discretising a follow-up $[0, \mathrm{B}]$ into T_{max} time intervals

$$[a_0, a_1), [a_1, a_2), \cdots, [a_{T_{max}}, \infty),$$

where $a_0 = 0$, and T_{max} is the period in which all individuals are assumed to be administratively censored. In practice, we often discretise the survival data collected in discrete-time units into a sufficiently smaller number of periods because the structure of many standard discrete-time survival models cannot handle large number of periods.

Remark 2.1.1. The proposed measures in this thesis are mainly designed for discrete-time units. The measures can generally be applied to continuous-times units even though some proposed measures require adaptations. In the continuous-time units, to compute the integrated Brier score, modified integrated Brier

scores, and pair calibration, we have to use integration instead of summation with respect to time over the follow-up. Meanwhile, time-dependent Uno's C-index can be applied to the continuous-time units as long as each survival time is considered as a unique quantity. However, in practice we still usually record the continuous survival times in discrete-time units (e.g. days, months, and years) so that we do not need to discretise the observed survival times, and hence the time-dependent Uno's C-index can be directly applied.

Let event time of individual $i(i = 1, \cdots, n)$ be denoted by T_i which is either a random variable or its realisation in the data, depending on the context. Throughout this thesis, our survival model is defined as a function

$$S : (t, \mathbb{Z}_i) \mapsto [0, 1],$$

where $\mathbb{Z}_i = [Z_{i1} \cdots Z_{ip}]'$ is $(p \times 1)$ vector of covariates of individual i, and $t \in \mathcal{T}$. The predicted survival curve, $S(t; \mathbb{Z}_i)$, is the predicted probability of $T_i > t$. It can be interpreted as the predicted probability of individual i surviving from the beginning of the follow-up time up to the end of period t.

We can obtain the predicted survival curve in various ways depending on the fitted survival models. For example, in most classical discrete-time survival models (Singer & Willett, 1993; Tutz & Schmid, 2016), the modelled quantities are the predicted hazard probabilities, namely the predicted probability of experiencing the event of interest in a given period t provided that the individual has survived to the beginning of period t as follows

$$h(t; \mathbb{Z}_i) = \mathbb{P}\left[T_i = t \middle| T_i \geq t, \mathbb{Z}_i\right]. \tag{2.1}$$

Then, the predicted survival curve is computed by

$$\begin{aligned} S(t; \mathbb{Z}_i) &= \prod_{r=1}^{t} \left(1 - h\left(r; \mathbb{Z}_i\right)\right) \\ &= S(t-1; \mathbb{Z}_i)\left(1 - h\left(t; \mathbb{Z}_i\right)\right). \end{aligned} \tag{2.2}$$

A number of artificial neural networks (ANN) (Brown *et al.*, 1997; Gensheimer & Narasimhan, 2019) also compute $h(t; \mathbb{Z}_i)$ or $1 - h(t; \mathbb{Z}_i)$ before obtaining (2.2). Discrete-time survival forests (Schmid *et al.*, 2020) compute $h(t; \mathbb{Z}_i)$ before we transform it to (2.2). However, in Kaplan-Meier (KM) estimator, the survival

curve is directly estimated from the data without any transformation. The predicted survival curve in the KM estimator is common for all individuals, that is $S(t; \mathbb{Z}_i) = S(t)$, which is computed as follows

$$S(t) = \prod_{q=1}^{t} \left(1 - \frac{d_q}{n_q}\right), \tag{2.3}$$

where for $t \in \mathcal{T}$, d_q is the number of events in the train data within period q, n_q is the number of individuals who have survived up to period q, and $S(0)$ is defined to equal one.

When we have population models, throughout this thesis, we assume that the data are from a population where covariates $\mathbb{Z}_i (i = 1, \cdots, n)$ are independent and identically distributed (i.i.d.) with an unknown distribution, and the event times T_i are independent between individuals. Given a realisation of individual covariates, event time follows a distribution which depends on the covariates, where this distribution is again unknown. All expectations and probabilities in the thesis will be with respect to such model. Although survival models can be extended to handle multiple events of interest, this thesis is restricted to survival models for single event of interest only.

2.1.1 Censoring

Although we expect to collect n independent random samples $T_1, T_2, \cdots, T_n$, some of them may not be observed. A number of individuals may not experience the event of interest so that we only know their partial information. Those 'surviving' individuals are known as censored data. Censoring is a crucial issue in survival analysis since excluding them completely from the analysis may lead to bias.

One of the most common types of censoring is right censoring. Three common causes of right censoring are:

1. The individuals do not experience the event of interest before the study ends. Because the study duration is usually limited, there is no guarantee that all individuals in the study experience the event of interest. They may survive within the study, but their exact event times are unknown.

2. Some individuals are lost to follow-up. Sometimes, we cannot manage all individuals to always be in the study. For instance, suppose we are interested in the time to death. However, some individuals have decided to move out of their current countries, so we cannot track them anymore.

3. The individuals decide to withdraw from the study. For instance, we are interested in time until death due to a heart attack from heart surgery patients in a hospital. However, some patients decide to exit the study to avoid possible side effects from medical treatment.

Although there are other types of censoring, e.g., left censoring and interval censoring (Gijbels, 2010; Kleinbaum & Klein, 2010), we only focus on right censoring throughout this thesis. Therefore, for simplicity, any censoring mechanism in this thesis henceforth refers to right censoring.

Denote by D_i the censoring time of individual i, where D_i is assumed to be i.i.d. between individuals, so in particular D_i is independent from event time T_i and covariates $\mathbb{Z}_i$. The notation D_i will represent a random variable or its realisation in the data, depending on the context. We then write $X_i = T_i \wedge D_i := \min(T_i, D_i)$ – this is the time available from the data; it is either the event time if it happens before censoring or, otherwise, the censoring time. We will use subscripts i, j to denote two randomly chosen individuals from the whole population or a sample, depending on the context. The independence assumption between D_i and $\mathbb{Z}_i$ is common in the literature. Although Gerds *et al.* (2013) have tried to relax the assumption and to incorporate a set of covariates into X_i and D_i, there was no significant improvement of their model performance.

We denote by $G(t) = \mathbb{P}(D_i > t)$, which is the probability of individual i to be censored after period t. The distribution of censoring times G is approximated by an empirical tail function $\widehat{G}_n$ based on observed censoring times $\{D_i, \cdots, D_n\}$ in the test data, which is the same data as the one in computing the estimator of the measure of performance. Due to the independence of the censoring times from the covariates and event times, this estimator is consistent. Alternatively, one can use the KM estimator in which the role of censors and event times is swapped; assuming that the test set is drawn from the population distribution, this estimator is also consistent (Wang, 1987) and makes better use of the available data.

In addition to its consistency, KM estimator is also the standard method for $\widehat{G}_n$ (Gerds *et al.*, 2013; Uno *et al.*, 2011). Throughout this thesis, we therefore use KM estimator to compute $\widehat{G}_n$.

Due to the end of the study observation, we will also perform administrative censoring throughout this thesis so that all individuals within the last period T_{max} are censored. In other words, if $X_i \geq a_{T_{max}}$, then $X_i = T_{max}$, and individual i will be administratively censored. As a result, the overall measures only assess the model performance over periods $\{1, \cdots, T_{max} - 1\}$.

2.1.2 Data Discretisation

Let $\mathscr{D} = \{a_0, a_1, \cdots, a_{T_{max}}, \infty\}$ be a set of discretisation points with respect to time intervals $[a_0, a_1), [a_1, a_2), \cdots, [a_{T_{max}}, \infty)$. We will have T_{max} possible periods for discrete-time survival models from $\mathscr{D}$. Varying the discretisation points in $\mathscr{D}$ results in different data structures because it determines the observed periods for each individual. As a result, we can obtain different models by varying the discretisation points in the train data. Note that throughout this thesis the data structure means the distribution of the observed periods in the data.

Gensheimer & Narasimhan (2019) have applied several discretisation setups for their proposed non-linear survival model, namely equally spaced time intervals and increasing width time intervals. They showed that Harrell's C-index is robust to their discretisation points setups. In other words, they concluded that Harrell's C-index does not depend on the change in the data structure. This thesis will also study whether our proposed measures are robust or not to data structure change by varying the splitting points setups.

The literature generally has no standard guidelines for the data discretisation method. The study of the discretisation method will require proper mathematical principles, extensive experiments, and appropriate measures of performance. Different measures will choose different discretisation setups. For examples, Gensheimer & Narasimhan (2019) used Harrell's C-index when for their proposed discretisation method. Meanwhile, Kvamme & Borgan (2021) applied time-dependent concordance (Antolini *et al.*, 2005) and average mean squared error (MSE) between the survival estimates and the true survival function for

their proposed methods. However, the main objective of this thesis is to evaluate the quality of the outputs of non-linear survival models regardless of the best method to train the models. Moreover, the discussion about the optimal discretisation setup is beyond the scope of this thesis. The chosen discretisation setup in this thesis only aims to describe various data structures (i.e. the observed period distributions) and how our measures behave in those discretisation setups.

2.1.3 Population Probability of Interest

When we evaluate the model performance using discrimination measures, we expect that the proposed estimators are convergent to the population probability. For instance, the population probability at $t \in \mathcal{T}$ of Harrell's C-index in Section 2.4.1 is defined as follows

$$\mathbb{P}\big[S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}\big]. \tag{2.4}$$

Probability (2.4) explains that the population consists of T_i with covariates vector $\mathbb{Z}_i (i = 1, \cdots, n)$. Therefore, when we have two individuals $(i \neq j)$ from the population, we take two independent random variables T_i and T_j and assess their concordance with the respective model's outputs $S(t;\mathbb{Z}_i)$ and $S(t;\mathbb{Z}_j)$. Given the order of their model's outcomes, it represents the quality of the order of the model's outputs of two different individuals in the population. The closer the value of the probability to one, the better the model performance. If it equals 0.5, the performance is similar to random guessing.

Throughout this thesis, when we prove the convergence of the proposed estimators, for $(i = 1, \cdots, n)$, we will have the following regularity conditions:

(R1): T_i are i.i.d. with an unknown distribution,

(R2): D_i are i.i.d. with an unknown distribution,

(R3): D_i are unconditionally independent to T_i, and

(R4): D_i do not depend on covariates $\mathbb{Z}_i$.

Note that the primary aim of this thesis is to propose measures for assessing the predictive performance of non-linear survival models for right-censored data with a single event of interest. Any measures discussed in this thesis are general and can be used properly to evaluate the model performance of any survival models with such characteristics. We choose Nnet-survival (Gensheimer & Narasimhan, 2019) as the example of the models evaluated throughout this thesis. We therefore introduce the basic of Nnet-survival in the subsequent section. Readers can skip the section if they want to evaluate the other relevant models.

2.2 Nnet-survival

Nnet-survival is a machine learning approach, particularly artificial neural networks (ANN), for discrete-time survival models, requiring the follow-up time to be discretised into a set of periods. We recommend the book by Goodfellow *et al.* (2016) to those interested in ANN's basic principles. Moreover, readers who are interested in more details about Nnet-survival, such as the likelihood function derivation, the model training method, and the codes, can refer to the original paper of Nnet-survival by Gensheimer & Narasimhan (2019). A number of available works in the literature have applied Nnet-survival as one of their proposed machine learning approaches for survival analysis prediction (Deng *et al.*, 2023; Suresh *et al.*, 2022). However, due to this thesis's scope and aims, the rest of this section will only discuss the basic architecture of Nnet-survival. Some benefits of using Nnet-survival for this thesis are:

1. Nnet-survival is a flexible approach for survival data. We can tune its hyper-parameters and the discretisation setups to fit the objectives of our numerical experiments.

2. Nnet-survival estimates the model's outputs $(1 - h(t; \mathbb{Z}_i))(i = 1, \cdots, n)$ for each $t \in \mathcal{T}$. This model structure is suitable for one of our proposed measures, namely pair calibration in Chapter 5, requiring the presence of outputs in each period over the follow-up to compute its probabilities of interest.

3. Since Nnet-survival is a type of discrete-time survival model, the non-strict inequalities sign in the pair calibration can be explored optimally.

Graphical representation, as shown by Figure 2.1, might be one of the straightforward approaches to describe ANN architectures. ANN contains one input layer, a number of hidden layers, and an output layer, where weights denoted by arrows linking those layers. Each layer has at least one node, which processes the incoming information passed to them. The nodes in hidden layers contain some linear operations of the incoming information, weights, and biases. The results of the linear operations are then transformed into the node's output by a non-linear activation function, e.g. sigmoid, softmax, and rectified linear unit (ReLU) functions. These non-linear functions are beneficial since they allow the ANN to learn complex patterns. After that, the nodes' outputs are passed on to the subsequent layer. These process aim to update the model's parameters, i.e. the weights, using an optimisation algorithm such that the loss function is minimised. Some common optimisation algorithm are stochastic gradient descent (SGD) and adaptive moment estimation (Adam). These algorithms require several hyper-parameters, such as learning rate, the batch size, the number of epochs, and regularisation method.

In Nnet-survival, the contribution of a period $t \in \mathcal{T}$ to the overall log-likelihood function is given by

$$\ell = \sum_{i=1}^{d_t} \log\left(h\left(t; \mathbb{Z}_i\right)\right) + \sum_{i=d_t+1}^{r_t} \log\left(1 - h\left(t; \mathbb{Z}_i\right)\right), \tag{2.5}$$

where d_t is the number of individuals who have experienced the event during the period t, and r_t is the number of individuals who have survived before period t. The optimisation algorithm minimises the negative log-likelihood function. Suppose that we have $(n \times p)$-dimensional matrix $\mathbb{D}$ as our design matrix as follows

$$\mathbb{D} = \begin{bmatrix} Z_{11} & \dots & Z_{1p} \\ \vdots & & \vdots \\ Z_{n1} & \dots & Z_{np} \end{bmatrix} = \begin{bmatrix} \mathbb{Z}_1 \\ \vdots \\ \mathbb{Z}_n \end{bmatrix}.$$

The outputs of Nnet-survival with one hidden layer are given by

$$1 - h(t; \mathbb{Z}_i) = f^{(2)}(f^{(1)}(\mathbb{Z}_i); \boldsymbol{w}_1, b_1); \boldsymbol{w}_2, b_2),$$

where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are vectors of weights, b_1 and b_2 are bias components, and $f^{(1)}$ and $f^{(2)}$ are non-linear activation functions. If we map the input $\mathbb{Z}_i$ to the output function by using r activation functions $\{f^{(1)}, \cdots, f^{(r)}\}$, the output is given by

$$1 - h(t; \mathbb{Z}_i) = f^{(r)}(f^{(r-1)}(\cdots(f^{(1)}(\cdots(\mathbb{Z}_i; \boldsymbol{w}_1, b_1); \boldsymbol{w}_2, b_2); \cdots); \boldsymbol{w}_r, b_r),$$

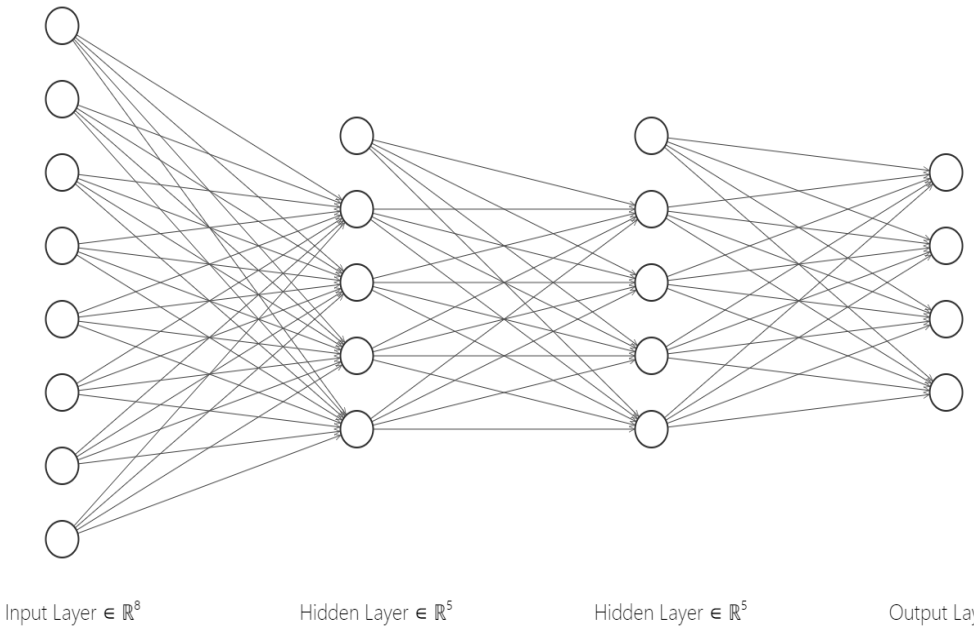
where $(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r)$ and $(b_1, \cdots, b_r)$ are sets of r vectors of weights and r biases, respectively. Nnet-survival tries to estimate the optimum values of those weights and biases that minimise the loss function using an optimisation algorithm. In particular, in all the numerical experiments of this thesis, we will apply SGD as the optimisation algorithm. Note that Nnet-survival is categorised as a non-linear survival model because the non-linear activation functions in the hidden layers have mapped the linear combination of input units to scalars.

The dimension of the vector of covariates $\mathbb{Z}_i (i = 1, \cdots, n)$, namely p, represents the length of the input layer. If $p = 1$, the number of nodes in the input layer equals one. Although Nnet-survival can handle large number of p, i.e. $p > n$, in this thesis we limit our work to the case $p < n$. Meanwhile, the number of nodes in the output layer must be equal to the number of all possible periods $\mathcal{T}$. The $(r-1)$ activation functions $\{f^{(1)}, \ldots, f^{(r-1)}\}$ process the incoming information to the hidden layers, and $f^{(r)}$ processes the incoming information from the final hidden layer to the output layer. We use the sigmoid function for $f^{(r)}$ to transform the inputs from the final hidden layer into the conditional probability of surviving at each $t \in \mathcal{T}$.

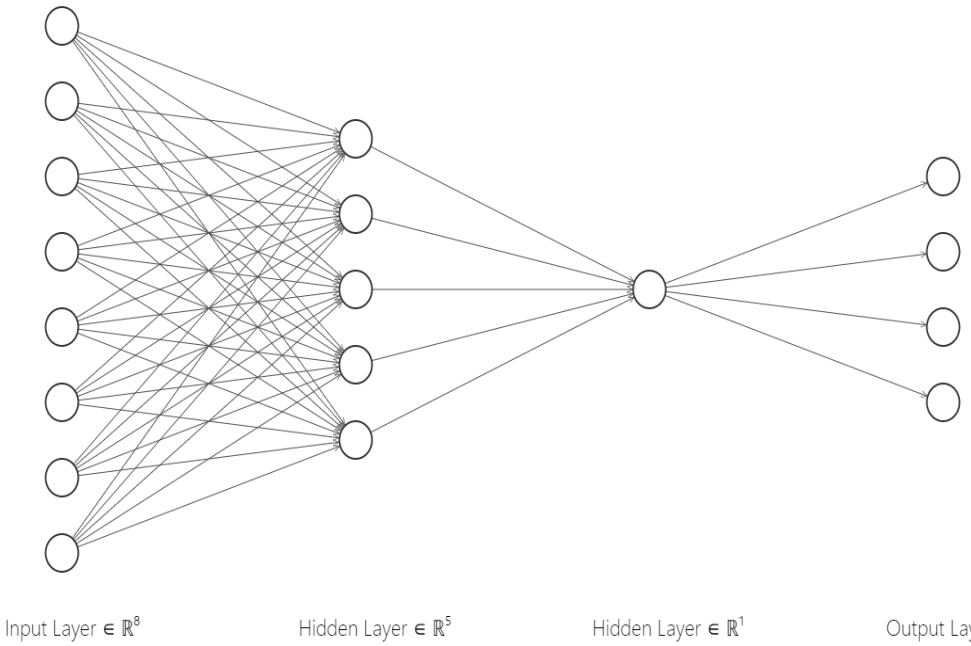
Nodes denote the units of all layers in Nnet-survival, and the weights are denoted by arrows connecting the nodes from one layer to the nodes in another layer (Figure 2.1). The training process learns the optimum values of weights and biases that minimise the negative of the log-likelihood function (2.5). Panel (a) of Figure 2.1 displays a fully connected network with one input layer (8 nodes), two hidden layers (5 nodes in each layer) where the nodes without incoming edge are the biases, and an output layer (4 nodes). Meanwhile, panel (b) of Figure 2.1 has one input layer (8 nodes), the first hidden layer (5 nodes), the final hidden layer (1 node) where there is no bias in the hidden layers, and an output layer (4 nodes).

Two versions of Nnet-survival are: (1) flexible Nnet-survival, and (2) PH Nnet-survival. In the flexible version, the final hidden layer must be fully connected to the output layer (see panel (a) of Figure 2.1). This approach allows the baseline hazard rate to vary over the follow-up time so that the PH assumption does not hold. Since we have five nodes, including the bias node in the final hidden layer, the outputs $(1 - h(t; \mathbb{Z}_i))$ can vary freely over the follow-up depending on the incoming information from the five nodes. On the other hand, in the PH version, the final hidden layer must contain only one node and the prior hidden layer is fully connected to the final hidden layer without biases (see panel (b) of Figure 2.1). Because we have only one node in the final hidden layer, the outputs $(1 - h(t; \mathbb{Z}_i))$ are only affected by a single fixed incoming information. If the incoming information results in a high risk of death at period $t = 1$, then the high risk of death also happens to the rest of periods $t = 2, \cdots, T_{max}$.

We can see also from the figures that the PH Nnet-survival is only a particular case of the flexible Nnet-survival. Throughout this thesis, we apply our proposed measures to evaluate the performance of the flexible version only. In addition, the flexible Nnet-survival will benefit one of the proposed measures, namely time-dependent Uno's C-index in Chapter 4, which is mainly developed for non-PH data. For the rest of this thesis, we refer to Nnet-survival as the flexible version that allows for the violation of the PH assumption.



(a) A flexible Nnet-survival with two hidden layers.



(b) A PH Nnet-survival with two hidden layers.

Figure 2.1: Examples of graphical representation of Nnet-survival.

2.3 Calibration Measure: Integrated Brier Score

In this section and the subsequent section, we will present standard performance measures adapted to discrete-time units, categorised into calibration and discrimination, which will be applied to evaluate the predictive performance of non-linear survival models. We first discuss calibration which assesses how well the predicted survival curves reflect the observed survival times.

Brier score (Brier & Allen, 1951) is one of the most applied measures in evaluating the performance of statistical model predictive performance. Brier score (and hence integrated Brier score) were adapted to cope with censored data by Graf *et al.* (1999) and commonly categorised as calibration measures (DeGroot & Fienberg, 1983; Haider *et al.*, 2020). We first consider the following expectation:

$$\mathbb{E}\left[\left(I_{\{T_i>t\}} - S(t;\mathbb{Z}_i)\right)^2\right], \tag{2.6}$$

which can be rewritten as follows

$$\mathbb{E}\left[\left(I_{\{T_i\leq t\}}\left(0 - S(t;\mathbb{Z}_i)\right)^2 + I_{\{T_i>t\}}\left(1 - S(t;\mathbb{Z}_i)\right)^2\right]. \tag{2.7}$$

To derive an estimator of (2.6) or (2.7) as proposed by Graf *et al.* (1999), we first transform (2.7) as follows

$$\mathbb{E}\left[I_{\{X_i\leq t\}}\frac{I_{\{T_i<D_i\}}}{G(T_i)}\left(0 - S(t;\mathbb{Z}_i)\right)^2 + I_{\{X_i>t\}}\frac{I_{\{t<D_i\}}}{G(t)}\left(1 - S(t;\mathbb{Z}_i)\right)^2\right], \tag{2.8}$$

where the expected value of $\left(I_{\{T_i<D_i\}}/G(T_i)\middle|T_i,\mathbb{Z}_i\right)$, where D_i is independent of T_i and $\mathbb{Z}_i$, is valid because

$$\begin{aligned}\mathbb{E}\left[\frac{I_{\{T_i<D_i\}}}{G(T_i)}\middle|T_i,\mathbb{Z}_i\right] &= \frac{1}{G(T_i)}\mathbb{E}\left[I_{\{T_i<D_i\}}\middle|T_i,\mathbb{Z}_i\right] \\ &= \frac{1}{G(T_i)}G(T_i) \\ &= 1.\end{aligned} \tag{2.9}$$

The analogous reasons also apply to the expected value of $\left(I_{\{t<D_i\}}/G(t)\middle|\mathbb{Z}_i\right)$, where D_i is independent of $\mathbb{Z}_i$. Thus, based on (2.8), the estimator of (2.7) is

defined as follows

$$\widehat{\mathrm{BS}}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left[I_{\{X_i\leq t\}}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\left(0-S(t;\mathbb{Z}_i)\right)^2 + I_{\{X_i>t\}}\frac{I_{\{t<D_i\}}}{\widehat{G}_n(t)}\left(1-S(t;\mathbb{Z}_i)\right)^2\right], \tag{2.10}$$

where $\widehat{G}_n$ is the estimator of G based on n observed censoring times in the test data. (2.10) is known as Brier score at t computed empirically on data with size n. Finally, the integrated Brier score over $\{1,\cdots,T_{max}-1\}$ is defined by

$$\widehat{\mathrm{IBS}}_n = \frac{1}{T_{max}-1}\sum_{t=1}^{T_{max}-1}\widehat{\mathrm{BS}}_n(t). \tag{2.11}$$

Brier score and integrated Brier score take values in $[0,1]$. The closer the values to zero, the better the model performance.

Remark 2.3.1. In practice, we need to carefully tune how far $\widehat{G}_n$ can be from 0. Too far from 0 will simplify the effect of G while too small $\widehat{G}_n$ will blow up the contributions of the respective individuals to the measure. We will introduce the minimum gap ($\epsilon > 0$) between $\widehat{G}_n(T_{max}-1)$ and 0 to the proposed measures when we show the convergence of time-dependent Uno's C-index in Chapter 4.

Remark 2.3.2. $G(T_i)$ accounts for the skewed distribution of event times T_i over follow-up. Smaller values of $G(T_i)$ indicate that the contribution of predictions at T_i on the model performance assessment is smaller than other predictions with large $G(T_i)$.

2.4 Discrimination Measures

Selecting proper measures of generalisation error is one of the essential tasks in statistical predictive modelling. The measures should match the characteristics of the evaluated models. Survival models may have different representations for the survival distribution, e.g. parametric, non-parametric, or may have censored observations. Unlike calibration measures, discrimination measures are sensitive to the choice of the parametrisation of the survival distribution. For instance, Harrell's C-index is appropriate for assessing the performance of the CPH model.

However, it is unsuitable for non-PH models, such as random survival forests and many other machine learning approaches (Antolini *et al.*, 2005; Sonabend *et al.*, 2022).

Calibration and discrimination use different approaches to assess the predictive performance of survival models. On the one hand, calibration assesses how well the predicted survival curves reflect the observed survival times. On the other hand, discrimination assesses how well a model determines which individual has a better-predicted survival curve in a pair of individuals. Therefore, their justifications for the performance of the same model can be different. For instance, we may have bad calibration but good discrimination, as shown later in Chapter 6.

This section is devoted to discrimination measures. In particular, we discuss three standard discrimination measures for assessing the predictive performance of survival models, namely Harrell's C-index, Uno's C-index, and time-dependent concordance. We first adapt them to their discrete-time versions. Finally, we show their convergence, especially for Harrell's C-index and time-dependent concordance.

2.4.1 Harrell's C-Index

Harrell's C-index (Harrell *et al.*, 1982) was developed for assessing the discrimination capability of the CPH model and aimed at estimating $\mathbb{P}\left[R_i < R_j | T_i > T_j\right]$, where R_i and R_j are risk scores of individuals i and j. Applying this idea to non-PH models requires specification of the risk score; we take the predicted survival probability at a fixed period $t \in \mathcal{T}$. Using $S(t; \mathbb{Z}_i)$ instead of R_i in C-index is common for the evaluation of survival models, such as the works by Antolini *et al.* (2005), Gensheimer & Narasimhan (2019), and Kvamme *et al.* (2019).

According to Harrell *et al.* (1982), an ordered pair of two different individuals $(i \neq j)$ with the respective observed times X_i and X_j is "usable" if both individuals are uncensored, or the first is uncensored and the censored time of the other is larger than the uncensored time. We define that the usable pair is concordant if $S(t; \mathbb{Z}_i)$ is less than $S(t; \mathbb{Z}_j)$ when $T_i < X_j$. Harrell's C-index at $t \in \mathcal{T}$ is given

by the ratio of the number of all concordant pairs to the number of all usable pairs:

$$\widehat{\mathrm{C}}_n^{\mathrm{har}}(t) = \frac{\sum_{i\neq j}^{n} I_{\{T_i<D_i\}} I_{\{S(t;\mathbb{Z}_i)<S(t;\mathbb{Z}_j)\}} I_{\{T_i<X_j\}}}{\sum_{i\neq j}^{n} I_{\{T_i<D_i\}} I_{\{T_i<X_j\}}}. \tag{2.12}$$

In the following theorem, we state the convergence of $\widehat{\mathrm{C}}_n^{\mathrm{har}}(t)$, where the proof will be discussed in Subsection 2.4.4.

Theorem 2.4.1 (Convergence of Harrell's C-Index)**.** *Suppose the regularity conditions (R1-R4) in Section 2.1.3 hold.*

$$\widehat{C}_n^{har}(t) \overset{wp1}{\to} \mathbb{P}\big[S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j)\big|T_i < T_j, T_i < D_i \wedge D_j\big], \tag{2.13}$$

where wp1 stands for convergence with probability one.

2.4.2 Uno's C-Index

Harrell's C-index (2.12) computes the concordance of pairs for which $T_i < D_i \wedge D_j$, which does not converge to the intended population probability (2.4). Therefore, Uno *et al.* (2011) proposed the IPCW-based C-index converging to (2.4). Uno's C-index at a spesific period $t \in \{1, \cdots, T_{max} - 1\}$ is defined as follows

$$\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t) = \frac{\sum_{i\neq j}^{n} I_{\{T_i<D_i\}} I_{\{S(t;\mathbb{Z}_i)<S(t;\mathbb{Z}_j)\}} I_{\{T_i<X_j, T_i<T_{max}\}} \widehat{G}_n^{-2}(T_i)}{\sum_{i\neq j}^{n} I_{\{T_i<D_i\}} I_{\{T_i<X_j, T_i<T_{max}\}} \widehat{G}_n^{-2}(T_i)},$$

where $\widehat{G}_n(T_i)$ is the predicted survival curve computed by KM estimator for censoring times based on n samples in test data, and T_{max} is a prespecified period such that $G(T_{max} - 1) > 0$. Uno's C-index is unbiased converging in probability as $n \to \infty$ to the population probability (Uno *et al.*, 2011) as follows

$$\mathbb{P}\left[S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j)\middle|T_i < T_j, T_i < T_{max}\right].$$

2.4.3 Time-Dependent Concordance

Time-dependent concordance was proposed by Antolini *et al.* (2005). It is based on time-dependent AUC (Heagerty & Zheng, 2005) and can be seen as an extension of Harrell's C-index explicitly designed for survival model with possibly

time-dependent covariates. Time-dependent AUC is a discrimination measure at a specific period t based on the probability of correctly classifying the individual outcomes conditional on the individual's status. Furthermore, the time-dependent concordance is defined as the weighted average of the time-dependent AUC (see the original paper by Antolini *et al.* (2005)) for more details about the measure derivation).

Following the notations used in this thesis, time-dependent concordance estimates

$$\begin{aligned} \mathrm{C} &= \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}\big] \\ &= \frac{\mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j), T_i < T_j, T_i < T_{max}\big]}{\mathbb{P}\big[T_i < T_j, T_i < T_{max}\big]}. \end{aligned} \tag{2.14}$$

The probabilities of interest (2.14) and (2.4) explain how well the order of any two predicted survival curves matches the order of their respective outcomes. If the model can correctly order the outputs, then it has good capability in discriminating the outputs. Note that there is a crucial difference between (2.14) and (2.4), namely the evaluated period of the survival curves. In the estimand of Harrell's C-index (2.4), we evaluate the (order) survival curves of $i \neq j$ at a fixed $t \in \mathcal{T}$. Meanwhile, (2.14), the survival curves are evaluated at the event times T_i. Hence, (2.14) is an overall measure instead of a time-specific measure as in (2.4). Note that (2.14) employs $S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)$ instead of $S(T_i;\mathbb{Z}_i) < S(T_j;\mathbb{Z}_j)$ for $(i \neq j)$. When the PH assumption is violated, the relationship between $S(T_i;\mathbb{Z}_i)$ and $S(T_j;\mathbb{Z}_j)$ can vary freely even though $T_i < T_j$. On the other hand, the relationship between $S(T_i;\mathbb{Z}_i)$ and $S(T_i;\mathbb{Z}_j)$ is always fixed regardless of the relationship between T_i and T_j.

The estimator of (2.14) can be derived by a direct approximation of the numerator and denominator of the second line of (2.14) in the presence of censoring as follows

$$\widehat{\mathrm{C}}_n = \frac{\sum_{i\neq j}^{n} I_{\{T_i<D_i\}} I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}} I_{\{T_i<X_j, T_i<T_{max}\}}}{\sum_{i\neq j}^{n} I_{\{T_i<D_i\}} I_{\{T_i<X_j, T_i<T_{max}\}}}. \tag{2.15}$$

In this thesis, we refer to $\widehat{\mathrm{C}}_n$ as a time-dependent concordance.

$\widehat{\mathrm{C}}_n$ is a direct adaptation of Antolini's measure proposed in Antolini *et al.* (2005) to discrete-time units and $T_i < T_{max}$. However, in their works, the convergence of their measure was not shown. In this thesis, we state that $\widehat{\mathrm{C}}_n$ still

depends on $T_i < D_i \wedge D_j$ and does not converge to (2.14) in Theorem 2.4.2, where the corresponding proof will be given in Section 2.4.4. In Chapter 4, we will amend $\widehat{\mathrm{C}}_n$ so that the correct convergence is obtained.

Theorem 2.4.2 (Convergence of Time-Dependent Concordance)**.** *Suppose the regularity conditions (R1-R4) in Section 2.1.3 hold. Then,*

$$\widehat{C}_n \xrightarrow{wp1} \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}, T_i < D_i \wedge D_j\big].$$

Time-dependent concordance is designed to evaluate the model performance based on predicted survival curves at event times and summarises the overall model performance. It provides a more flexible measure which can be used for more general survival models in which the PH assumption is not imposed. However, the time-dependent concordance is strongly related to Harrell's C-index since they are developed under the same principles except for the evaluated times and the PH assumption. Furthermore, time-dependent concordance may have the same disadvantages as Harrell's C-index, where one of its notorious drawbacks is that its bias increases as the censoring rate increases (Gerds *et al.*, 2013), where Uno's C-Index (Uno *et al.*, 2011) administered an IPCW approach to remove the bias in Harrell's C-index.

2.4.4 Proofs of The Convergence

In this section, we provide the proofs of Theorem 2.4.1 and Theorem 2.4.2. We do not discuss the proof of Uno's C-index in this thesis since Uno *et al.* (2011) had discussed it and $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$ follows their results. To this end, we first present the statistical theory of U-statistics and its convergence results. In the following definition and theorem, we do not present the original versions, but we adapt and extend their notations without losing their essence so that they can fit into our discussions throughout this thesis.

Definition 2.4.1 (U-statistic (Nolan & Pollard, 1987))**.** Let $x_1, x_2, \cdots$ be independent samples from a distribution F on $\mathscr{X}$, and $\mathscr{H}$ be a class of symmetric functions on $\mathscr{X} \otimes \mathscr{X}$. For $h \in \mathscr{H}$, we define

$$\zeta_n(h) = \sum_{1 \le i < j \le n} h(x_i, x_j). \tag{2.16}$$

If h is a symmetric function, $\zeta_n(h)/n(n-1)$ is its corresponding U-statistic. If it is not symmetric, we can replace h by its symmetric version $\bar{h}$ so that the corresponding U-statistic is defined by (Serfling, 1980)

$$\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \bar{h}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Theorem 2.4.3 (Theorem A on page 190 in the book of Serfling (1980))**.** *Let* $\frac{\zeta_n}{n(n-1)}$ *be the corresponding U-statistic for a symmeric kernel* $h \in \mathscr{H}$*, where* ζ_n *is defined by* (2.16)*. Let*

$$F \otimes F(h) = \iint h(\boldsymbol{x}_i, \boldsymbol{x}_j)\, dF(\boldsymbol{x}_i)\, dF(\boldsymbol{x}_j)$$

be the expected value of $h(\boldsymbol{x}_i, \boldsymbol{x}_j)$ *with respect to* $F \otimes F$*. If*

$$F \otimes F(|h|) < \infty,$$

then

$$\frac{\zeta_n}{n(n-1)} \xrightarrow{wp1} F \otimes F(h).$$

To prove the theorems, our approach is slightly different to the available literature, such as the works by Uno *et al.* (2011) and Gerds *et al.* (2013), because we adapt Harrell's C-index to evaluate the predicted survival curves instead of the predicted risk scores as in the original Harrell's C-index (Harrell *et al.*, 1982). As a consequence, (2.12) only applies to a single period $t \in \mathcal{T}$ and not to all periods except when the fitted survival models hold the PH assumption. The proof of Theorem 2.4.1 is given as follows

Proof of Theorem 2.4.1. We first rewrite the numerator of (2.12) as follows

$$\sum_{i \neq j}^{n} I_{\{S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j)\}} I_{\{T_i < T_j\}} I_{\{T_i < D_i\}} I_{\{T_i < D_j\}}.$$

We then denote $\boldsymbol{x}_i = (\mathbb{Z}_i, T_i, D_i)$ and $\boldsymbol{x}_j = (\mathbb{Z}_j, T_j, D_j)$ so that $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$ are independent samples from distribution F on $\mathscr{X}$. Let $\mathscr{H} = \{h\}$ be a class of functions on $\mathscr{X} \otimes \mathscr{X}$, where

$$h(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j)\}} I_{\{T_i < T_j\}} I_{\{T_i < D_i\}} I_{\{T_i < D_j\}}.$$

Because h is not symmetric, we need to symmetrise it as follows

$$\bar{h}\left(x_i, x_j\right) = \frac{1}{2}\left(h(x_i, x_j) + h(x_j, x_i)\right)$$

so that

$$\frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \bar{h}(x_i, x_j)$$

is the corresponding U-statistic with kernel $\bar{h}$, where $\bar{\mathscr{H}} = \{\bar{h} : (h+h)/2 : h \in \mathscr{H}\}$. We then denote

$$\bar{h}(x_i, x_j) = \frac{1}{2}\left(\mathscr{K}\left(x_i, x_j\right) + \mathscr{K}\left(x_j, x_i\right)\right),$$

where $\mathscr{K}\left(x_i, x_j\right) = I_{\{S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j)\}} I_{\{T_i < T_j\}} I_{\{T_i < D_i\}} I_{\{T_i < D_j\}}$. By the regularity conditions (R1-R4) in Section 2.1.3, we have

$$\begin{aligned}
&F \otimes F(|\bar{h}|) \\
&= \frac{1}{2}\left[F \otimes F\left(\left|I_{\{S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j)\}} I_{\{T_i < T_j\}} I_{\{T_i < D_i\}} I_{\{T_i < D_j\}}\right|\right)\right. \\
&\quad \left. + F \otimes F\left(\left|I_{\{S(t;\mathbb{Z}_j) < S(t;\mathbb{Z}_i)\}} I_{\{T_j < T_i\}} I_{\{T_j < D_j\}} I_{\{T_j < D_i\}}\right|\right)\right] \\
&= \frac{1}{2}\left[\mathbb{P}\left[S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j), T_i < T_j, T_i < D_i \wedge D_j\right]\right. \\
&\quad \left. + \mathbb{P}\left[S(t;\mathbb{Z}_j) < S(t;\mathbb{Z}_i), T_j < T_i, T_j < D_j \wedge D_i\right]\right] \\
&< \infty,
\end{aligned}$$

where the second equality holds because $I_{\{S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j)\}}$, $I_{\{T_i < T_j\}}$, $I_{\{T_i < D_i\}}$, and $I_{\{T_i < D_j\}}$ are independent events, and $\mathbb{E}_{F \otimes F}[I_{\{\varepsilon\}}] = \mathbb{P}[\varepsilon]$. Therefore, the condition of Theorem 2.4.3 is satisfied so that

$$\frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \bar{h}(x_i, x_j) \overset{wp1}{\to} F \otimes F(\bar{h}) \tag{2.17}$$

where

$$\begin{aligned}
&F \otimes F(\bar{h}) \\
&= \frac{1}{2}\left[\mathbb{P}\left[S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j), T_i < T_j, T_i < D_i \wedge D_j\right]\right. \\
&\quad \left. + \mathbb{P}\left[S(t;\mathbb{Z}_j) < S(t;\mathbb{Z}_i), T_j < T_i, T_j < D_i \wedge D_j\right]\right].
\end{aligned}$$

We now rewrite the denominator of (2.12) as follows

$$\sum_{i \neq j}^{n} I_{\{T_i < T_j\}} I_{\{T_i < D_i\}} I_{\{T_i < D_j\}}.$$

Because the only difference between the numerator and the denominator of Harrell's C-index is the indicator function $I_{\{S(t;\mathbb{Z}_i)<S(t;\mathbb{Z}_j)\}}$ whose values is either 1 or 0, then we can use the same approach and arguments used in obtaining the convergence in (2.17) to obtain

$$\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{\psi}(\boldsymbol{x}_i, \boldsymbol{x}_j) \overset{wp1}{\to} F \otimes F(\bar{\psi}), \tag{2.18}$$

where $\bar{\psi}$ is the symmetrised version of ψ,

$$\psi(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{T_i<T_j\}} I_{\{T_i<D_i\}} I_{\{T_i<D_j\}},$$

and

$$\begin{aligned} & F \otimes F(\bar{\psi}) \\ &= \frac{1}{2}\left[\mathbb{P}\left[T_i < T_j, T_i < D_i \wedge D_j\right]\right. \\ &\quad \left. + \mathbb{P}\left[T_j < T_i, T_j < D_i \wedge D_j\right]\right]. \end{aligned}$$

Since we have shown the convergence of the standardised U-statistics with kernels $\bar{h}$ and $\bar{\psi}$ in (2.17) and (2.18) , respectively, then the convergence of the standardised U-statistics with their respective asymmetric kernels also holds. That is,

$$n(n-1)^{-1} \sum_{1\le i<j\le n} h(\boldsymbol{x}_i, \boldsymbol{x}_j) \overset{wp1}{\to} F \otimes F(h), \tag{2.19}$$

and

$$n(n-1)^{-1} \sum_{1\le i<j\le n} \psi(\boldsymbol{x}_i, \boldsymbol{x}_j) \overset{wp1}{\to} F \otimes F(\psi). \tag{2.20}$$

Dividing the left hand side of (2.19) by the left hand side of (2.20), and by applying the Continuous Mapping Theorem (Shao, 2003), we have

$$\begin{aligned} & \frac{\sum_{1\le i<j\le n} h(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sum_{1\le i<j\le n} \psi(\boldsymbol{x}_i, \boldsymbol{x}_j)} \\ & \overset{wp1}{\to} \frac{F \otimes F(h)}{F \otimes F(\psi)} \\ &= \frac{F \otimes F\left(I_{\{S(t;\mathbb{Z}_i)<S(t;\mathbb{Z}_j)\}} I_{\{T_i<T_j\}} I_{\{T_i<D_i\}} I_{\{T_i<D_j\}}\right)}{F \otimes F\left(I_{\{T_i<T_j\}} I_{\{T_i<D_i\}} I_{\{T_i<D_j\}}\right)} \\ &= \mathbb{P}\left[S(t;\mathbb{Z}_i) < S(t;\mathbb{Z}_j) \middle| T_i < T_j, T_i < D_i \wedge D_j\right], \end{aligned}$$

which confirms the convergence of $\widehat{\mathrm{C}}_n^{\mathrm{har}}(t)$. □

To prove Theorem 2.4.2, we can easily adapt the proof of Theorem 2.4.1 into the context of Theorem 2.4.2. The proof is presented as follows

Proof of Theorem 2.4.2. We rewrite the numerator of (2.15) as follows

$$\sum_{i\neq j}^{n} I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}} I_{\{T_i<T_j, T_i<T_{max}\}} I_{\{T_i<D_i\}} I_{\{T_i<D_j\}}.$$

By the same assumption and arguments in showing the convergence (2.17) of the numerator $\widehat{\mathrm{C}}_n^{\mathrm{har}}(t)$, we obtain

$$\frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \bar{h}(\boldsymbol{x}_i, \boldsymbol{x}_j) \overset{wp1}{\to} F\otimes F(\bar{h}) \tag{2.21}$$

where $\bar{h}$ is the symmetrised $h(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}} I_{\{T_i<T_j\}} I_{\{T_i<D_i\}} I_{\{T_i<D_j\}}$ for $\boldsymbol{x}_i = (\mathbb{Z}_i, T_i, D_i)$ and $\boldsymbol{x}_j = (\mathbb{Z}_j, T_j, D_j)$ so that $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$ are independent samples from distribution F on $\mathscr{X}$, $\mathscr{H} = \{h\}$ is a class of functions on $\mathscr{X}\otimes\mathscr{X}$, and

$$\begin{aligned}
&F\otimes F(\bar{h}) \\
&= \frac{1}{2}\left[\mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j), T_i < T_j, T_i < T_{max}, T_i < D_i \wedge D_j\big]\right. \\
&\quad \left. + \mathbb{P}\big[S(T_i;\mathbb{Z}_j) < S(T_i;\mathbb{Z}_i), T_j < T_i, T_j < T_{max}, T_j < D_i \wedge D_j\big]\right].
\end{aligned}$$

The denominator of time-dependent concordance is the same as the denominator of Harrell's C-index except we now have $I_{\{T_i<T_{max}\}}$. Thus, by adapting the convergence of the denominator of $\widehat{\mathrm{C}}_n^{\mathrm{har}}(t)$ in Theorem 2.4.1, we obtain

$$\frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \bar{\psi}(\boldsymbol{x}_i, \boldsymbol{x}_j) \overset{wp1}{\to} F\otimes F(\bar{\psi}), \tag{2.22}$$

where $\bar{\psi}$ is the symmetrised version of

$$\psi(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{T_i<T_j, T_i<T_{max}\}} I_{\{T_i<D_i\}} I_{\{T_i<D_j\}},$$

and

$$\begin{aligned}
&F\otimes F(\bar{\psi}) \\
&= \frac{1}{2}\left[\mathbb{P}\big[T_i < T_j, T_i < T_{max}, T_i < D_i \wedge D_j\big]\right. \\
&\quad \left. + \mathbb{P}\big[T_j < T_i, T_j < T_{max}, T_j < D_i \wedge D_j\big]\right].
\end{aligned}$$

Because in (2.21) and (2.22) we have shown the convergence of the standardised U-statistics with kernels $\bar{h}$ and $\bar{\psi}$, respectively, then we also have

$$n(n-1)^{-1} \sum_{1 \leq i < j \leq n} h(\boldsymbol{x}_i, \boldsymbol{x}_j) \overset{wp1}{\to} F \otimes F(h), \tag{2.23}$$

and

$$n(n-1)^{-1} \sum_{1 \leq i < j \leq n} \psi(\boldsymbol{x}_i, \boldsymbol{x}_j) \overset{wp1}{\to} F \otimes F(\psi). \tag{2.24}$$

Dividing the left hand side of (2.23) by the left hand side of (2.24), and by applying the Continuous Mapping Theorem (Shao, 2003), we finally have

$$\begin{aligned}
&\frac{\sum_{1 \leq i < j \leq n} h(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sum_{1 \leq i < j \leq n} \psi(\boldsymbol{x}_i, \boldsymbol{x}_j)} \\
&\overset{wp1}{\to} \frac{F \otimes F(h)}{F \otimes F(\psi)} \\
&= \frac{F \otimes F\left(I_{\{S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\}} I_{\{T_i < T_j, T_i < T_{max}\}} I_{\{T_i < D_i\}} I_{\{T_i < D_j\}}\right)}{F \otimes F\left(I_{\{T_i < T_j, T_i < T_{max}\}} I_{\{T_i < D_i\}} I_{\{T_i < D_j\}}\right)} \\
&= \mathbb{P}\left[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j) \middle| T_i < T_j, T_i < T_{max}, T_i < D_i \wedge D_j\right],
\end{aligned}$$

which concludes the convergence of $\widehat{\mathrm{C}}_n$ to a probability that still depends on censoring distribution $T_i < D_i \wedge D_j$. □

In this chapter, we have presented basic notations and some fundamental concepts in survival analysis, including the predicted survival curves, Nnet-survival, and two types of performance measures, namely calibration and discrimination, that will be used throughout this thesis. We have adapted the standard proposed measures to discrete-time units and have shown the convergences of Harrell's C-index and time-dependent concordance. In the subsequent chapter, we will discuss some weaknesses of the integrated Brier score and then provide potential solutions to such drawbacks. Furthermore, we will also propose several modified versions of the integrated Brier score.

Chapter 3

The Integrated Brier Score's Pitfalls and Modified Integrated Brier Scores

One measure for evaluating survival model prediction capability is calibration, which assesses how close the model's outputs, i.e., the predicted survival curves, are to their respective model's outcomes. As calibration does not depend on the assumptions of the fitted models, they can be used properly on PH or non-PH data. One prominent example of calibration measures is integrated Brier score as introduced in Chapter 2.

This chapter will first discuss several potential pitfalls of the integrated Brier score. Then, we use the integrated Brier score for KM estimator as the reference value to cope with such drawbacks. We finally introduce some alternative versions of integrated Brier score, namely normalised integrated Brier score, centered integrated Brier score, and normalised centered integrated Brier score. In particular, we would show how the integrated Brier score and its modified versions behave in good and overfitted models.

3.1 Pitfalls of Integrated Brier Score

In this section, we will demonstrate the pitfalls of the integrated Brier score via case studies and real-world examples. We rigorously evaluated the model predic-

tive performance of Nnet-survival on train or test data. Our simulation presents two scenarios, each with its unique setup. In the first scenario, we have one discretisation setup for the train data and three for the test data. This scenario aims to show that the integrated Brier score depends on the chosen discretisation setups in the test data. The second scenario increases the complexity with three discretisation setups for train and test data. Its objective is to demonstrate that the values of the integrated Brier score may confuse us in evaluating the performance of two or more models.

In the real-world examples, we apply the integrated Brier score to evaluate the model performance in TCGA data and breast cancer data, showcasing the relevance and importance of our findings. The objective of our numerical experiment in the TCGA data is the same as the first scenario in Simulation 1. Meanwhile, in the breast cancer data, we would show that the integrated Brier score highly depends on data structure where almost all individuals are censored so that the predicted survival curves are always close to one regardless of the fitted models.

In most settings of the simulation studies and the real data examples of this chapter, we will train models with different qualities, namely the non-overfitted (good), overfitted, and underfitted (e.g. high regularisation values) models. Then, our proposed measures will be applied to evaluate those models' performance. This scenario aims to assess whether our proposed measures align with the qualities of the models.

3.1.1 Simulation 1: PH Data

In this simulation study, we generated PH data based on the CPH model and employed two simulation scenarios. For each scenario, we differentiated the good and bad models based on whether they overfit the data. To obtain the simulated data, we first generated right-censored event times by inverting the cumulative hazard function of the CPH model (Bender *et al.*, 2005), where the baseline hazard rate follows Gompertz distribution with scale parameter (α) and shape parameter (γ). In particular, the event times were generated using the following formula:

$$T = \frac{1}{\alpha} \log \left(1 - \frac{\alpha \log(v)}{\gamma \exp\left(\boldsymbol{\beta}\mathbb{Z}\right)} \right), \tag{3.1}$$

where $v \sim U(0,1)$, and $\boldsymbol{\beta} = [\beta_1 \cdots \beta_5]$ is coefficient vector of covariate vector $\mathbb{Z} = [Z_1 \cdots Z_5]'$. A total of 1000 independent event times were generated based on (3.1) with $\alpha = 0.01$ and $\gamma = 0.05$. For the CPH model, we used five covariates observed at baseline: three were generated from Bernoulli distributions (i.e. $Z_1 \sim \text{Ber}(1, 0.1), Z_2 \sim \text{Ber}(1, 0.5), Z_3 \sim \text{Ber}(1, 0.3)$), and two were from normal distributions (i.e. $Z_4 \sim \mathcal{N}(0,1)$ and $Z_5 \sim \mathcal{N}(0, 0.5)$), where $\beta_1 = 3, \beta_2 = 0.5, \beta_3 = 0.8, \beta_4 = 0.25$, and $\beta_5 = 0.95$ were their respective effect sizes.

The follow-up ended at $T^* = 70$ such that all individuals with event times greater than or equal to T^* were administratively censored. Denote by n_{surv} the number of individuals surviving until T^*. To obtain censoring within $[0, T^*)$, $(1000 - n_{\text{surv}})$ independent right censored times were generated from Weibull distribution with three parameters (shape, scale, and location). The values of these parameters were fine-tuned to achieve the desired proportions of censored individuals. However, in this study, we set the censoring rates close to 0% (i.e. in the range 0.01%-0.017%) to minimise the $\widehat{G}_n$ effects. Then, the $(1000 - n_{\text{surv}})$ observed times X in $[0, T^*)$ were obtained by taking the minimum between the survival times and the censoring times. For individuals with survival times greater than or equal to T^*, the observed times would equal T^*. Finally, we discretised all observed times into T_{max} periods, where all individuals in T_{max} were administratively censored, and their observed times were set to equal T_{max}.

In this work, we only fitted the models to a single fixed train data ($n_{\text{train}} = 1000$). Then, we evaluated the model predictive performance on 100 independent test data ($n_{\text{test}} = 1000$) generated with the same setting as the training data. We developed two main models that were differentiated by their fitting quality to the train data, namely good and overfitted models. The good model was obtained by fitting the good Nnet-survival architecture in Table B.1 (Appendix B.1) to the train data containing the discretised observed times, the status, and $\mathbb{Z}_i = [Z_{i1} \cdots Z_{i5}]'$ for $i = 1, \cdots, 1000$. The overfitted model was developed by incorporating more additional covariates $\bar{\mathbb{Z}}_i = [\bar{Z}_{i1} \cdots \bar{Z}_{i5}]'$. For each $p (p = 1, \cdots, 5)$, covariate $\bar{Z}_{ip}$ was obtained by randomly permuting Z_{ip} over the 1000 individuals so that Z_{ip} had random relationship with the observed time X_i. We fitted the overfitted Nnet-survival architecture in Table B.1 (Appendix B.1)

to the train data containing the discretised observed times X_i, the status, and $(\mathbb{Z}_i \setminus Z_{i1}) \cup \bar{\mathbb{Z}}$. Since Z_{i1} has the highest influence on T_i, indicated by $\beta_1 = 3$, we removed it from the analysis to increase the overfitting.

Scenario 1: Performance Evaluation of A Fixed Model

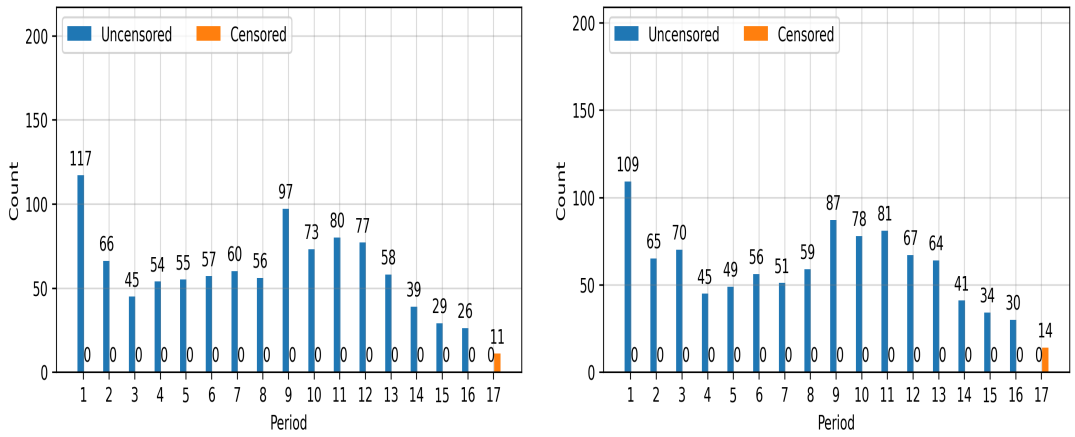
The main objective of this simulation is to show how the integrated Brier score behaves in assessing the predictive performance of a fixed model evaluated on various discretisation setups. In particular, we first discretised the generated observed times in the train data based on a set of discretisation points $\mathscr{D}_1$={0, 2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20, 25, 30, 35, 40, 45, 50, 55, 65, ∞}. We fitted both the good and overfitted models in Table B.1 (Appendix B.1) to their respective fixed train data (n_{train}=1000). Then, we evaluated their model performance on their respective test data (n_{test}=1000) discretised by using $\mathscr{D}_1$, $\mathscr{D}_2$={0, 2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20, 25, 30,∞}, and $\mathscr{D}_3$={0, 2.5, 5,∞}. All sets of discretisation points are dependent amongst others, that is $\mathscr{D}_3 \subset \mathscr{D}_2 \subset \mathscr{D}_1$ so that we can obtain the prediction on $\mathscr{D}_2$ and $\mathscr{D}_3$ from the results of $\mathscr{D}_1$.

The distribution of the train data and the first test data after the discretisation process can be seen in Figure 3.1. Panels (a) and (b) show the distributions of the train and test data discretised by $\mathscr{D}_1$. Meanwhile, panels (c) and (d) display the distribution of the first test data discretised by $\mathscr{D}_2$ and $\mathscr{D}_3$, respectively. Regardless of the last period T_{max}, their distributions are right-skewed following the distribution obtained by $\mathscr{D}_1$. The censoring rate for train data is close to zero, namely 1.1%. Meanwhile, the censoring rates in the test data are about 1.4%, 33.1%, and 82.6% with $\mathscr{D}_1$, $\mathscr{D}_2$, and $\mathscr{D}_3$, respectively. All the censoring has occurred in the last period only T_{max} due to the end of the study (administrative censoring).

To find out whether our model indicates overfitting or not before carrying out the model evaluation, we can conduct prediction on both train and test data, and we compare the difference between their model's outputs $S(t; \mathbb{Z}_i)$ and their model's outcomes $I_{\{T_i > t\}}$. If such a difference in the train data is much smaller than in the test data, the models have overfitted the data. Figure 3.2 contains six panels (a-f), where each panel show the individuals' predicted survival curves

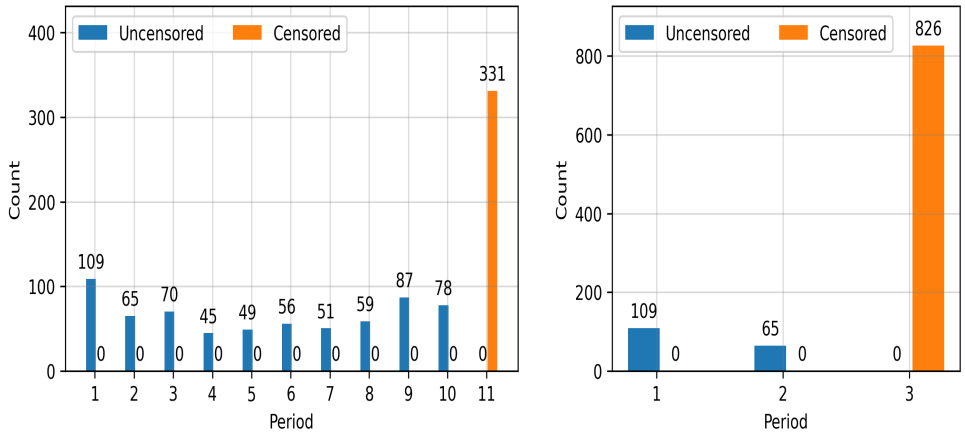
over the follow-up $\{1, \cdots, T_{max} - 1\}$ computed on train or test data for a specific discretisation setup (i.e. $\mathscr{D}_1$, $\mathscr{D}_2$, or $\mathscr{D}_3$). Each panel contains two sub-panels, where the left and right sub-panels are for individuals with the model's outcomes $I_{\{T_i > t\}} = 0$ and $I_{\{T_i > t\}} = 1$, respectively. As we can see from Figure 3.2, the shapes of the predicted survival curves (outputs) from the test data and train data for all discretisation setups (panels a-f) in the good models are almost similar, indicating that the models do not overfit the train data. On the other hand, the output shapes from the test and train data in Figure 3.3 are dissimilar. The output shapes from the train data are very close to the outcomes, indicating that the models have overfitted the data. Meanwhile, the outputs from the test data spread more evenly over $[0, 1]$.

The main results of this simulation scenario can be seen in Figure 3.4. It is clear from the figure that $\widehat{\text{IBS}}_n$ can differentiate between the good and overfitted models. For the same discretisation setup in test data, $\widehat{\text{IBS}}_n$ in the overfitted model (panel (b)) are much higher than in the good model (panel (a)), as expected. However, from the results, we can see a drawback of the integrated Brier score. Although we fitted a single fixed model to the train data by applying one discretisation setup $\mathscr{D}_1$, we obtained various values of $\widehat{\text{IBS}}_n$ that are significantly different amongst them depending on the used discretisation setups in the test data. In the beginning, $\widehat{\text{IBS}}_n$ are around 0.12 for $\mathscr{D}_1$. After that, they increase to around 0.15 for $\mathscr{D}_2$, then decrease drastically to less than 0.1 for $\mathscr{D}_3$. This situation is because $\widehat{\text{IBS}}_n$ from $\mathscr{D}_2$ and $\mathscr{D}_3$ are the (unweighted) average over the first ten $\widehat{\text{BS}}_n(t)$ and the first two $\widehat{\text{BS}}_n(t)$ of $\mathscr{D}_1$, respectively (Figure 3.5). As the values of $\widehat{\text{BS}}_n(t)$ oscillate over the follow-up, depending on how close the outputs and the outcomes at each period, $\widehat{\text{IBS}}_n$ can be easily manipulated by defining the discretisation points in the test data. Since our proposed integrated Brier score is defined as the (unweighted) average of Brier scores over the evaluated periods $\{1, \cdots, T_{max} - 1\}$, it may significantly vary depending on the Brier score at each period $t \in \{1, \cdots, T_{max} - 1\}$.



(a) $\mathcal{D}_1$ (Train Data)

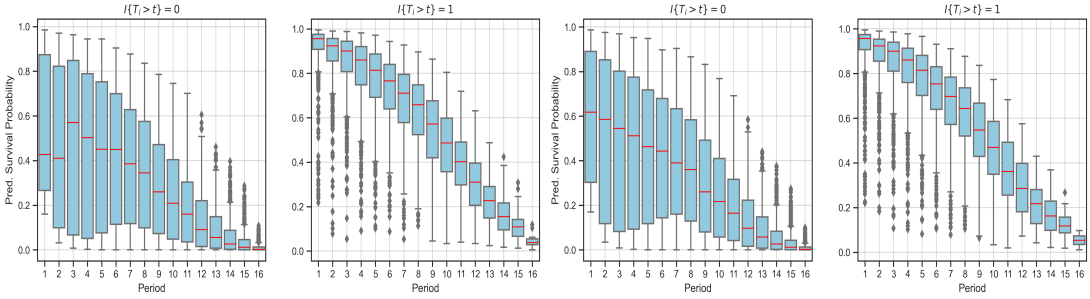
(b) $\mathcal{D}_1$ (Test Data)



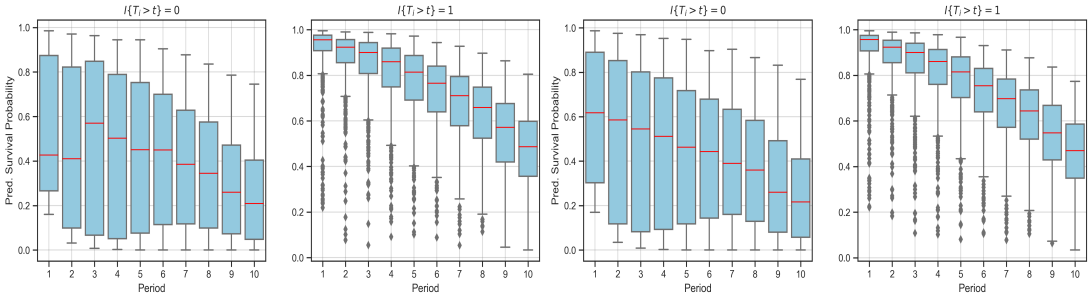
(c) $\mathcal{D}_2$ (Test Data)

(d) $\mathcal{D}_3$ (Test Data)

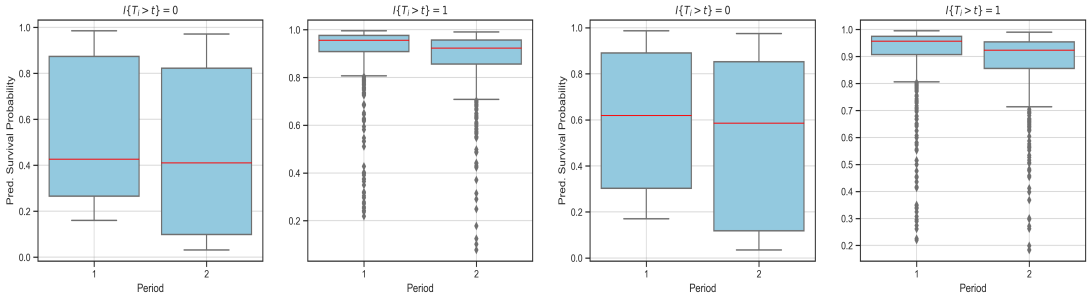
Figure 3.1: Distributions of the simulated train data and test data in Scenario 1 of Simulation 1 for each discretisation setup over $\mathcal{T}$, where all individuals in T_{max} were administratively censored.



(a) $\mathscr{D}_1$ (test data) (b) $\mathscr{D}_1$ (train data)

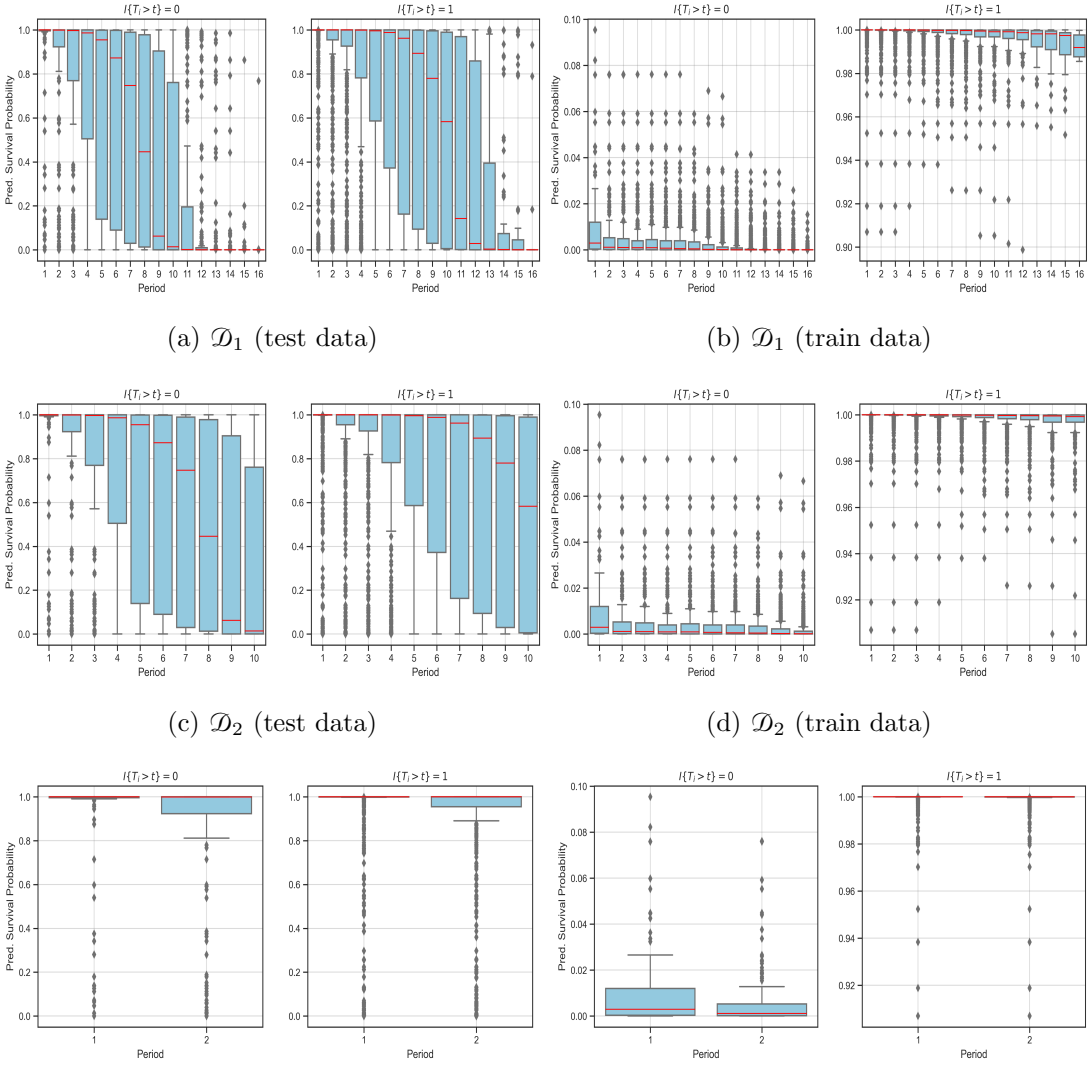


(c) $\mathscr{D}_2$ (test data) (d) $\mathscr{D}_2$ (train data)



(e) $\mathscr{D}_3$ (test data) (f) $\mathscr{D}_3$ (train data)

Figure 3.2: The predicted survival curves of each individual $i(i = 1, \cdots, 1000)$ categorised by $I_{\{T_i>t\}}$ over $\{1, \cdots, T_{max} - 1\}$ in Scenario 1 of Simulation 1. The good Nnet-survival architecture fitted to a fixed train data ($n_{\text{train}} = 1000$). Then, the predictions were conducted on the first test data ($n_{\text{test}} = 1000$) and the train data.



(a) $\mathscr{D}_1$ (test data)

(b) $\mathscr{D}_1$ (train data)

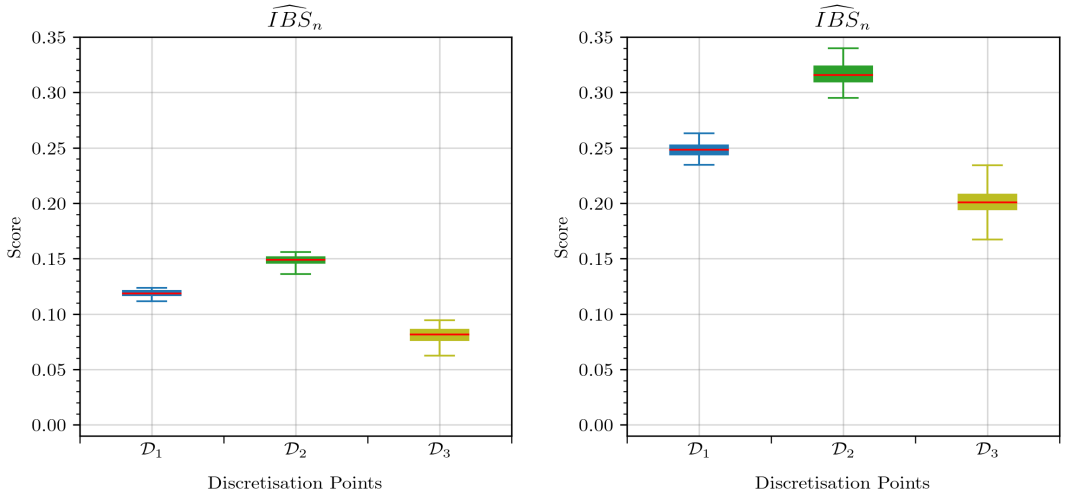
(c) $\mathscr{D}_2$ (test data)

(d) $\mathscr{D}_2$ (train data)

(e) $\mathscr{D}_3$ (test data)

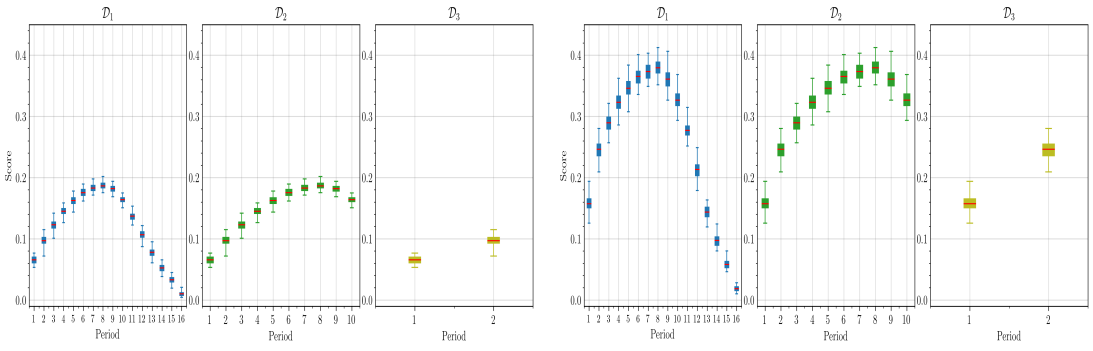
(f) $\mathscr{D}_3$ (train data)

Figure 3.3: The predicted survival curves of each individual $i(i = 1, \cdots, 1000)$ categorised by $I_{\{T_i>t\}}$ over $\{1, \cdots, T_{max} - 1\}$ in Scenario 1 of Simulation 1 . The overfitted Nnet-survival architecture fitted to a fixed train data ($n_{\text{train}} = 1000$). Then, the predictions were conducted on the first test data ($n_{\text{test}} = 1000$) and the train data.



(a) Good Model (b) Overfitted Model

Figure 3.4: Integrated Brier Score over $\{1, \cdots, T_{max} - 1\}$ of the good model (a) and the overfitted model (b) for each discretisation setup in the test data in Scenario 1 of Simulation 1. They were estimated on 100 independent test data (n_{test}=1000) from models fitted to a single fixed train data (n_{train}=1000).



(a) Good Model (b) Overfitted Model

Figure 3.5: Brier Score at each $t \in \{1, \cdots, T_{max} - 1\}$ of (a) the good model and (b) the overfitted model over the three sets of discretisation setups for the test data in Scenario 1 of Simulation 1. They were estimated on 100 independent test data (n_{test}=1000) from the models fitted to a single fixed train data (n_{train}=1000).

Scenario 2: : Performance Evaluation of Different Models

The main objective of this simulation scenario is to study how the integrated Brier score behaves in assessing different models fitted to a single fixed train data. We can obtain various models from Nnet-survival by tuning hyper-parameters, such as regularisation, number of hidden layers, and number of nodes in the hidden layer. However, in this scenario, varying discretisation setup was our method for obtaining different models. In particular, we used three sets of discretisation points, namely $\mathscr{D}_4$=$\{0, 2.5, 5, 55, \infty\}$, $\mathscr{D}_5$=$\{0, 55, 60, 65, 70, \infty\}$, and $\mathscr{D}_6$=$\{0, 5, 7.5, 10, 12.5, 15, 17.5, 20, 40, 45, 65, \infty\}$. These sets were used to discretise the training and testing data. We selected these three discretisation setups to obtain different event times distribution with various numbers of periods.

Figure 3.6 shows the distribution of the generated train data for each discretisation setup. We can see from the figure that a small proportion of censoring only exists in the last period T_{max} due to administrative censoring. The event time distribution in $\mathscr{D}_4$ is left-skewed. In $\mathscr{D}_5$, the event time distribution is right-skewed, where most lie within the first period. The event times distribution obtained from $\mathscr{D}_6$ is not as extreme as in $\mathscr{D}_4$ and $\mathscr{D}_5$ although most of them have occurred in the right-tail of the follow-up time. The identical discretisation setups are also applied to the test data, resulting in almost identical event time distributions and censoring rates as in the train data.

We fitted the good and overfitted Nnet-survival architectures in Table B.1 (Appendix B.1) to a fixed train data discretised by $\mathscr{D}_4, \mathscr{D}_5$, and $\mathscr{D}_6$. Then, we computed the predicted survival curves from the train and test data for each discretisation setup. We finally evaluated the model's predictive performance from the train and test data using an integrated Brier score. The prediction on the train data was conducted only once. However, we repeated the prediction procedure in test data 100 times for 100 independently generated test data.

Figure 3.7 and Figure 3.8 show $S(t; \mathbb{Z}_i)$ for each $i(i = 1, \cdots, 1000)$ over $\{1, \cdots, T_{max} - 1\}$ based on the first test data and the single train data for each discretisation setup. In Figure 3.7, the outputs of the good model in the train and the test data are almost the same, indicating that the models do not overfit the data. On the other hand, the outputs from the train data for each discretisation

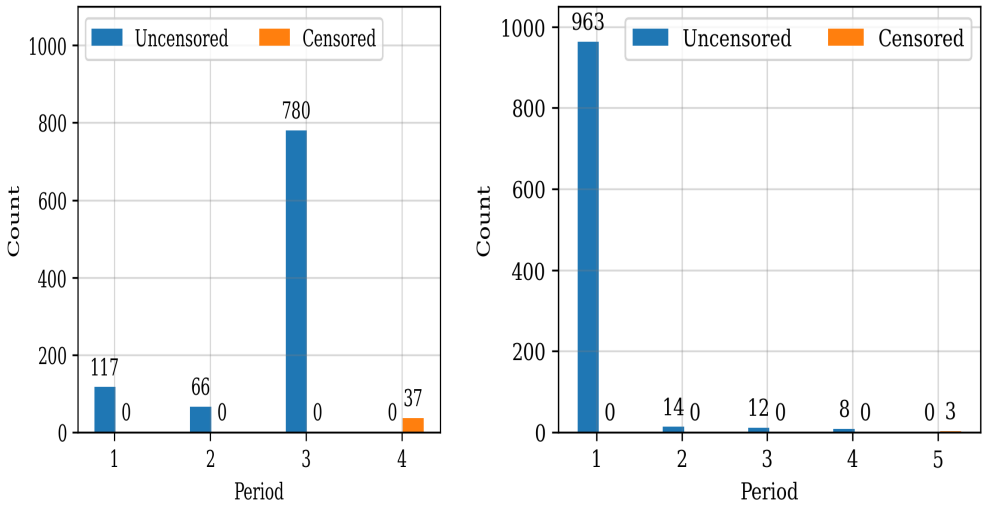
setup (i.e. panels (b), (d), and (f) in Figure 3.8 are very close to their respective outcomes. Meanwhile, panels (a), (c), and (e) in Figure 3.8 show that the outputs from the test data are not close to the outcomes. Most of the outputs in $\mathscr{D}_4$ and $\mathscr{D}_5$ oppose their outcomes when $I_{\{T_i>t\}} = 0$ and $I_{\{T_i>t\}} = 1$, respectively. In addition, the outputs from $\mathscr{D}_6$ in the test data are mostly spread out over $[0, 1]$. In contrast, the respective outputs in the train data are so close to their outcomes. The results from Figure 3.8 clearly demonstrated that the models have overfitted the data.

The results of the model evaluation of the good and the overfitted Nnet-survival using integrated Brier score for each discretisation setup are displayed in Figure 3.9. Although in $\mathscr{D}_4$ and $\mathscr{D}_6$ the differences between $\widehat{\text{IBS}}_n$ in the good and the overfitted models are significant, their differences in $\mathscr{D}_5$ are minimal. The values of $\widehat{\text{IBS}}_n$ in $\mathscr{D}_5$ for the good and overfitted models are between 0.02-0.04. These results are not as expected since the $\widehat{\text{IBS}}_n$ values below 0.05 are usually considered to have good model performance while we have the overfitted models. This result is because in $\mathscr{D}_5$ most event times occur in the first period. As a consequence, the predicted survival curves are mainly close to zero (see panels (c) and (d) of Figure 3.7 and Figure 3.8). Even though in the overfitted models, most predicted survival curves are close to one when $I_{\{T_i>t\}} = 0$ (see panel (d)), they are only a small number compared to $I_{\{T_i>t\}} = 1$. For $\mathscr{D}_6$, $\widehat{\text{IBS}}_n$ (i.e. around 0.13) in the good models are close to $\widehat{\text{IBS}}_n$ from $\mathscr{D}_4$ (i.e. around 0.15) in the overfitted model. This result may be confusing because it is difficult to differentiate two models with $\widehat{\text{IBS}}_n$ values equal to 0.13 and 0.15.

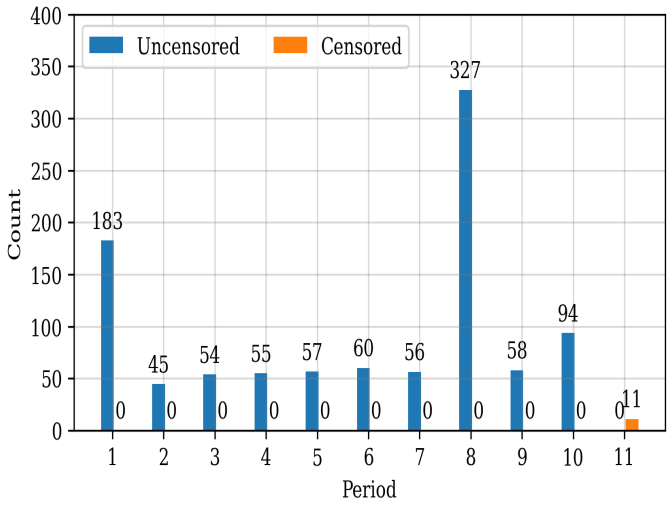
The $\widehat{\text{IBS}}_n$ in Figure 3.9 are obtained by the respective $\widehat{\text{BS}}_n(t)$ in Figure 3.10. $\widehat{\text{IBS}}_n$ is highly determined by the distribution of $\widehat{\text{BS}}_n(t)$ over the follow-up. We can obtain various models by varying the discretisation setups in train data. The distribution of event times obtained by the discretisation process dictates how the predictions over the follow-up behave. For example, if we see $\mathscr{D}_4$ in panel (a) of Figure 3.10, $\widehat{\text{BS}}_n(t)$ slightly increase from $t = 1$ to $t = 2$, before drastically fall at $t = 3$. In the train data distribution of $\mathscr{D}_4$ in Figure 3.6, the number of event times slightly down from the first period (117 events) to the second period (66 events) resulting in the prediction at $t = 2$ less stable as shown by larger standard deviation of the boxplot at $t = 2$ (panels (a) and (b) in Figure 3.7). However,

Figure 3.6 shows that the number of events increases significantly from the second period (66 events) to the third period (780 events). The predictions at $t = 3$ are much more stable and very close to their outcomes, especially when $I_{\{T_i>3\}} = 0$ (panels (a) and (b) in Figure 3.7). Although the predictions for $I_{\{T_i>3\}} = 1$ are mostly wrong, these errors are only a small proportion, that is 37/(37+780) = 0.045, so that $\widehat{BS}_n(3)$ are between 0.04-0.05 (panel (a) in Figure 3.7).

In summary, our experiments have revealed a crucial finding: the data structure of the train data plays a pivotal role in determining the integrated Brier score. This result means that a change in the event times distribution can significantly alter the values of the respective Brier scores. As a result, even in a well-performing model (e.g., a non-overfitted model), we may observe a relatively high integrated Brier score, making it challenging to assess the model's predictive performance accurately. In the next section, we will study how $\widehat{\text{IBS}}_n$ behaves in two real-world data, namely TCGA data and breast cancer data.

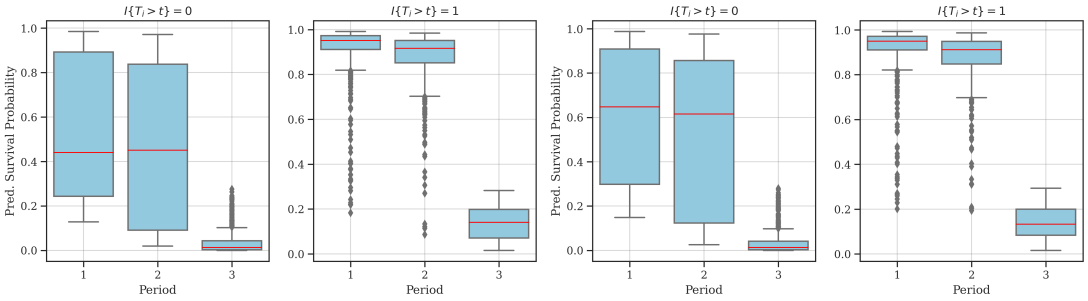


(a) $\mathcal{D}_4$ (b) $\mathcal{D}_5$

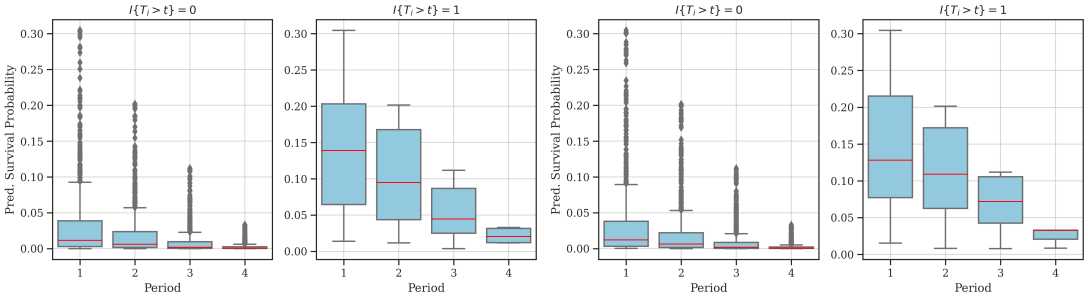


(c) $\mathcal{D}_6$

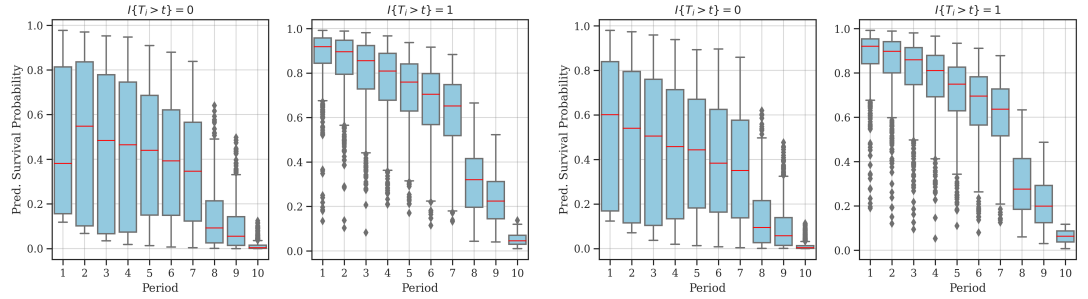
Figure 3.6: Distribution of the simulated train data in Scenario 2 of Simulation 1 for each discretisation setup over $\mathcal{T}$, where all individuals in T_{max} were administratively censored.



(a) $\mathcal{D}_4$ (test data) (b) $\mathcal{D}_4$ (train data)

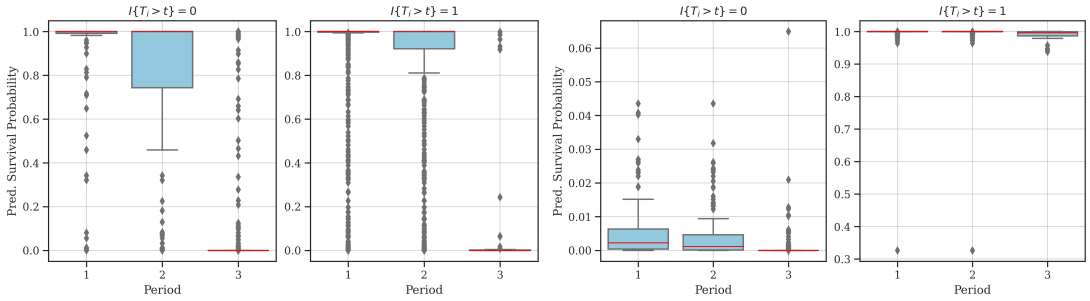


(c) $\mathcal{D}_5$ (test data) (d) $\mathcal{D}_5$ (train data)

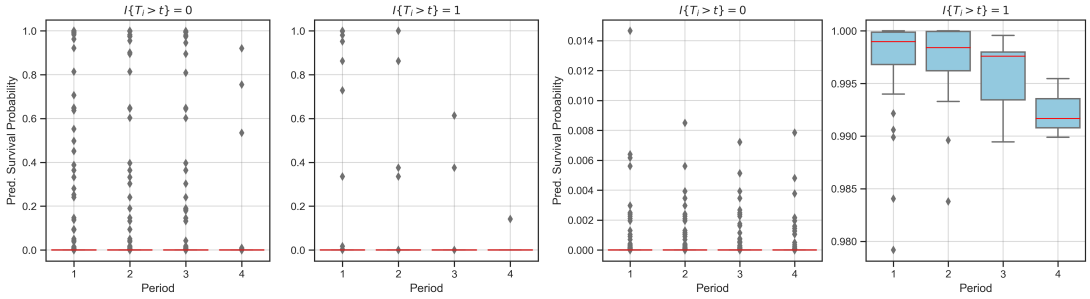


(e) $\mathcal{D}_6$ (test data) (f) $\mathcal{D}_6$ (train data)

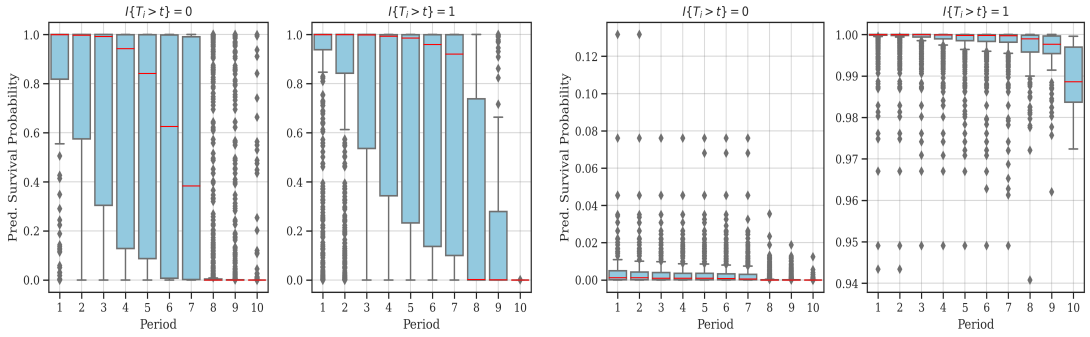
Figure 3.7: The predicted survival curves of each individual $i(i = 1, \cdots, 1000)$ categorised by $I_{\{T_i>t\}}$ over $\{1, \cdots, T_{max} - 1\}$ in Scenario 2 of Simulation 1. The good Nnet-survival architecture fitted to a fixed train data ($n_{\text{train}} = 1000$). Then, the predictions were conducted on the first test data ($n_{\text{test}} = 1000$) and the train data.



(a) $\mathcal{D}_4$ (test data) (b) $\mathcal{D}_4$ (train data)

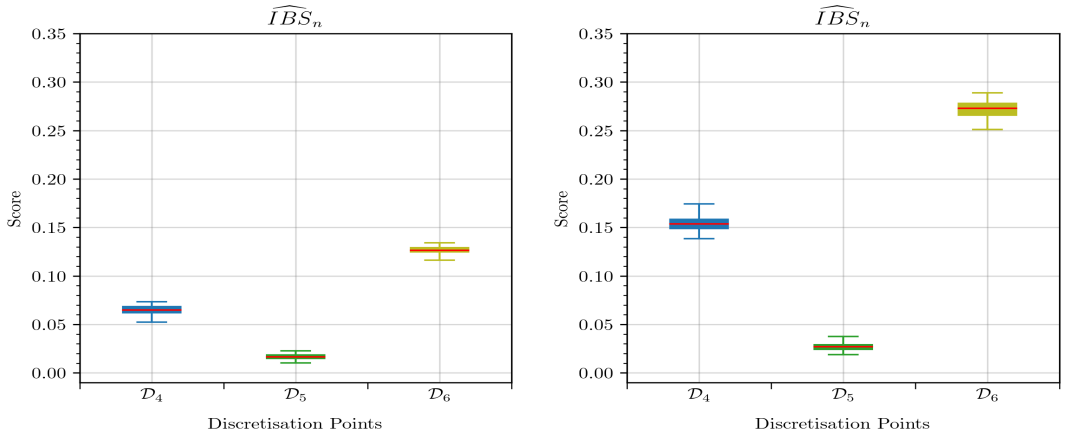


(c) $\mathcal{D}_5$ (test data) (d) $\mathcal{D}_5$ (train data)



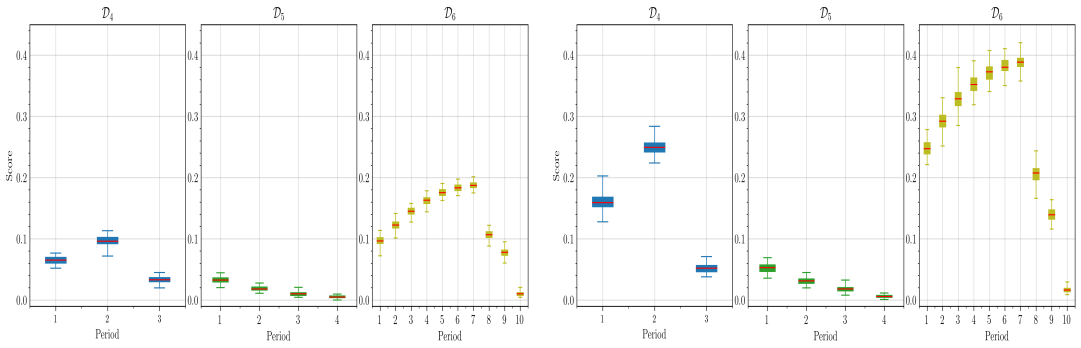
(e) $\mathcal{D}_6$ (test data) (f) $\mathcal{D}_6$ (train data)

Figure 3.8: The predicted survival curves of each individual $i (i = 1, \cdots, 1000)$ categorised by $I_{\{T_i > t\}}$ over $\{1, \cdots, T_{max} - 1\}$ in Scenario 2 of Simulation 1. The overfitted Nnet-survival architecture fitted to a fixed train data ($n_{\text{train}} = 1000$). Then, the predictions were conducted on the first test data ($n_{\text{test}} = 1000$) and the train data.



(a) Good Models (b) Overfitted Models

Figure 3.9: Integrated Brier score over $\{1, \cdots, T_{max} - 1\}$ of the good model (a) and the overfitted model (b) for each discretisation setup in the test data in Scenario 2 of Simulation 1. They were estimated on 100 independent test data (n_{test}=1000) from models fitted to a single fixed train data (n_{train}=1000).



(a) Good Models (b) Overfitted Models

Figure 3.10: Brier score at each $t \in \{1, \cdots, T_{max} - 1\}$ of the good model (a) and the overfitted model (b) over the three sets of discretisation setups for the test data in Scenario 2 of Simulation 1. They were estimated on 100 independent test data (n_{test}=1000) from models fitted to a single fixed train data (n_{train}=1000).

3.1.2 Real-World Examples

This section will explore the pitfalls of the integrated Brier score through real-world examples using two real data, namely TCGA data and breast cancer data.

TCGA Data

TCGA mutation data, or simply TCGA data, was a part of TCGA project (Weinstein *et al.*, 2013) and was published by Kandoth *et al.* (2013). We extracted the data from R package **dnet** (Fang & Gough, 2014), containing 12 major cancer types for cancer patients with their respective time to death within the study period. The data contains clinical information of 3,096 cancer patients with 19,428 baseline variables, including the time until the death of the patients (in days) during the study. The follow-up was defined from zero up to the largest recorded observed time (6,975 days), where 2,099 (around 68%) patients were censored, and 997 (around 32%) patients died.

Most variables are categorical except 'Age' and 'time', which are continuous variables. We first checked the number of missing values within each variable. Because variables 'Tumor stage' and 'Tumor grade' contain too many missing values (i.e. 526 and 1760, respectively), we excluded them from further analysis. There were no missing values in the variable 'time' as the observed time for the event of interest (death). We imputed the missing values in the variable 'Age' using the mean imputation method (Jadhav *et al.*, 2019), and standardised the result. Meanwhile, we imputed the missing data for the categorical variables by using the mode imputation method (Memon *et al.*, 2023). Next, we randomly split the data into 70% train data and 30% test data, where the censoring rates in both train and test data were kept similar to the original data (i.e. around 68%). After that, we did feature selection using random survival forests (RSF) (Ishwaran *et al.*, 2008) and a Python package **scikit-learn** (Pedregosa *et al.*, 2011) with 'eli5' and 'PermutationImportance' functions. To train the RSF, we mostly used the default hyper-parameters from the package. For example, 'n estimators'=1000, 'min samples split'=10, 'min samples leaf'=15, 'max features'= sqrt, 'n jobs'=-1, and 'random state'=random state. Then, using 15 iterations, we obtained the weights showing the importance of each variable. We selected

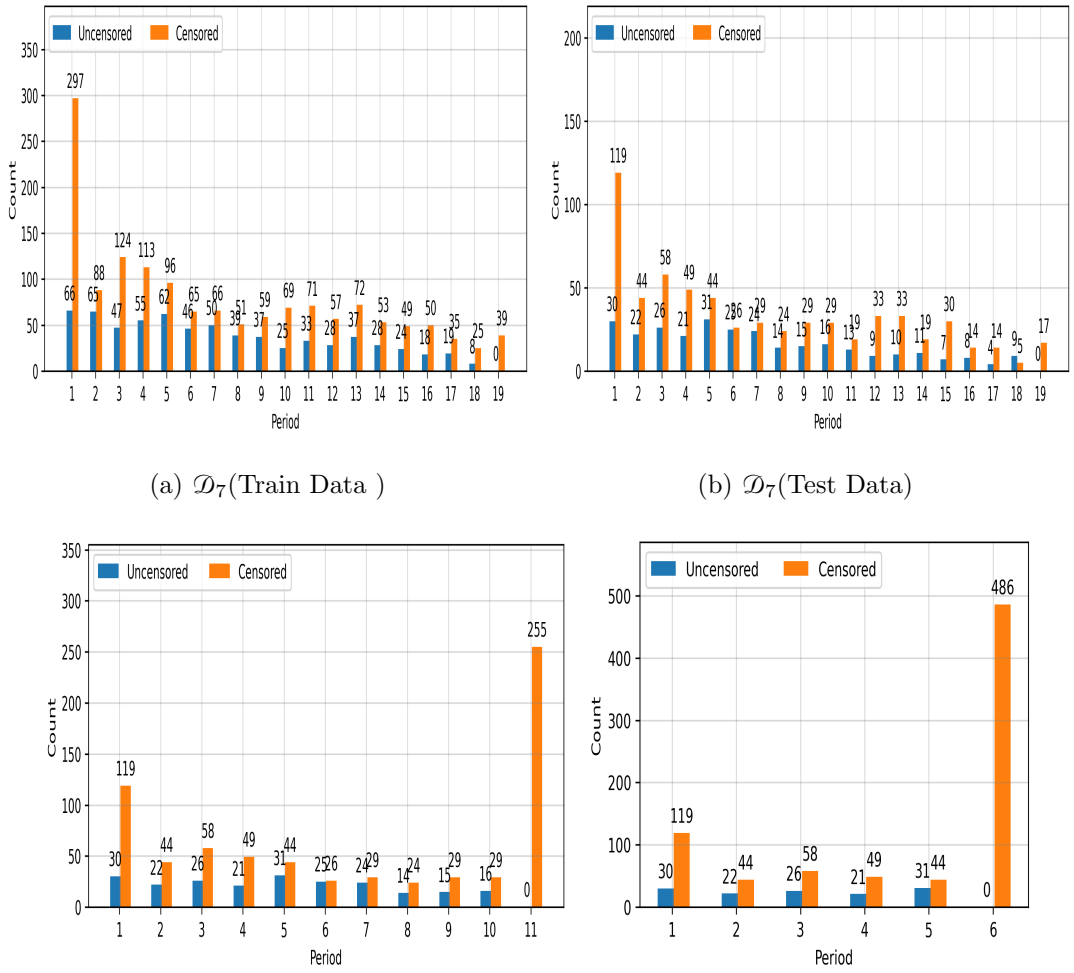
variables whose weight was greater than 0.003. After that, we evaluated the performance of the fitted model to the train data containing the selected variables on the test data using the C-index as the default measure of the package. We then repeated this procedure until the C-index did not decrease. Finally, we only employed 13 variables as the model's covariates denoted by 'CSMD3', 'EGFR', 'FLG', 'MUC16', 'MUC4', 'PIK3CA', 'PTEN', 'TP53', 'TTN', 'USH2A', 'Age', 'Gender', and 'TCGA tumor type'.

We generated 100 pairs of independent train and test data. Then, the train data were discretised using $\mathscr{D}_7$ = {0, 74, 152.003, 234.465, 321.928, 415.039, 514.573, 621.488, 736.966, 862.497, 1000, 1152.003, 1321.928, 1514.573, 1736.966, 2000, 2321.928, 2736.96, 3321.928, ∞}. Meanwhile, $\mathscr{D}_7$, $\mathscr{D}_8$ = {0, 74, 152.003, 234.465, 321.928, 415.039, 514.573, 621.488, 736.966, 862.497, 1000, ∞}, and $\mathscr{D}_9$ = {0, 74.001, 152.003, 234.465, 321.928, 415.038, ∞} were used for the test data discretisation. $\mathscr{D}_7$ followed the 'half-life' approach by Gensheimer & Narasimhan (2019) so that we have increasing width of intervals. $\mathscr{D}_8$ and $\mathscr{D}_9$ are the truncated version of $\mathscr{D}_7$ by ignoring all points over 1000 except the ∞. After the discretisation process, the censoring rates of the data have changed to around 68.28%, 77.01%, and 77.01% for $\mathscr{D}_7$, $\mathscr{D}_8$, and $\mathscr{D}_9$, respectively. From Figure 3.11, regardless of the last period T_{max}, the data distributions are right-skewed.

After we fitted good and overfitted Nnet-survival architectures in Table B.2 (Appendix B.1) to each train data, we computed the predicted survival probabilities from the train and the respective test data. This prediction procedure was repeated 100 times with respect to the 100 pairs of train and test data. In Figure 3.12, the outputs from the good models in the first train and the first test data are almost the same. On the other hand, Figure 3.13 shows that the predicted survival curves from the overfitted models in the train data are mostly closer to their outcomes. However, the outputs in the test data are not. These results indicate that overfitting has happened in the overfitted Nnet-survival.

The complete results of model evaluation using integrated Brier score on 100 test data for the good and overfitted models are given in panels (a) and (b) of Figure 3.14, respectively. We can see from the figure that the integrated Brier score can vary depending on the chosen discretisation setup in the test data. The smallest $\widehat{\text{IBS}}_n$ are in $\mathscr{D}_9$. In contrast, $\widehat{\text{IBS}}_n$ in $\mathscr{D}_7$ are the largest in both panels.

By taking the discretisation points as the subsets of $\mathcal{D}_7$, we only employed $\widehat{\mathrm{BS}}_n(t)$ with a smaller follow-up resulting in the change of $\widehat{\mathrm{IBS}}_n$. These results are in line with Scenario 1 of Simulation 1. This result also indicates the pitfall of $\widehat{\mathrm{IBS}}_n$, where we may 'cheat' to report good $\widehat{\mathrm{IBS}}_n$ only by manipulating the test data structure (and hence the $\widehat{\mathrm{BS}}_n(t)$).



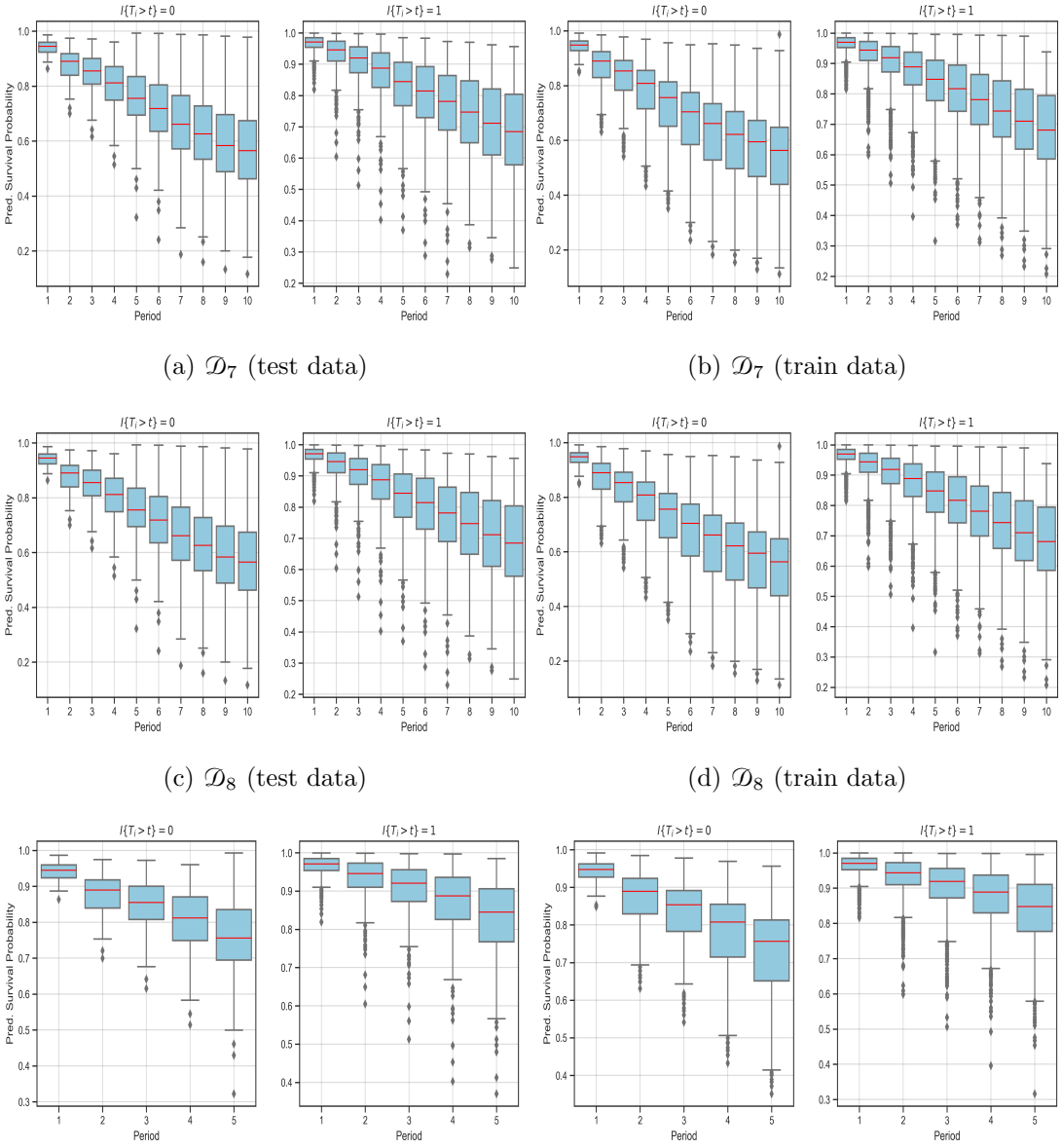
(a) $\mathcal{D}_7$(Train Data)

(b) $\mathcal{D}_7$(Test Data)

(c) $\mathcal{D}_8$(Test Data)

(d) $\mathcal{D}_9$(Test Data)

Figure 3.11: The distributions of event and censoring times of TCGA Data for each discretisation setup over $\mathcal{T}$, where all individuals in T_{max} were administratively censored.



(a) $\mathscr{D}_7$ (test data)

(b) $\mathscr{D}_7$ (train data)

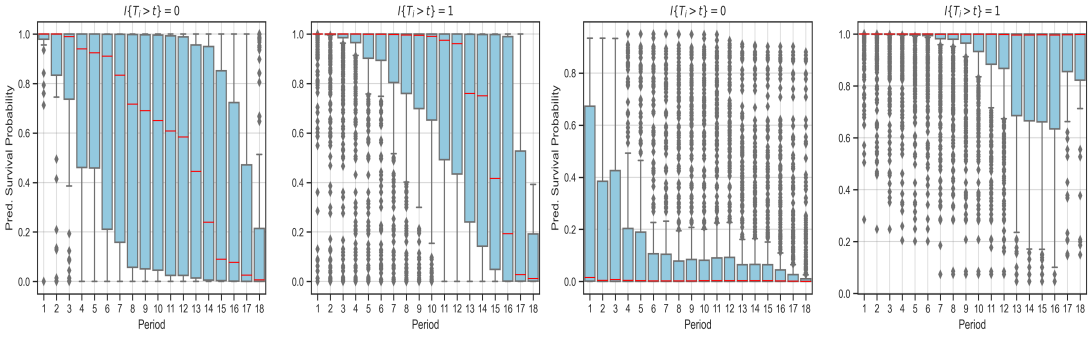
(c) $\mathscr{D}_8$ (test data)

(d) $\mathscr{D}_8$ (train data)

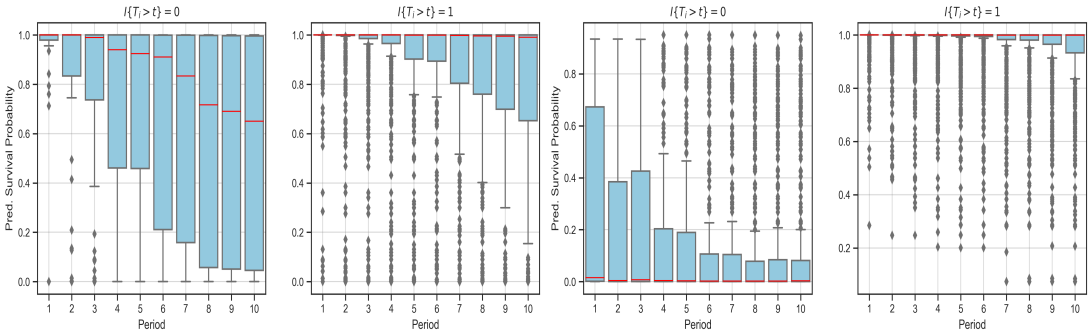
(e) $\mathscr{D}_9$ (test data)

(f) $\mathscr{D}_9$ (train data)

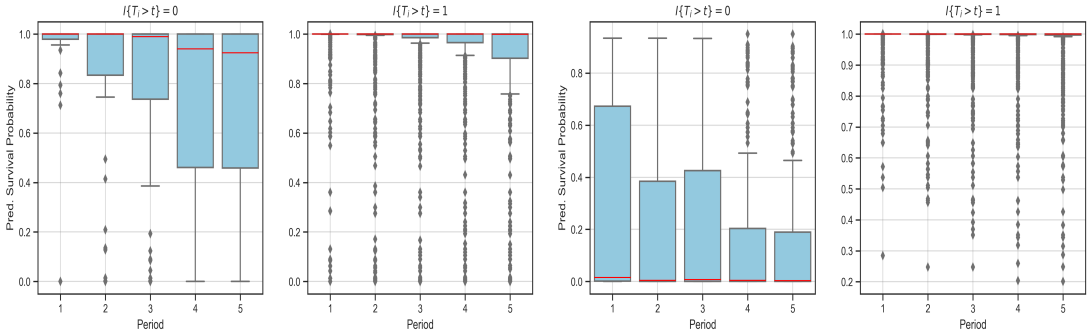
Figure 3.12: The predicted survival probabilities for all 1000 individuals from the good Nnet-survival in the first test data (panels (a),(c),(e)) and the first train data (panels (b),(d),(f)) of TCGA data. The outputs are categorised by the outcomes ($I_{\{T_i>t\}}$) over ($T_{max}-1$) for each splitting points setup.



(a) $\mathcal{D}_7$ (test data) (b) $\mathcal{D}_7$ (train data)

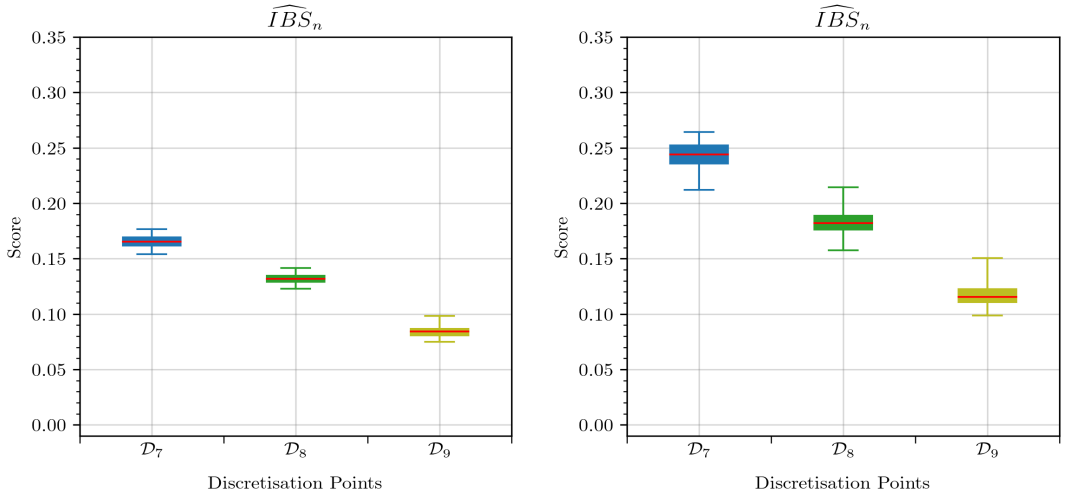


(c) $\mathcal{D}_8$ (test data) (d) $\mathcal{D}_8$ (train data)



(e) $\mathcal{D}_9$ (test data) (f) $\mathcal{D}_9$ (train data)

Figure 3.13: The predicted survival probabilities for all 1000 individuals from the overfitted Nnet-survival in the first test data (panel(a),(c),(e)) and the first train data (panels (b),(d),(f)) of TCGA data. The outputs are categorised by the outcomes ($I_{\{T_i>t\}}$) over ($T_{max}-1$) for each splitting points setup.



(a) Good Models

(b) Overfitted Models

Figure 3.14: Integrated Brier Scores of the good (a) and overfitted models over periods ($T_{max} - 1$) for the three splitting points estimated on 100 TCGA test data (n_{test}=30% of the Data) and 100 train data (n_{train}=1000). Censoring rates of the train and test data are close to 0%.

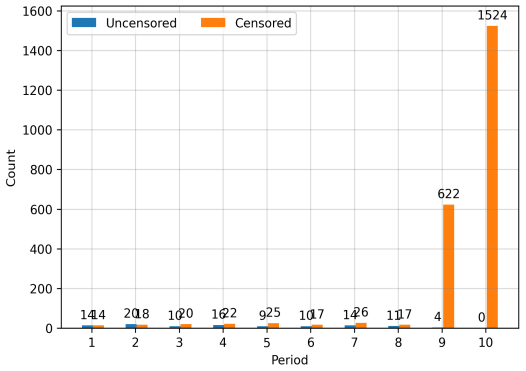
Breast Cancer Data

The breast cancer data used in this section was obtained from the results of the UK Women's Cohort Study (UKWCS) in 1990 (Cade *et al.*, 2017). The data contains information on 35,372 women, including individual characteristics, chronic disease, lifetime lifestyle, and dietary patterns. In this thesis, we are interested in the duration from the subject's first time joining the study until the incidence of breast cancer as the event of interest, where the used time units are in years. We included some of the variables at the baseline used in the work by Aivaliotis *et al.* (2021) as covariates: age, alcohol consumption, folate intake, walking time in summer and winter, and height. Since breast cancer incidence information was not always available for each individual, we only employed 34,493 women in our numerical implementation.

The used breast cancer data has unique characteristic since the proportion of event times is so tiny while the majority of individuals have survived up to

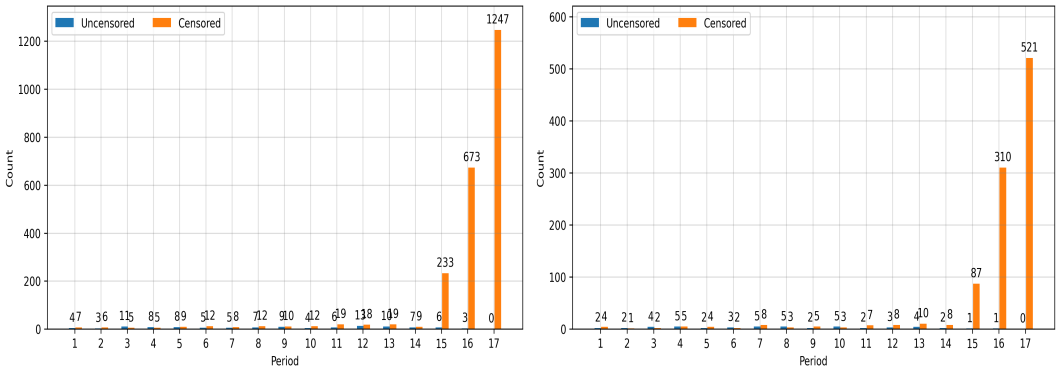
the end of the follow-up. In particular, the number of events and censoring in the data are 1,571 (4.5%) and 32,922 (95.5%), respectively. The distribution of the event times spreads almost evenly between 0 and 16, and there is no event beyond 16 years. However, we only used 30% of the data for computational convenience, where we randomly took 30% individuals from both censored and uncensored individuals and then combined them into a data sample. After that, we randomly grouped the sample data into 70% train data and 30% test data, where the censoring rates in both train and test data were kept similar to the sample data (i.e. around 0.045%). This random splitting was repeated 100 times to obtain 100 pairs of independent train and test data.

We fitted Nnet-survival architecture in Table B.3 to the 100 train data discretised by three sets of discretisation points, namely $\mathscr{D}_{10}$={0, 2.75, 5, 7, 9, 10.5, 11.75, 13, 14.5, 15.5, ∞}, $\mathscr{D}_{11}$={0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ∞}, and $\mathscr{D}_{12}$={0, 10, 16, ∞}. We chose $\mathscr{D}_{10}$ because it distributes the event times almost evenly amongst the periods. $\mathscr{D}_{11}$ represents a discretisation setup where each element is the number of years in the data. Meanwhile, $\mathscr{D}_{12}$ is applied to separate the observed times into two main periods only, namely $[0, 10)$ and $[10, 16)$, so that we have a simpler model to be fitted. The distributions of the last train and test data for each discretisation setup are given in Figure 3.15. We can see from the figure that the number of uncensored individuals is so small compared to the censored individuals, where most of the censoring happened due to the end of the study (administrative censoring). After fitting the models to the train data, we computed $\widehat{\text{IBS}}_n$ from their respective 100 independent test data. Note that the train and the respective test data were discretised by the same discretisation setups, namely $\mathscr{D}_{10}$, $\mathscr{D}_{11}$, and $\mathscr{D}_{12}$.



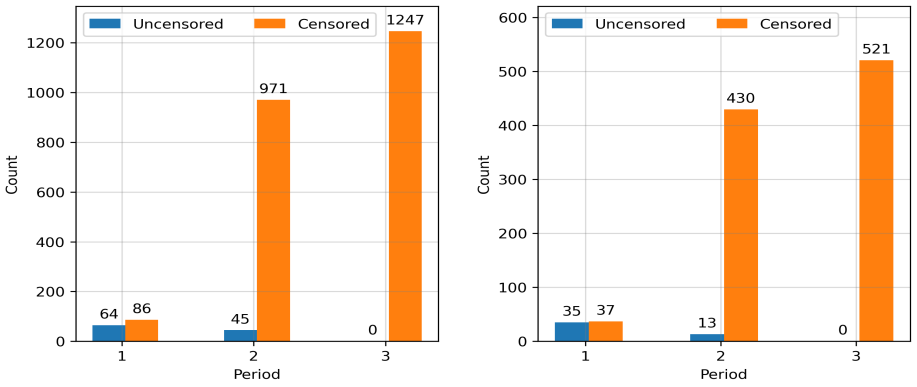
(a) Train Data with $\mathscr{D}_{10}$

(b) Test Data with $\mathscr{D}_{10}$



(c) Train Data with $\mathscr{D}_{11}$

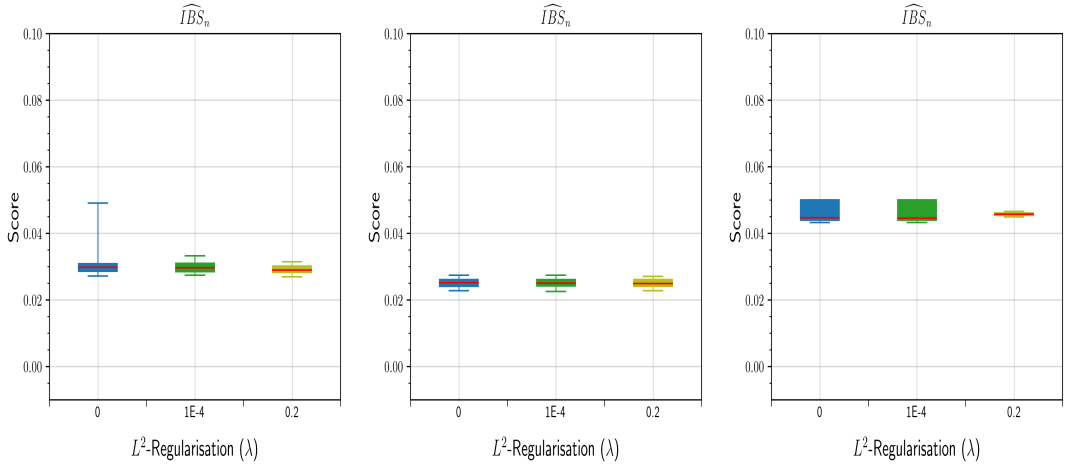
(d) Test Data with $\mathscr{D}_{11}$



(e) Train Data with $\mathscr{D}_{12}$

(f) Test Data with $\mathscr{D}_{12}$

Figure 3.15: The event and censoring times distributions of the last train and test data from the 100 repetitions discretised by $\mathscr{D}_{10}$, $\mathscr{D}_{11}$, and $\mathscr{D}_{12}$.



(a) $\mathscr{D}_{10}$ (b) $\mathscr{D}_{11}$ (c) $\mathscr{D}_{12}$

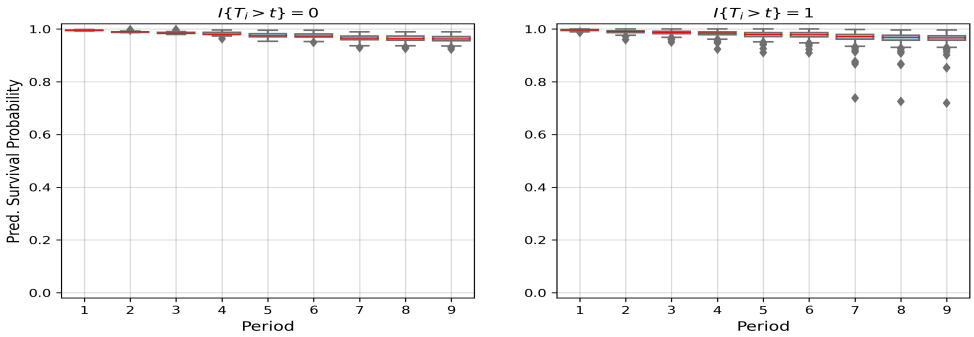
Figure 3.16: Integrated Brier score over different values of L^2- regularisation (λ) for three splitting points, i.e. $\mathscr{D}_{10}$, $\mathscr{D}_{11}$, and $\mathscr{D}_{12}$. They were estimated on 100 breast cancer test data from the models fitted to 100 breast cancer train data.

The main results of this simulation are given in Figure 3.16. We have nine different models based on the combinations of three different discretisation setups and three values of λ. As we can see from the figure, there is no significant change of $\widehat{\text{IBS}}_n$ when λ is varied. In particular, the three different values of L^2 regularisation do not significantly affect the models' predictive performance. Due to the small proportion of events in the data, even the more complex model (e.g. model with $\lambda = 0$) does not make $\widehat{\text{IBS}}_n$ worse. Figure 3.17 shows the predicted survival curves for all individuals over $\{1, \cdots, T_{max}-1\}$ from a test data discretised by $\mathscr{D}_{10}$. For different values of λ, the differences between the outputs and the outcomes are always close to zero. The same results are also shown by the predicted survival curves in the test data discretised by $\mathscr{D}_{11}$ and $\mathscr{D}_{12}$ (see Figure A.1 and Figure A.2, respectively, in Appendix A.1). The predicted survival curves are almost flat and close to one over the follow-up time, disregarding their respective outcomes. Because around 95.5% individuals are censored, most of the outcomes are $I_{\{T_i>t\}} = 1$, resulting in the tiny differences between $S(t; \mathbb{Z}_i)$ and $I_{\{T_i>t\}} = 1$ have dominant contributions to $\widehat{\text{BS}}_n(t)$ for $t \in \{1, \cdots, T_{max} - 1\}$. Thus, the values of $\widehat{\text{BS}}_n(t)$ (and hence $\widehat{\text{IBS}}_n$) are always close to zero regardless

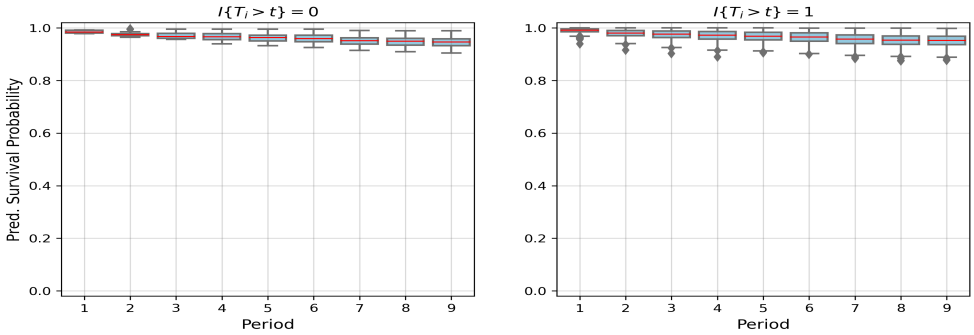
of the fitted models.

Through numerical experiments on the breast cancer data, we have explored other pitfalls of the Brier score (and hence integrated Brier score). These measures have a problem in assessing the prediction performance of the fitted models when the event times are scarce while, on the contrary, the proportion of censoring is significantly large. In this case, we find that: (1) it is difficult to differentiate two or more different models fitted to the same data because their $\widehat{\text{IBS}}_n$ will always be close to zero, and (2) one might be confused to evaluate the model's performance since even for the more complex model we still have relatively small $\widehat{\text{IBS}}_n$. Our results are consistent with the works by Aivaliotis *et al.* (2021), who reported that $\widehat{\text{IBS}}_n$ in the breast cancer data was also close to zero.

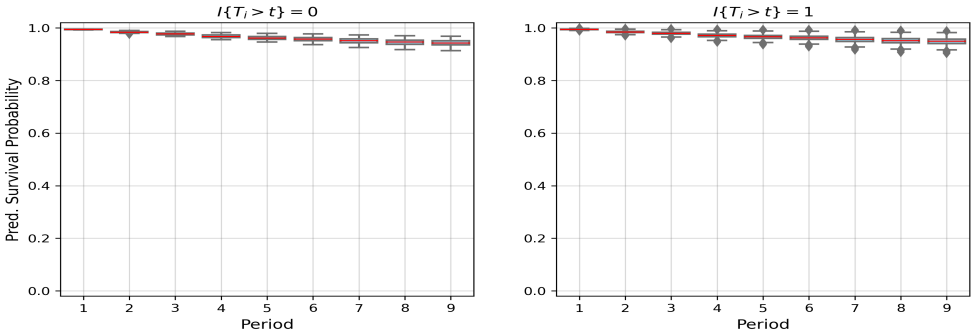
In summary, based on our experiments, we have shown that the integrated Brier score cannot evaluate how well the model predictive performance is when the censoring proportion is vast, most of which is administrative censoring. The models will always output high predicted survival curves. At the same time, most outcomes satisfy the primary indicator function of the integrated Brier score, $I_{\{T_i>t\}} = 1$, so that $\widehat{\text{IBS}}_n$ is always good. The used breast cancer data is a perfect example to explore such weaknesses, as we have shown in this section. Therefore, in the next section, we will propose the reference value for the integrated Brier score to cope with the pitfalls we have demonstrated through the case studies and real-world examples.



(a) $\mathcal{D}_{10}$ with $\lambda = 0$



(b) $\mathcal{D}_{10}$ with $\lambda = 1E-4$



(c) $\mathcal{D}_{10}$ with $\lambda = 0.2$

Figure 3.17: $S(t; \mathbb{Z}_i)$ for $i = 1, \cdots, 1000$ categorised by $I_{\{T_i > t\}}$ over $\{1, \cdots, T_{max} - 1\}$ in the breast cancer data. The prediction was conducted in a test data based on a model fitted to the respective train data with three values of L^2-regularisation (λ), i.e. $0, 1E-4$, and 0.2. The train and test data were discretised by $\mathcal{D}_{10}$.

3.2 Integrated Brier Score for KM Estimator

In the previous section, we have explored some pitfalls of the integrated Brier score. Using this measure, we sometimes cannot evaluate whether a model has good predictive performance or not, and the value can be manipulated easily by changing the data structure. The values of the Brier score (or the integrated Brier score) between 0 and 0.1 indicated a very accurate prediction performance (Graf *et al.*, 1999). However, as we have shown in some numerical experiments in Section 3.1, we may still confuse to evaluate the model performance although the values of the integrated Brier score are close to zero. To cope with such weaknesses, in this section, we propose the reference value for the integrated Brier score as its complement measure in evaluating the model performance.

KM estimator outputs the predicted survival curves that are common for all individuals $i(i = 1, \cdots, n)$. This estimator assumes that event time T_i is independent of $\mathbb{Z}_i$, so that the predicted survival curve at each period t is the same for all individuals regardless of their unique characteristics. The assumption underestimates the effect of covariates in modelling individual-specific quantities. Although KM estimator is a poor model to estimate individual-specific survival curve, it is sometimes useful when we are interested in survival modelling regardless of the individual characteristics. For instance, we assume that censoring time D_i is independent of $\mathbb{Z}_i$ since the effects $\mathbb{Z}_i$ are sometimes insignificant to the censoring time (Gerds *et al.*, 2013). Hence, the KM estimator is usually used to compute $\widehat{G}_n$ (Gerds *et al.*, 2013; Graf *et al.*, 1999; Song *et al.*, 2012; Uno *et al.*, 2011).

Due to its poor quality for individual-specific quantity prediction, we can say that KM estimator is a 'naive' prediction in survival modelling. KM survival curve had been utilised for long time as the reference for the predicted survival curves (Gensheimer & Narasimhan, 2019; Royston & Parmar, 2002). Furthermore, the use of Brier score (and integrated Brier score) for KM estimator as the benchmark values had already been discussed by Schumacher *et al.* (2003). The smaller the values of Brier score than the benchmark values, the smaller the model prediction error. However, our motivation of using them as the reference values is highly different. Some crucial drawbacks of integrated Brier score that

have been investigated in comprehensive numerical experiments (see Section 3.1) have motivated us to employ the integrated Brier score for KM estimator as the reference value. In particular, we define the integrated Brier score for KM estimator as follows

$$\widehat{\text{IBSKM}}_n = \frac{1}{T_{max}-1} \sum_{t=1}^{T_{max}-1} \widehat{\text{BSKM}}_n(t),$$

where

$$\widehat{\text{BSKM}}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \left[I_{\{X_i \leq t\}} \frac{I_{\{T_i < D_i\}}}{\widehat{G}_n(T_i)} (0 - S(t))^2 + I_{\{X_i > t\}} \frac{I_{\{t < D_i\}}}{\widehat{G}_n(t)} (1 - S(t))^2 \right],$$

and $S(t)$ the predicted survival curve at $t \in \mathcal{T}$ using KM estimator (2.3). We expect that $\widehat{\text{IBS}}_n >> \widehat{\text{IBSKM}}_n$ from a really bad model, $\widehat{\text{IBS}}_n << \widehat{\text{IBSKM}}_n$ from an excellent model, and $\widehat{\text{IBS}}_n \approx \widehat{\text{IBSKM}}_n$ from a poor model.

We now apply $\widehat{\text{IBSKM}}_n$ to all simulation scenarios and the real-world examples discussed in Section 3.1. The settings, Nnet-survival architectures, and the data are still the same. We will report the differences between the original integrated Brier score and its respective integrated Brier score for KM estimator, i.e. $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$, estimated from 100 independent test data. The same as previous implementations, we display the results by boxplots.

The results for Scenario 1 of Simulation 1 are shown by Figure 3.18. The figure contains two panels, namely the results from the good models (panel (a)) and the overfitted models (panel (b)). As we can see from panel (a), $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ are less than zero for all discretisation setups, meaning that the models' predictive performance is good. On the other hand, panel (b) shows that $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ are much greater than zero, meaning that the model's predictive performance is really bad. The model performance is also consistent for each model.

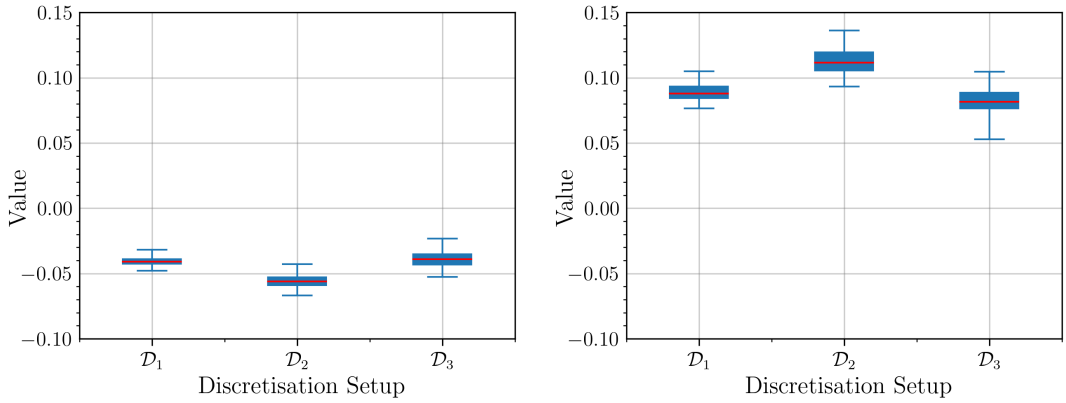
Figure 3.19 shows the results for Scenario 2 of Simulation 1. $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ for $\mathcal{D}_4$ and $\mathcal{D}_6$ are relatively far from zero, where we have negative values for the good models and positive values for the overfitted models. However, for $\mathcal{D}_5$, $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ are very close to zero, although their medians (i.e. the red horizontal lines within the boxplots) are slightly below or above zero for the good and overfitted models, respectively. Consequently, we can say that the model performance obtained from $\mathcal{D}_5$ is poor, the same as the KM estimator.

Figure 3.20 displays the results for TCGA data. We can see from the figure that $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ are negative for the good models (panel (a)) and positive for the overfitted models (panel (b)). These results are consistent with the fitted Nnet-survival architectures. We can see also from the table that $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ gets closer to zero as the number of used periods decreases, where $\mathscr{D}_9 \subset \mathscr{D}_8 \subset \mathscr{D}_7$. This result means that the smaller the number of periods, the worse the model performance. This trend does not happen in Scenario 1 of Simulation 1, which also uses a fixed model, due to the difference in data structure.

The result of the breast cancer data are given by Figure 3.20 containing of three panels with respect to the used discretisation setups. As we can see from the figure, the values of $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ are very close to zero for combinations of discretisation setup and λ, indicating that the model performances are poor the same as the quality of KM estimator. This justification contradicts the conclusions when we only integrated the Brier score without the integrated Brier score for KM estimator. We may conclude that the model performances are excellent if we only use $\widehat{\text{IBS}}_n$ solely.

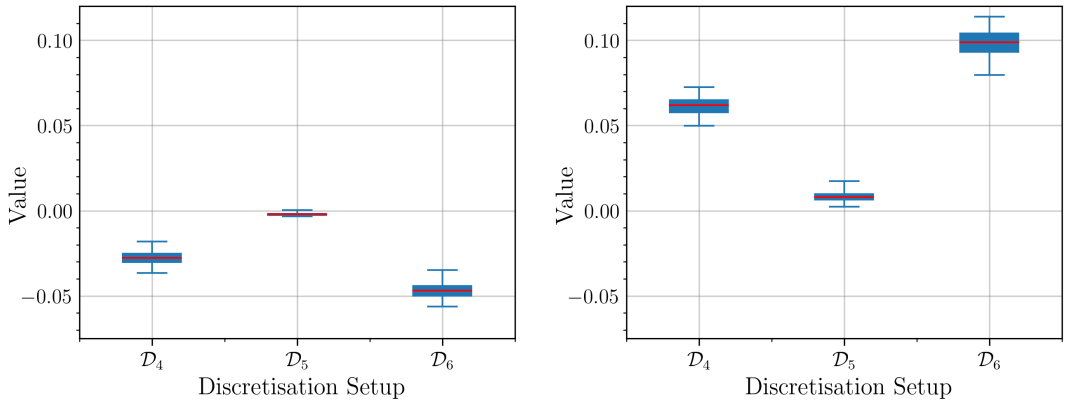
In this section, we have demonstrated the advantage of using the integrated Brier score for KM estimator as the reference value for the integrated Brier score. One of the significant limitations of the integrated Brier score is its restricted ability to evaluate the performance of a single model. Due to the data structure effects, a close-to-zero integrated Brier score may not necessarily indicate a good model performance. However, by incorporating the integrated Brier score for KM estimator, we can effectively overcome this limitation and provide a more comprehensive evaluation of the model performance.

While we have demonstrated the main benefit of using the integrated Brier score and the integrated Brier score for KM estimator altogether, the integrated Brier score can still be challenging to interpret. Therefore, in the subsequent section, we will introduce some alternative measures, namely modified integrated Brier scores. These measures offer improved interpretations compared to the integrated Brier score, providing a more optimistic outlook for model performance evaluation.



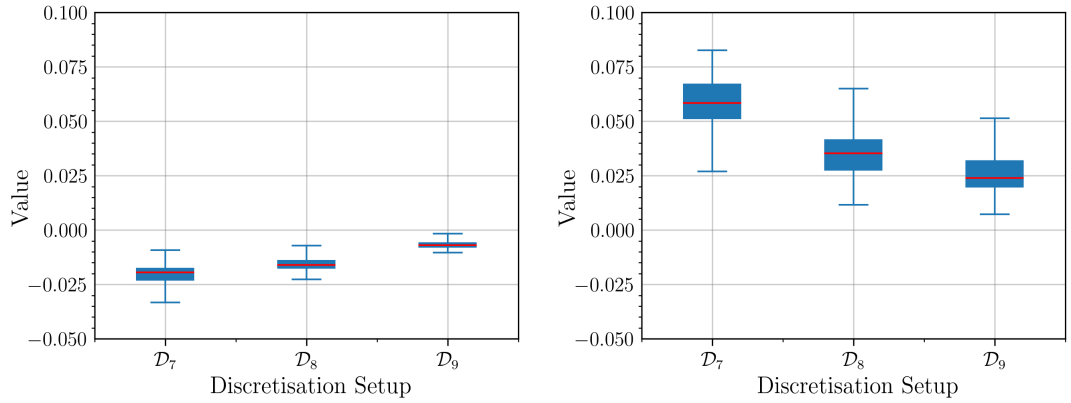
(a) Good Models (b) Overfitted Models

Figure 3.18: Boxplots of 100 $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ for each discretisation setup in (a) the good models and (b) the overfitted models from Scenario 1 of Simulation 1.



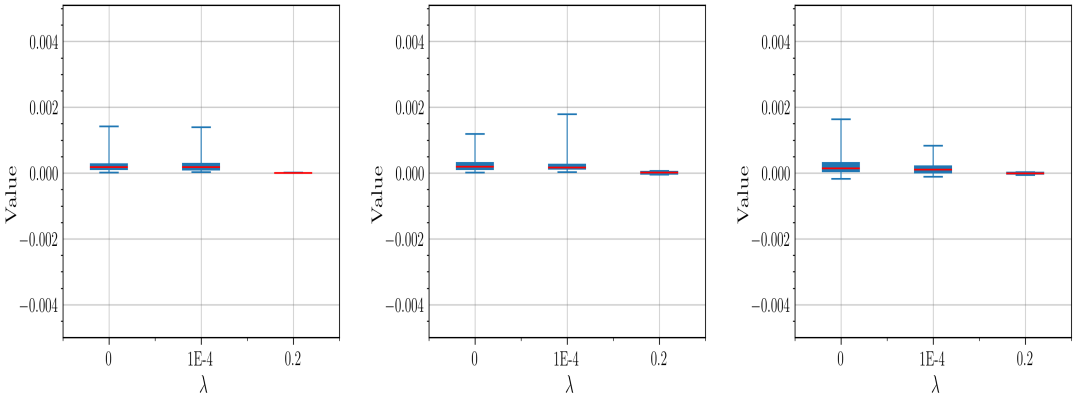
(a) Good Models (b) Overfitted Models

Figure 3.19: Boxplots of 100 $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ for each discretisation setup in (a) the good models and (b) the overfitted models from Scenario 2 of Simulation 1.



(a) Good Models (b) Overfitted Models

Figure 3.20: Boxplots of 100 $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ for each discretisation setup in (a) the good models and (b) the overfitted models from TCGA data.



(a) $\mathcal{D}_{10}$ (b) $\mathcal{D}_{11}$ (c) $\mathcal{D}_{12}$

Figure 3.21: Boxplots of 100 $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ for each λ with discretisation setups (a) $\mathcal{D}_{10}$, (b) $\mathcal{D}_{11}$, and (c) $\mathcal{D}_{12}$ from the breast cancer data.

3.3 Modified Integrated Brier Scores

In the previous section, we have applied the integrated Brier score for KM estimator to complement the "classical" integrated Brier score. We can gain a more comprehensive understanding of our survival prediction performance by considering both measures. However, it is essential to note that the interpretation of the integrated Brier score is limited due to their nature as mean squared error between the outputs and the outcomes. We have improved the interpretability by comparing it to the integrated Brier score for KM estimator. In this section, we will improve the interpretability of the integrated Brier score using another approach, namely by involving the variability of the outputs to the measure, leading us to propose several new measures that are based on the main principles of the integrated Brier score, including normalised integrated Brier score, centered integrated Brier score, and normalised centered integrated Brier score. These new measures have practical implications for survival prediction modelling, enhancing the interpretability and applicability of the integrated Brier score in real-world scenarios.

3.3.1 Normalised Integrated Brier Score

For notational convenience, we denote

$$\psi_i^t = I_{\{T_i > t\}},$$

and

$$\pi_i^t = S(t; \mathbb{Z}_i),$$

for $t \in \mathcal{T}$. These notations do not apply only in this section but also for the rest of this chapter. Throughout this thesis, we assume that ψ_i^t are independent to each other for all individuals $i (i = 1, \cdots, n)$. It is clear that

$$\psi_i^t \sim \text{Bern}(\pi_i^t), \tag{3.2}$$

where

$$\hat{\mu}_{i;t} = \pi_i^t \tag{3.3}$$

and

$$\hat{\sigma}^2_{i;t} = \pi^t_i(1-\pi^t_i) \tag{3.4}$$

are the estimated values of $\mathbb{E}[\psi^t_i]$ and $\mathrm{Var}[\psi^t_i]$, respectively.

By taking the expectation of the ratio of the squared difference between ψ^t_i and $\hat{\mu}_{i;t}$ to $\hat{\sigma}^2_{i;t}$ as follows

$$\begin{aligned}
&\mathbb{E}\left[\frac{\left(\psi^t_i - \hat{\mu}_{i;t}\right)^2}{\hat{\sigma}^2_{i;t}}\right] \\
&= \mathbb{E}\left[\frac{\left(\psi^t_i - \pi^t_i\right)^2}{\pi^t_i\left(1-\pi^t_i\right)}\right] \\
&= \frac{\mathbb{E}\left[(\psi^t_i)^2 - 2\psi^t_i\pi^t_i + (\pi^t_i)^2\right]}{\mathbb{E}\left[\pi^t_i\left(1-\pi^t_i\right)\right]} \\
&= \frac{\mathbb{E}\left[(\psi^t_i)^2\right] - 2\mathbb{E}\left[\psi^t_i\right]\pi^t_i + \mathbb{E}\left[(\pi^t_i)^2\right]}{\pi^t_i\left(1-\pi^t_i\right)} \\
&= \frac{\pi^t_i - 2\pi^t_i\pi^t_i + (\pi^t_i)^2}{\pi^t_i - (\pi^t_i)^2} \\
&= 1,
\end{aligned} \tag{3.5}$$

we normalise the squared error term $\left(\psi^t_i - \hat{\mu}_{i;t}\right)^2$ by $\hat{\sigma}^2_{i;t}$.

Dividing (2.10) by (3.4), we propose

$$\begin{aligned}
&\widehat{\mathrm{NBS}}_n(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[I_{\{X_i\le t\}}I_{\{T_i<D_i\}}\frac{\left(0-\pi^t_i\right)^2}{\hat{\sigma}^2_{i;t}}\widehat{G}^{-1}_n(T_i) + I_{\{X_i>t\}}\frac{\left(1-\pi^t_i\right)^2}{\hat{\sigma}^2_{i;t}}\widehat{G}^{-1}_n(t)\right]
\end{aligned} \tag{3.6}$$

as the estimator of (3.5) called normalised Brier score at a fixed $t \in \mathcal{T}$. The normalised Brier score should be more informative than the Brier score because it provides information about how well our prediction relative to the variance is. Furthermore, we propose

$$\widehat{\mathrm{NIBS}}_n = \frac{1}{T_{max}-1}\sum_{t=1}^{T_{max}-1}\widehat{\mathrm{NBS}}_n(t) \tag{3.7}$$

as normalised integrated Brier score over $\{1,\cdots,T_{max}-1\}$. A value close to one is what we expect from a good model performance in the estimator; otherwise, a far value from one indicates a bad model performance.

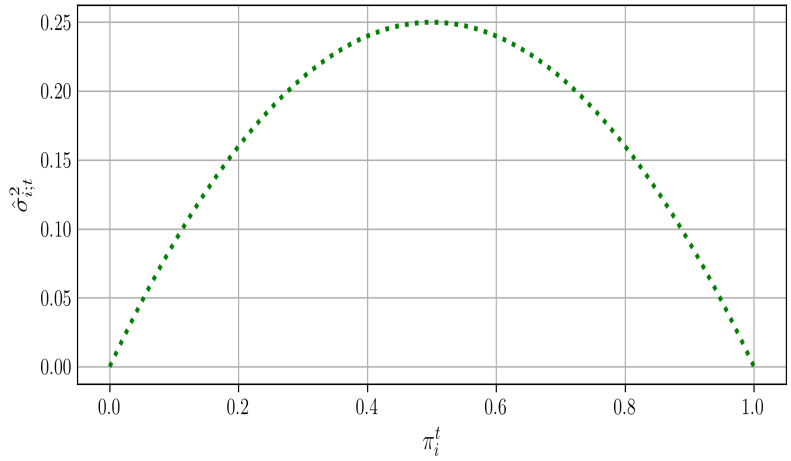


Figure 3.22: $\hat{\sigma}^2_{i;t}$ as a function of π_i^t for a fixed $t \in \mathcal{T}$.

One crucial drawback of the normalised Brier score (and hence the normalised integrated Brier score) is that it has no upper bound since its values are in $[0, \infty)$, depending on $\hat{\sigma}^2_{i;t}$ whose values are within $[0, 0.25]$ (see Figure 3.22). The closer $\hat{\sigma}^2_{i;t}$ to zero, the higher the contribution of i to $\widehat{\text{NBS}}_n(t)$. The contribution of i will be undefined, when $\hat{\sigma}^2_{i;t}$ equals zero. This situation is depicted by Figure 3.23 which shows the possible contributions of an individual i to $\widehat{\text{NBS}}_n(t)$ over $\hat{\sigma}^2_{i;t}$ for different values of squared error, $(\psi_i^t - \hat{\mu}_{i;t})^2$, when $\widehat{G}_n$ is completely specified, that is $\widehat{G}_n = G$. Note that $\widehat{G}_n = G$ happens when no censored individual exists in the data. However, when $\hat{\sigma}^2_{i;t}$ is very close to zero, $\hat{\mu}_{i;t} = \pi_i^t$ is either close zero or one (see Figure 3.22). If a model only outputs $\pi_i^t = 0$ or $\pi_i^t = 1$ for all individuals $i (i = 1, \cdots, n)$ and $t \in \mathcal{T}$, it might be a sign that we have an overfitted models. In this situation, $\widehat{\text{NIBS}}_n$ might be blown up because its denominator is nearly zero. We will discuss such cases and propose a possible solution for the problem in Section 3.3.5.

In this section, we developed a new calibration measure, the normalised integrated Brier score. The subsequent section will propose another type of calibration measure developed by centering the squared error in the Brier score.

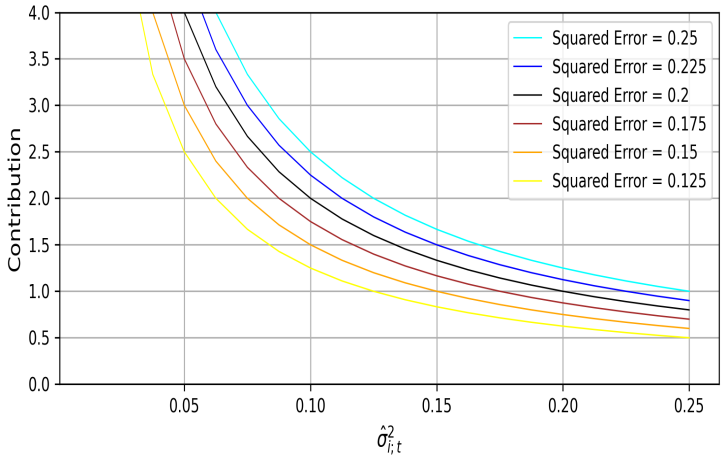


Figure 3.23: Contribution of an individual i to $\widehat{\mathrm{NBS}}_n(t)$ over $\hat{\sigma}^2_{i;t}$ for for a fixed $t \in \mathcal{T}$ and different values of squared error.

3.3.2 Centered Integrated Brier Score

In this section, we will develop another modified integrated Brier score. Instead of dividing the squared error term of the Brier score by the variance as in normalised Brier score (3.5), we subtract the squared error by the variance, $\hat{\sigma}^2_{i;t}$. In particular, we first define the following expectation:

$$\begin{aligned}
&\mathbb{E}\big[\big(\psi_i^t - \hat{\mu}_{i;t}\big)^2 - \hat{\sigma}^2_{i;t}\big)\big] \\
&= \mathbb{E}\big[\big(\psi_i^t - \pi_i^t\big)^2 - \pi_i^t\big(1-\pi_i^t\big)\big] \\
&= \mathbb{E}\big[(\psi_i^t)^2 - 2\psi_i^t\pi_i^t + (\pi_i^t)^2 - \pi_i^t\big(1-\pi_i^t\big)\big] \\
&= \mathbb{E}\big[(\psi_i^t)^2\big] - 2\mathbb{E}\big[\psi_i^t\big]\pi_i^t + \mathbb{E}\big[(\pi_i^t)^2\big] - \mathbb{E}\big[\pi_i^t\big(1-\pi_i^t\big)\big] \\
&= \pi_i^t - 2\pi_i^t\pi_i^t + (\pi_i^t)^2 - \pi_i^t + (\pi_i^t)^2 \\
&= 0.
\end{aligned} \tag{3.8}$$

Subtracting the squared errors $(0-\pi_i^t)^2$ and $(1-\pi_i^t)^2$ within the Brier score (2.10) by $\hat{\sigma}_{i;t}^2$, we propose

$$\begin{aligned}&\widehat{\text{CBS}}_n(t)\\&= \frac{1}{n}\sum_{i=1}^{n}\Big[I_{\{X_i\le t\}}I_{\{T_i<D_i\}}\big((0-\pi_i^t)^2-\hat{\sigma}_{i;t}^2\big)\widehat{G}_n^{-1}(T_i)\\&\quad + I_{\{X_i>t\}}\big((1-\pi_i^t)^2-\hat{\sigma}_{i;t}^2\big)\widehat{G}_n^{-1}(t)\Big],\end{aligned} \tag{3.9}$$

as the estimator of (3.8) called centered Brier score at a fixed period $t \in \mathcal{T}$. Finally, we define

$$\widehat{\text{CIBS}}_n = \frac{1}{T_{max}-1}\sum_{t=1}^{T_{max}-1}\widehat{\text{CBS}}_n(t) \tag{3.10}$$

as centered integrated Brier score.

When $\widehat{G}_n(t)$ is completely specified, that is $\widehat{G}_n(t) = G(t)$, the contribution of an individual i to $\widehat{\text{CBS}}_n(t)$ is

$$(\pi_i^t)^2 - \pi_i^t(1-\pi_i^t)$$

or

$$(1-\pi_i^t)^2 - \pi_i^t(1-\pi_i^t),$$

depending on whether $I_{\{X_i\le t\}}$ equals one or not. The possible contributions of an individual i in such cases lie within $[-0.125, 1]$ as shown by Figure 3.24. In a good model, we expect that π_i are close to zero for $I_{\{X_i\le t\}}$ and one for $I_{\{X_i>t\}}$. The contribution of an individual whose $I_{\{X_i\le t\}}$ and $\pi_i^t \in [0.5, 1]$ lies only within $[-0.125, 0]$. It is also the same as the contribution of an individual whose $I_{\{X_i>t\}}$ and $\pi_i^t \in [0, 0.5]$. In practice, however, it is unlikely to happen that π_i^t much greater than 0.5 when $X_i \le t$ and much less than 0.5 when $X_i > t$ since we usually optimise the model to avoid this situation. In other words, in non-extreme situations, individuals' contributions will be mostly around zero. Moreover, when $I_{\{X_i\le t\}}$, the contributions will be zero if i is censored, resulting in even smaller $\widehat{\text{CBS}}_n(t)$. This situation is not as expected for a measure because the small number, which is very close to zero, may confuse us in evaluating the model performance.

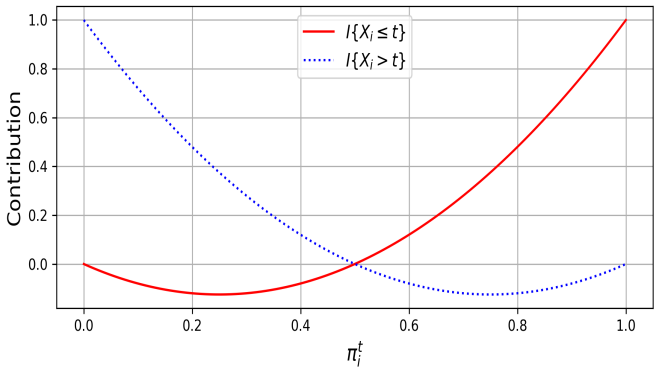


Figure 3.24: Contributions of individual i to $\widehat{\text{CBS}}_n(t)$ for a completely specified $\widehat{G}_n$ as functions of π_i^t.

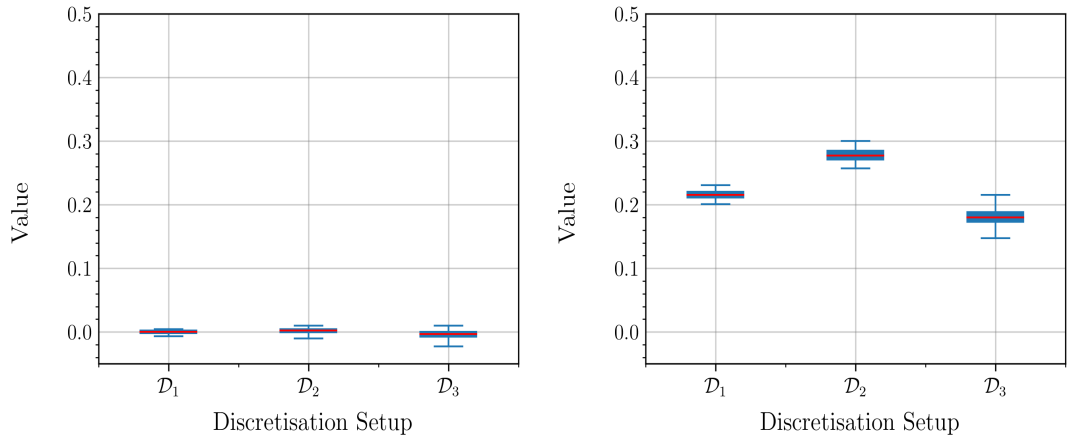
Numerical Implementations

This section aims to apply $\widehat{\text{CIBS}}_n$ to Simulation 1, TCGA data, and breast cancer data. We still utilise the model fitting and prediction that were conducted in the previous sections. In other words, the fitted Nnet-survival architectures and the obtained train and test data are still the same. We only evaluate the model's predictive performance using the centered integrated Brier score.

The result for Scenario 1 and 2 of Simulation 1, TCGA data, and breast cancer data are displayed by the boxplots in Figure 3.25, Figure 3.26, Figure 3.27, and Figure 3.28, respectively. As we can see from panel (a) in all four figures, the values of $\widehat{\text{CIBS}}_n$ for the good models are close to zero. They are obtained by subtracting $\hat{\sigma}^2_{i;t}$ from the squared error term within $\widehat{\text{IBS}}_n$. As a result, $\widehat{\text{CIBS}}_n$ will be smaller than the respective $\widehat{\text{IBS}}_n$. As expected, the same behaviour also happens in the overfitted models, as shown by panel (b) in each figure.

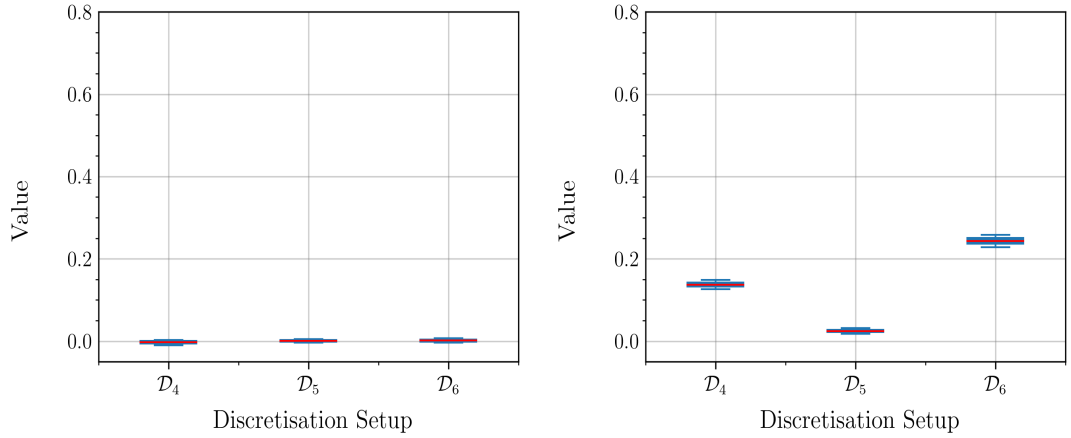
$\widehat{\text{CIBS}}_n$ can be interpreted as the values of the transformed $\widehat{\text{IBS}}_n$ whose mean value equals zero mean. Therefore, a perfect model is achieved when $\widehat{\text{CIBS}}_n$ equals the mean, namely zero. A perfect model means that the estimator (i.e. $\widehat{\text{CIBS}}_n$) has exactly the same value as the estimand (3.8). A value far from zero indicated a worse model performance. However, as we can see from the good models, it is not easy to differentiate the model performance based on $\widehat{\text{CIBS}}_n$ because the values are very close to zero. These results show that the centered integrated Brier score has improved the interpretability of the integrated Brier

score. However, we will have difficulty in comparing the performance of two or more different models.



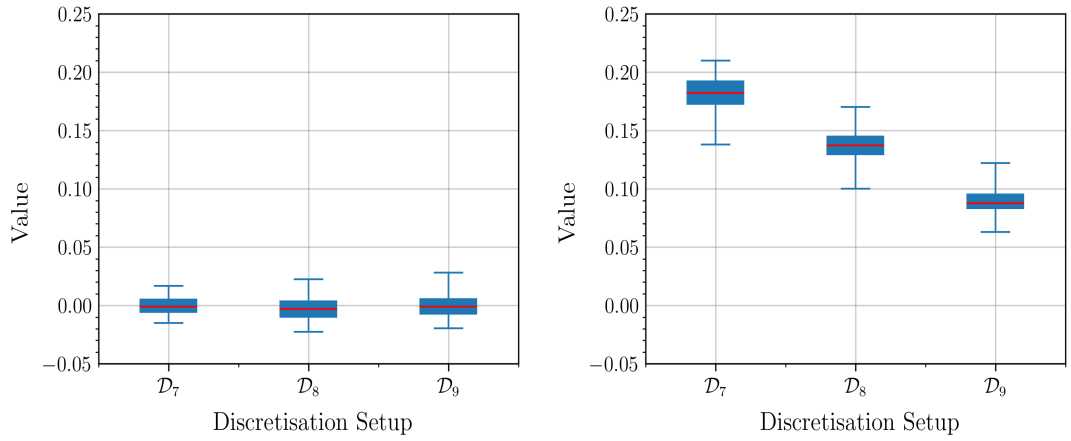
(a) Good Models (b) Overfitted Models

Figure 3.25: Boxplots of 100 $\widehat{\text{CIBS}}_n$ were obtained from the 100 test data over the three discretisation setups in (a) the good models and (b) the overfitted models from Scenario 1 of Simulation 1.



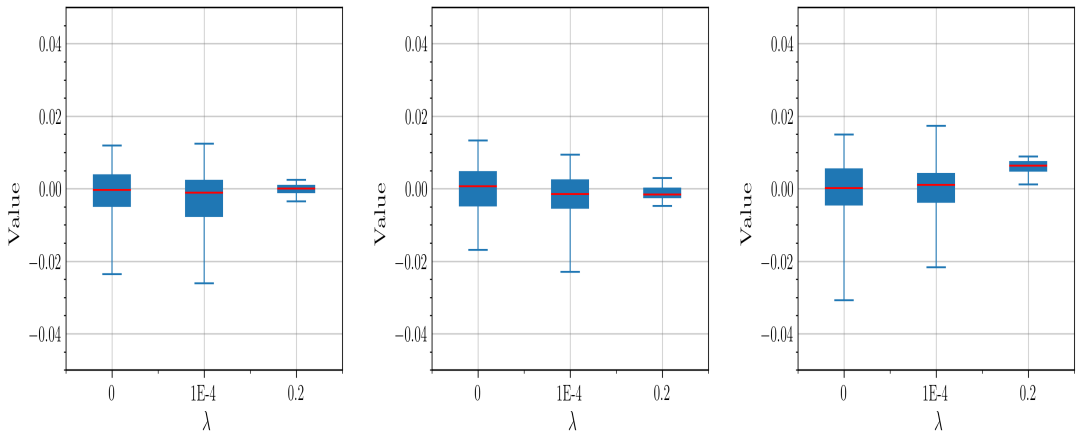
(a) Good Models (b) Overfitted Models

Figure 3.26: Boxplots of 100 $\widehat{\text{CIBS}}_n$ were obtained from the 100 test data over the three discretisation setups in (a) the good models and (b) the overfitted models from Scenario 2 of Simulation 1.



(a) Good Models (b) Overfitted Models

Figure 3.27: Boxplots of 100 $\widehat{\text{CIBS}}_n$ were obtained from the 100 test data over the three discretisation setups in (a) the good models and (b) the overfitted models from the TCGA data.



(a) $\mathcal{D}_{10}$ (b) $\mathcal{D}_{11}$ (c) $\mathcal{D}_{12}$

Figure 3.28: Boxplots of 100 $\widehat{\text{CIBS}}_n$ were obtained from the 100 test data over the three values of λ in (a) $\mathcal{D}_{10}$, (b) $\mathcal{D}_{11}$, and (c) $\mathcal{D}_{12}$ from the breast cancer data.

3.3.3 Normalised Centered Integrated Brier Score

To cope with the disadvantage of $\widehat{\text{CIBS}}_n$, we will propose a measure based on the first standardised moment to remove the high dependency of $\widehat{\text{CBS}}_n(t)$ on π_i^t. To derive the measure, we first consider the term inside the expectation of (3.8) as follows

$$\left(\psi_i^t - \pi_i^t\right)^2 - \pi_i^t\left(1 - \pi_i^t\right). \tag{3.11}$$

The variance of (3.11) can be computed by

$$\begin{aligned}
&\text{Var}\left[\left(\psi_i^t - \pi_i^t\right)^2 - \pi_i^t\left(1 - \pi_i^t\right)\right] \\
&= \mathbb{E}\left[\left(\left(\psi_i^t - \pi_i^t\right)^2 + \left(\left(\pi_i^t\right)^2 - \pi_i^t\right)\right)^2\right] \\
&\quad - \left(\mathbb{E}\left[\left(\psi_i^t - \pi_i^t\right)^2 - \pi_i^t\left(1 - \pi_i^t\right)\right]\right)^2 \\
&= \mathbb{E}\left[\left(\psi_i^t - \pi_i^t\right)^4 + 2\left(\psi_i^t - \pi_i^t\right)^2\left(\left(\pi_i^t\right)^2 - \pi_i^t\right) + \left(\left(\pi_i^t\right)^2 - \pi_i^t\right)^2\right] \\
&= \mathbb{E}\Big[\left(\psi_i^t\right)^4 - 4\left(\psi_i^t\right)^3\pi_i^t + 8\left(\psi_i^t\right)^2\left(\pi_i^t\right)^2 - 8\psi_i^t\left(\pi_i^t\right)^3 + 4\left(\pi_i^t\right)^4 \\
&\quad - 2\left(\psi_i^t\right)^2\pi_i^t + 4\psi_i^t\left(\pi_i^t\right)^2 - 4\left(\pi_i^t\right)^3 + \left(\pi_i^t\right)^2\Big] \\
&= \left(\pi_i^t\right)^2\left(\pi_i^t - 1\right)^2,
\end{aligned} \tag{3.12}$$

where the third equality is due to

$$\mathbb{E}\left[\psi_i^t\right] = \mathbb{E}\left[\left(\psi_i^t\right)^2\right] = \mathbb{E}\left[\left(\psi_i^t\right)^3\right] = \mathbb{E}\left[\left(\psi_i^t\right)^4\right] = \pi_i^t$$

and

$$\mathbb{E}\left[\left(\psi_i^t - \pi_i^t\right)^2 - \pi_i^t\left(1 - \pi_i^t\right)\right] = 0,$$

which is (3.8). The standard deviation of (3.11) can be obtained as follows

$$\sqrt{\left(\pi_i^t\right)^2\left(\pi_i^t - 1\right)^2} = \pi_i^t\left(1 - \pi_i^t\right). \tag{3.13}$$

Then, we divide (3.8) by standard deviation (3.13) so that we have

$$\mathbb{E}\left[\frac{\left(\psi_i^t - \pi_i^t\right)^2 - \pi_i^t\left(1 - \pi_i^t\right)}{\pi_i^t\left(1 - \pi_i^t\right)}\right] = 0. \tag{3.14}$$

To estimate (3.14), we divide the estimator of centered Brier score (3.9) by (3.13) so that we propose

$$\begin{aligned}&\widehat{\text{NCBS}}_n(t)\\&= \frac{1}{n}\sum_{i=1}^{n}\left[I_{\{X_i\leq t\}}I_{\{T_i<D_i\}}\left(\frac{(0-\pi_i^t)^2-\pi_i^t(1-\pi_i^t)}{\pi_i^t\,(1-\pi_i^t)}\right)\widehat{G}_n^{-1}(T_i)\right.\\&\left.+\,I_{\{X_i>t\}}\left(\frac{(1-\pi_i^t)^2-\pi_i^t(1-\pi_i^t)}{\pi_i^t\,(1-\pi_i^t)}\right)\widehat{G}_n^{-1}(t)\right]\end{aligned} \tag{3.15}$$

as the estimator of (3.14) called normalised centered Brier score at a fixed period $t \in \mathcal{T}$. It is clear that the closer $\widehat{\text{NCBS}}_n(t)$ to zero, the better the model's performance. By normalising the centered Brier score with its standard deviation, we expect that its normalised version (3.15) would be more informative because it shows how well the model performance relative to the mean and the standard deviation. We moreover define

$$\widehat{\text{NCIBS}}_n = \frac{1}{T_{max}-1}\sum_{t=1}^{T_{max}-1}\widehat{\text{NCBS}}_n(t) \tag{3.16}$$

as normalised centered integrated Brier score over $\{1,\cdots,T_{max}-1\}$.

Recall the formula of (3.14). Then, we can rewrite it as follows

$$\mathbb{E}\left[\frac{(\psi_i^t-\pi_i^t)^2-\pi_i^t\,(1-\pi_i^t)}{\pi_i^t\,(1-\pi_i^t)}\right] = \mathbb{E}\left[\frac{\left(\psi_i^t-\pi_i^t\right)^2}{\pi_i^t\,(1-\pi_i^t)}\right]-1,$$

resulting in the relationship between the estimators of each expected value as follows

$$\widehat{\text{NCIBS}}_n = \widehat{\text{NIBS}}_n - 1. \tag{3.17}$$

We can only compute one of the two measures from the relationship (3.17). Therefore, in all numerical experiments throughout this thesis, we only report $\widehat{\text{NCIBS}}_n$.

The plot of the possible contributions from an individual i at a fixed period t to $\widehat{\text{NCBS}}_n(t)$ for a perfectly specified $\widehat{G}_n(t)$ is given by Figure 3.29. Contrary to $\widehat{\text{CBS}}_n(t)$, the contribution of the first term with $\pi_i^t \in [0.5, 1]$ and the second term with $\pi_i^t \in [0, 0.5]$ lies in the range $[0, \infty)$. This contribution range is much wider than individual contribution in $\widehat{\text{CBS}}_n(t)$ with the same range of π_i^t, which is only limited in a narrow range that is close to zero. Although the normalised

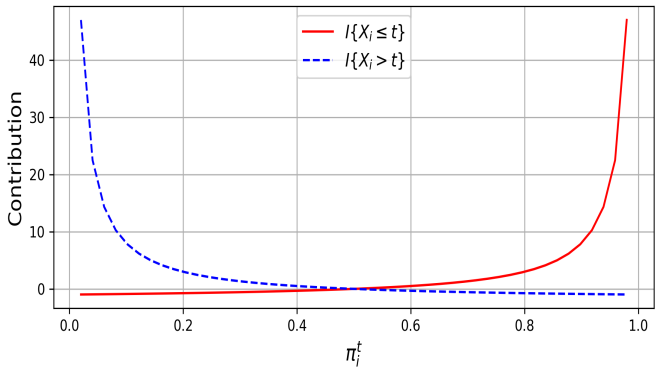


Figure 3.29: The contributions of an individual i to $\widehat{\mathrm{NCBS}}_n(t)$ for a perfectly specified $\widehat{G}_n(t)$ as functions of π_i^t.

centered Brier score faces the same problem as the normalised Brier score, whose values do not have an upper bound, it is also as good as the normalised Brier score regarding interpretation capability. It can be interpreted as how close our prediction to the mean relative to the standard deviation is.

3.3.4 Normalised Integrated Brier Score with κ-Truncation

We have developed two modifications of Brier score based on normalisation over the outputs variability, namely normalised Brier score and normalised centered Brier score. However, those measures have two crucial potential issues:

1. If $\pi_i^t = 1$ or $\pi_i^t = 0$, then the contribution of i is undefined,
2. If $\pi_i^t \approx 1$ or $\pi_i^t \approx 0$, then the contribution of i is too large.

We would discuss one potential solution to those problems by including additional condition $I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}$ for $\kappa \in [0,1]$ to the estimator of (3.6). We only discuss how the condition affect normalised Brier score since the results will also apply to normalised centered Brier score. By adding the condition to normalised

Brier score, we obtain

$$\begin{aligned}
&\widehat{\mathrm{NBS}}_n^{\kappa}(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[I_{\{X_i\le t\}} I_{\{T_i<D_i\}} I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}} \frac{\left(0-\pi_i^t\right)^2}{\pi_i^t(1-\pi_i^t)} \widehat{G}_n^{-1}(T_i) \right. \\
&\quad \left. + I_{\{X_i>t\}} I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}} \frac{\left(1-\pi_i^t\right)^2}{\pi_i^t(1-\pi_i^t)} \widehat{G}_n^{-1}(t) \right],
\end{aligned} \tag{3.18}$$

and

$$\widehat{\mathrm{NIBS}}_n^{\kappa} = \frac{1}{T_{max}-1}\sum_{t=1}^{T_{max}-1} \widehat{\mathrm{NBS}}_n^{\kappa}(t),$$

as the estimators of normalised Brier score and integrated normalised Brier score with κ-truncation, respectively.

However, we find issues if we want to apply some scenarios to (3.18) as follows:

(i) It is a biased estimator of (3.5) since we do not add penalty due to loss information from individuals who do not satisfy $I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}$.

(ii) If we apply IPCW due to $I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}$, we need to estimate

$$\mathbb{E}\left[I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}\right] = \mathbb{P}\left[I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}\right],$$

and then we include $1/\mathbb{P}\left[I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}\right]$ to (3.18). However, $\widehat{\mathbb{P}}\left[I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}\right]$ is not based on the outcomes T_i as we estimate G. In particular, it is a proportion of individuals whose $(1-\kappa)\le\pi_i^t\le\kappa$ over n individuals, which cannot be determined before we have π_i^t. So, this approach may be neither practical nor proper way to do.

(iii) Another approach is that we divide the numerator of (3.5) by (weighted) number of usable pairs instead of n, namely

$$\begin{aligned}
n_\kappa = \sum_{i=1}^{n} & I_{\{X_i\le t\}} I_{\{T_i<D_i\}} I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}} \widehat{G}_n^{-1}(T_i) \\
& + I_{\{X_i>t\}} I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}} \widehat{G}_n^{-1}(t),
\end{aligned}$$

so that we have

$$\begin{aligned}
&\widehat{\mathrm{NBS}}^{\kappa}_{n_\kappa}(t) \\
&= \frac{1}{n_\kappa}\sum_{i=1}^{n}\Bigg[I_{\{X_i\le t\}}I_{\{T_i<D_i\}}I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}\frac{\left(0-\pi_i^t\right)^2}{\pi_i^t(1-\pi_i^t)}\widehat{G}_n^{-1}(T_i) \\
&\quad + I_{\{X_i>t\}}I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}\frac{\left(1-\pi_i^t\right)^2}{\pi_i^t(1-\pi_i^t)}\widehat{G}_n^{-1}(t)\Bigg]
\end{aligned}$$

as an estimator of the following measure:

$$\begin{aligned}
&\mathrm{NBS}^{\kappa}(t) \\
&= \mathbb{E}\left[\frac{\left(\psi_i^t-\pi_i^t\right)^2 I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}}{\pi_i^t\left(1-\pi_i^t\right)}\right] \\
&= \mathrm{NBS}(t)\mathbb{E}\left[I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}\right] \\
&= \mathrm{NBS}(t)\mathbb{P}[I_{\{(1-\kappa)\le\pi_i^t\le\kappa\}}].
\end{aligned}$$

As we can see that $\mathrm{NBS}^{\kappa}(t)$ is a fraction of $\mathrm{NBS}(t)$. Thus, $\widehat{\mathrm{NBS}}^{\kappa}_{n_\kappa}(t)$ still have similar problems to $\widehat{\mathrm{NBS}}_n(t)$.

In this section, we have discussed a possible solution to the issues in modified integrated Brier scores, especially in the normalised versions, by ignoring individuals whose $(\pi_i^t \le 1-\kappa)$ or $(\pi_i^t \le \kappa)$ for some $\kappa \in [0,1]$. This approach has resulted in some crucial drawbacks so that the proposed κ-truncation modified integrated Brier score measures in this section are not proper as the alternative measures. Therefore, in the subsequent section, we will propose another approach that can cope with the issues of the modified integrated Brier scores.

3.3.5 Truncated Normalised Integrated Brier Score

Instead of removing some individuals from the sample as in the previous section, the measures proposed in this section still keep all individuals in the sample, but their predicted survival curves are dependent on some values of $\epsilon \in [0,1]$. As a result, they do not suffer from the problems occurred in (3.18) and (3.3.4). In

particular, we define

$$\begin{aligned}&\widehat{\text{NBS}}_n^{\varepsilon}(t)\\&= \frac{1}{n}\sum_{i=1}^{n}\left[I_{\{X_i\le t\}}I_{\{T_i<D_i\}}\frac{\left(0-\pi_i^{t;\varepsilon}\right)^2}{\pi_i^{t;\varepsilon}(1-\pi_i^{t;\varepsilon})}\widehat{G}_n^{-1}(T_i)\right.\\&\left.+\, I_{\{X_i>t\}}\frac{\left(1-\pi_i^{t;\varepsilon}\right)^2}{\pi_i^{t;\varepsilon}(1-\pi_i^{t;\varepsilon})}\widehat{G}_n^{-1}(t)\right]\end{aligned} \tag{3.19}$$

as truncated normalised Brier score, where

$$\pi_i^{t;\varepsilon} = \begin{cases}1-\epsilon, & \text{if } \pi_i^t \ge 1-\varepsilon\\ \epsilon, & \text{if } \pi_i^t \le \varepsilon\\ \pi_i^t, & \text{otherwise,}\end{cases} \tag{3.20}$$

for some $\varepsilon \in [0,1]$. Hence, we define

$$\widehat{\text{NIBS}}_n^{\varepsilon} = \frac{1}{T_{max}-1}\sum_{t=1}^{T_{max}-1}\widehat{\text{NBS}}_n^{\varepsilon}(t) \tag{3.21}$$

as truncated normalised integrated Brier score.

Following the same definition of $\pi_i^{t;\varepsilon}$ given in (3.20), we can replace π_i^t by $\pi_i^{t;\varepsilon}$ in $\widehat{\text{NCBS}}_n(t)$ and $\widehat{\text{NCIBS}}_n$ to obtain their truncated versions $\widehat{\text{NCBS}}_n^{\varepsilon}(t)$ and $\widehat{\text{NCIBS}}_n^{\varepsilon}$, respectively. We do not explicitly write the formulas since they are all the same except the predicted survival curves $\pi_i^{t;\varepsilon}$. Note that the truncated versions are exactly the same to the original non-truncated measures when $\varepsilon = 0$.

Numerical Implementations

This section is devoted to investigate how to determine the value of ε for the truncated normalised centered integrated Brier score. We moreover apply $\widehat{\text{NCIBS}}_n^{\varepsilon}$ to the first five test data obtained in Scenario 1 and 2 of Simulation 1, TCGA data, and breast cancer data, where their results are given by Table 3.1, Table 3.2, Table 3.3, and Table 3.4, respectively. Note that we still apply the same settings, Nnet-survival architectures, the train and test data in the previous sections.

The results for Scenario 1 of Simulation 1 can be seen in Table 3.1. There are two main models reported in the table, namely good models and overfitted models, for different values of ε and three discretisation setups, i.e. $\mathscr{D}_1, \mathscr{D}_2$, and

$\mathscr{D}_3$. The table's hyphen sign "-" means that the measures cannot be computed due to the division by zero in the denominator of the measures. We can see from Table 3.1 that generally, the values of $\widehat{\text{NCIBS}}_n^{\varepsilon}$ decrease as ε increase. If we continue the increase of ε, $\widehat{\text{NCIBS}}_n^{\varepsilon}$ will equal zero when $\varepsilon = 0.5$. We know that the closer $\widehat{\text{NCIBS}}_n^{\varepsilon}$ to zero, the better the model performance. On the other hand, the greater ε, the more $\pi_i^{t;\varepsilon}$ underestimates the original predicted survival curves π_i^t. Therefore, we must balance the trade-off between π_i^t and ε. We do not need high values of ε for the good models since $\widehat{\text{NCIBS}}_n^{\varepsilon}$ is slightly worse when $\varepsilon > 0.01$. Meanwhile, increasing ε reduces $\widehat{\text{NCIBS}}_n^{\varepsilon}$ in the overfitted models. We recommend 0.1 as the value for ε. By using $\varepsilon = 0.1$, $\widehat{\text{NCIBS}}_n^{\varepsilon}$ has a minimal change in the good models; meanwhile, in the overfitted models, $\widehat{\text{NCIBS}}_n^{\varepsilon}$ is much lower when $\varepsilon = 0.1$, and we can still see the apparent difference between the good and bad models.

For Scenario 2 of Simulation 1 (see Table 3.2), by the same approach in tuning the best ε, we recommend 0.01 as the value for ε because some values of $\widehat{\text{NCIBS}}_n^{\varepsilon}$ in the overfitted models, i.e. in $\mathscr{D}_4$ and $\mathscr{D}_5$, are very close to the values of $\widehat{\text{NCIBS}}_n^{\varepsilon}$ in the good models. Meanwhile, in the overfitted models, we can distinguish well between the good and bad models when $\varepsilon = 0.01$. For TCGA data (see Table 3.3), by the same arguments, we also recommend 0.01 as the best value for ε. For breast cancer data, we only report the results for $\lambda = 0$ with varying discretisation setups (see Table 3.4). In the breast cancer data, we do not have bad or good models since we only focus on $\lambda = 0$. Therefore, we will choose $\varepsilon = 0.01$ since it provides the closest $\widehat{\text{NCIBS}}_n^{\varepsilon}$ to zero.

To sum up, by tuning ε, we can find the best value of ε for $\widehat{\text{NCIBS}}_n^{\varepsilon}$. We do not have any rigid method to find the best ε. We have to conduct deeper analysis and to balance the trade-off between π_i^t and ε. In the next section, we will conduct further analysis regarding KM estimator as the reference value.

ε	Good Models			Overfitted Models		
	$\mathscr{D}_1$	$\mathscr{D}_2$	$\mathscr{D}_3$	$\mathscr{D}_1$	$\mathscr{D}_2$	$\mathscr{D}_3$
0	0.198	-0.101	-0.149	-	-	-
	-0.204	-0.039	-0.073	-	-	-
	-0.183	-0.023	0.017	-	-	-
	-0.167	0.026	0.172	-	-	-
	-0.221	-0.072	0.016	-	-	-
0.01	-0.198	-0.101	-0.149	17.212	22.223	15.937
	-0.204	-0.039	-0.073	15.872	20.714	14.203
	-0.186	-0.022	0.017	16.954	22.675	15.151
	-0.166	0.026	0.172	16.960	22.991	17.254
	-0.220	-0.072	0.016	16.040	21.791	16.148
0.1	-0.273	-0.182	-0.427	1.344	1.950	1.042
	-0.274	-0.142	-0.360	1.162	1.723	0.734
	-0.252	-0.099	-0.273	1.302	1.973	0.885
	-0.252	-0.101	-0.253	1.262	1.949	1.137
	-0.301	-0.165	-0.305	1.210	1.869	1.006
0.2	-0.316	-0.241	-0.490	0.268	0.538	0.111
	-0.313	-0.215	-0.458	0.188	0.441	-0.034
	-0.296	-0.178	-0.402	0.257	0.559	0.056
	-0.300	-0.188	-0.395	0.231	0.540	0.154
	-0.333	-0.226	-0.406	0.205	0.498	0.087
0.3	-0.293	-0.239	-0.422	-0.034	0.108	-0.122
	-0.288	-0.223	-0.408	-0.076	0.055	-0.202
	-0.279	-0.198	-0.372	-0.038	0.122	-0.146
	-0.283	-0.207	-0.373	-0.053	0.109	-0.099
	-0.302	-0.228	-0.370	-0.066	0.087	-0.135

*) If there is no number in the shell, the estimator cannot be computed.

Table 3.1: $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ for the first five test data in Scenario 1 Simulation 1.

ε	Good Models			Overfitted Models		
	$\mathscr{D}_4$	$\mathscr{D}_5$	$\mathscr{D}_6$	$\mathscr{D}_4$	$\mathscr{D}_5$	$\mathscr{D}_6$
0	-0.259	-0.459	-0.162	-	-	-
	-0.253	-0.758	-0.143	-	-	-
	-0.186	-0.672	-0.121	-	-	-
	-0.065	-0.627	-0.108	-	-	-
	-0.102	-0.458	-0.179	-	-	-
0.01	-0.257	-0.505	-0.163	11.690	1.319	18.909
	-0.252	-0.752	-0.142	10.205	1.243	17.734
	-0.185	-0.697	-0.123	11.627	0.655	19.523
	-0.064	-0.624	-0.107	11.609	1.179	20.010
	-0.124	-0.508	-0.178	11.434	0.763	18.827
0.1	-0.476	-0.722	-0.238	0.579	-0.632	1.437
	-0.435	-0.760	-0.229	0.401	-0.604	1.337
	-0.400	-0.792	-0.191	0.516	-0.704	1.487
	-0.380	-0.738	-0.198	0.600	-0.635	1.511
	-0.412	-0.748	-0.253	0.563	-0.683	1.478
0.2	-0.520	-0.666	-0.293	-0.086	-0.631	0.298
	-0.498	-0.683	-0.283	-0.168	-0.620	0.267
	-0.474	-0.700	-0.254	-0.126	-0.668	0.327
	-0.463	-0.671	-0.265	-0.079	-0.638	0.329
	-0.474	-0.682	-0.302	-0.096	-0.655	0.321
0.3	-0.443	-0.528	-0.282	-0.226	-0.507	-0.020
	-0.432	-0.536	-0.273	-0.264	-0.502	-0.035
	-0.416	-0.545	-0.255	-0.245	-0.528	-0.004
	-0.414	-0.529	-0.264	-0.220	-0.513	-0.004
	-0.416	-0.536	-0.286	-0.230	-0.522	-0.009

*) If there is no number in the shell, the estimator cannot be computed.

Table 3.2: $\widehat{\text{NCIBS}}_n^{\varepsilon}$ for the first five test data in Scenario 2 Simulation 1.

ε	Good Models			Overfitted Models		
	$\mathscr{D}_7$	$\mathscr{D}_8$	$\mathscr{D}_9$	$\mathscr{D}_7$	$\mathscr{D}_8$	$\mathscr{D}_9$
0	0.034	-0.030	-0.132	-	-	-
	-0.067	-0.096	-0.186	-	-	-
	-0.122	-0.174	-0.284	-	-	-
	-0.046	-0.120	-0.206	-	-	-
	-0.092	-0.135	-0.141	-	-	-
0.01	0.033	-0.032	-0.134	13.972	10.659	6.543
	-0.068	-0.097	-0.188	2.908	2.329	1.264
	-0.122	-0.175	-0.287	2.393	1.750	0.815
	-0.047	-0.121	-0.208	3.728	2.531	1.411
	-0.092	-0.135	-0.141	3.032	1.728	1.019
0.1	-0.053	-0.159	-0.299	1.210	0.692	0.071
	-0.151	-0.227	-0.358	-0.137	-0.348	-0.555
	-0.188	-0.270	-0.393	-0.299	-0.452	-0.627
	-0.128	-0.235	-0.351	-0.077	-0.297	-0.526
	-0.148	-0.237	-0.338	-0.146	-0.375	-0.550
0.2	-0.128	-0.237	-0.398	0.276	0.014	-0.289
	-0.194	-0.287	-0.431	-0.334	-0.467	-0.578
	-0.218	-0.303	-0.442	-0.416	-0.520	-0.617
	-0.180	-0.287	-0.426	-0.315	-0.436	-0.560
	-0.189	-0.297	-0.428	-0.338	-0.465	-0.570
0.3	-0.128	-0.237	-0.398	-0.009	-0.154	-0.317
	-0.194	-0.287	-0.431	-0.331	-0.411	-0.474
	-0.218	-0.303	-0.442	-0.372	-0.443	-0.496
	-0.180	-0.287	-0.426	-0.321	-0.394	-0.462
	-0.189	-0.297	-0.428	-0.332	-0.406	-0.468

*) If there is no number in the shell, the estimator cannot be computed.

Table 3.3: $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ for the first five test data in TCGA data.

ε	$\mathscr{D}_{\mathbf{10}}$	$\mathscr{D}_{\mathbf{11}}$	$\mathscr{D}_{\mathbf{12}}$
0	-	-	-
	-	-	-
	-	-	-
	-	-	-
	-	-	-
0.001	1.0020	0.0273	0.0473
	0.7388	-0.0002	0.0241
	0.9709	0.3774	0.3850
	0.8311	0.0300	0.0786
	0.6659	0.2179	0.3950
0.01	0.5411	-0.1189	0.0473
	0.4813	-0.1310	0.0242
	0.6094	-0.0043	0.0850
	0.4932	-0.1097	0.0787
	0.4470	-0.1231	0.0943
0.1	-0.6339	-0.6747	-0.4925
	-0.6243	-0.6665	-0.4917
	-0.6193	-0.6579	-0.4905
	-0.6332	-0.6742	-0.4922
	-0.6314	-0.6727	-0.4977
0.2	-0.6424	-0.6593	-0.5821
	-0.6384	-0.6559	-0.5811
	-0.6363	-0.6522	-0.5816
	-0.6421	-0.6592	-0.5818
	-0.6414	-0.6586	-0.5841
0.3	-0.5168	-0.5254	-0.4859
	-0.5147	-0.5236	-0.4857
	-0.5137	-0.5217	-0.4859
	-0.5166	-0.5253	-0.4860
	-0.5163	-0.5250	-0.4870

*) If there is no number in the shell, the estimator cannot be computed.

Table 3.4: $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ for the first five test data in the breast cancer data.

3.3.6 The Relationship between Normalised (Centered) Integrated Brier Score and KM Estimator

As discussed in Section 3.2, the integrated Brier score for KM estimator can be used as the reference value for integrated Brier score. The lesser integrated Brier score than the integrated Brier score for KM estimator, the better the predictive performance. In this section, we will show that the normalised Brier score and the normalised centered Brier score have the same calibration capability as their KM versions.

We first recall $\psi_i^t \in \{0, 1\}$. The predicted survival curve of KM estimator can be computed by

$$\pi = \frac{\sum_{i=1}^{n} \psi_i^t}{n},$$

for $i = 1, \cdots, n$. Thus, $\widehat{\text{NBS}}_n(t)$ for KM estimator at period $t \in \mathcal{T}$ can be computed as follows

$$\begin{aligned}
\widehat{\text{NBSKM}}_n(t) &= \frac{1}{n} \sum_{i=1}^{n} \frac{(\psi_i^t - \pi)^2}{\pi(1-\pi)} \\
&= \frac{1}{n\pi(1-\pi)} \sum_{i=1}^{n} (\psi_i^t - \pi)^2 \\
&= \frac{1}{\pi(1-\pi)} \left(\pi(1-\pi)^2 + (1-\pi)\pi^2\right) \\
&= (1-\pi) + \pi \\
&= 1.
\end{aligned} \tag{3.22}$$

Hence, the $\widehat{\text{NCBS}}_n(t)$ for KM estimator is defined as follows

$$\widehat{\text{NCBSKM}}_n(t) = \widehat{\text{NBSKM}}_n(t) - 1 = 0. \tag{3.23}$$

From the results in (3.22) and (3.23), we can obtain the normalised integrated Brier score for KM estimator ($\widehat{\text{NIBSKM}}_n = 1$) and the normalised centered integrated Brier score for KM estimator ($\widehat{\text{NCIBSKM}}_n = 0$). Since $\widehat{\text{NIBSKM}}_n = 1$ and $\widehat{\text{NCIBSKM}}_n = 0$ for any models, we will not compute them in all numerical experiments throughout this thesis. Furthermore, we are only interested in $\widehat{\text{NIBS}}_n$ or $\widehat{\text{NCIBS}}_n$.

Next, we will compare the relationships of the integrated Brier score and the normalised Brier score with their respective KM versions through a simple model as the illustration. Suppose we have only one covariate $Z_i \in \{0, 1\}$. We then assume that a fraction p of the data has covariate $Z_i = 1$ so that $\mathbb{P}[Z_i = 1] = p$ and $\mathbb{P}[Z_i = 0] = 1 - p$, where

$$\mathbb{P}\left[T_i > t \middle| Z_i = 1\right] = \widetilde{\pi}_{i_1}^t$$

and

$$\mathbb{P}\left[T_i > t \middle| Z_i = 0\right] = \widetilde{\pi}_{i_2}^t$$

are the true probabilities. Therefore, the expected value of Brier score at a fixed $t \in \mathcal{T}$ as defined in (2.6) for exact model is computed as follows

$$\begin{aligned}
&\mathbb{E}\left[(\psi_i^t - \pi^t)^2\right] \\
&= \mathbb{E}\left[(\psi_i^t - \widetilde{\pi}_{i_1}^t)^2 \middle| Z_i = 1\right] \mathbb{P}[Z_i = 1] \\
&\quad + \mathbb{E}\left[(\psi_i^t - \widetilde{\pi}_{i_2}^t)^2 \middle| Z_i = 0\right] \mathbb{P}[Z_i = 0] \\
&= \left[\widetilde{\pi}_{i_1}^t (1 - \widetilde{\pi}_{i_1}^t)^2 + (1 - \widetilde{\pi}_{i_1}^t)(1 - \widetilde{\pi}_{i_1}^t)^2\right] p + \widetilde{\pi}_{i_2}^t (1 - \widetilde{\pi}_{i_2}^t)(1 - p) \\
&= \widetilde{\pi}_{i_1}^t (1 - \widetilde{\pi}_{i_1}^t) p + \widetilde{\pi}_{i_2}^t (1 - \widetilde{\pi}_{i_2}^t)(1 - p).
\end{aligned}$$

We also need to calculate the following unconditional probability:

$$\begin{aligned}
&\mathbb{P}[T_i > t] \\
&= \mathbb{P}\left[T_i > t \middle| Z_i = 1\right] \mathbb{P}[Z_i = 1] + \mathbb{P}\left[T_i > t \middle| Z_i = 0\right] \mathbb{P}[Z_i = 0] \\
&= \widetilde{\pi}_{i_1}^t p + \widetilde{\pi}_{i_2}^t (1 - p) \\
&= \zeta^t.
\end{aligned}$$

Then, Brier score for KM estimator at a fixed $t \in \mathcal{T}$ is

$$(1 - \zeta^t)^2 \zeta^t + (0 - \zeta^t)^2 (1 - \zeta^t) = (1 - \zeta^t)\zeta^t.$$

If we denote $\mathbb{P}\left[Q = \widetilde{\pi}_{i_1}^t\right] = p$ and $\mathbb{P}\left[Q = \widetilde{\pi}_{i_2}^t\right] = 1 - p$, then we have

$$\mathbb{E}[Q(1 - Q)] = \mathbb{E}\left[\zeta^t(Q)\right],$$

and

$$(1 - \mathbb{E}[Q])\mathbb{E}[Q] = \zeta^t(\mathbb{E}[Q]),$$

where $\zeta^t(q) = (1-q)q = q - q^2$ is a strictly concave function. By Jensen's inequality, we have

$$\mathbb{E}\left[\zeta^t(Q)\right] < \zeta^t(\mathbb{E}\left[Q\right]), \tag{3.24}$$

but we can obviously quantify the difference for specific values of p, $\widetilde{\pi}_{i_1}^t$, and $\widetilde{\pi}_{i_1}^t$ to see the extent of the difference. From (3.24), we can see that the Brier score from the exact model is always lower than the Brier score for the KM estimator. Consider the normalised Brier score at a period t from the exact model as follows

$$\begin{aligned}
&\mathbb{E}\left[\frac{(\psi_i^t - \pi_i^t)^2}{\pi_i^t(1-\pi_i^t)}\right] \\
&= \mathbb{E}\left[\frac{(\psi_i^t - \pi_i^t)^2}{\pi_i^t(1-\pi_i^t)}\Big| Z_i = 1\right]\mathbb{P}\left[Z_i = 1\right] \\
&\quad + \mathbb{E}\left[\frac{(\psi_i^t - \pi_i^t)^2}{\pi_i^t(1-\pi_i^t)}\Big| Z_i = 0\right]\mathbb{P}\left[Z_i = 0\right] \\
&= \frac{\mathbb{E}\left[(\psi_i^t - \widetilde{\pi}_{i_1}^t)^2\big| Z_i = 1\right]}{\widetilde{\pi}_{i_1}^t(1-\widetilde{\pi}_{i_1}^t)}p \\
&\quad + \frac{\mathbb{E}\left[(\psi_i^t - \widetilde{\pi}_{i_2}^t)^2\big| Z_i = 0\right]}{\widetilde{\pi}_{i_2}^t(1-\widetilde{\pi}_{i_2}^t)}(1-p) \\
&= 1 \cdot p + 1 \cdot (1-p) \\
&= 1,
\end{aligned} \tag{3.25}$$

which is the same value as $\widehat{\text{NBSKM}}_n(t)$. Analogously, $\widehat{\text{NCBS}}_n(t)$ from the exact model equals $\widehat{\text{NCBSKM}}_n(t)$. The results show that both models, i.e. the exact model and the KM estimator, are perfectly calibrated.

In this chapter, we have discussed some pitfalls of the 'classical' integrated Briers score, leading us to propose the integrated Brier Score for the KM estimator as the reference value. While we typically interpret the integrated Brier score as the closer the value to zero, the better the model prediction performance, this is not always the case when using the integrated Brier score for the KM estimator. For example, we may observe the integrated Brier score close to zero in breast cancer data. However, the model performance is poor when considering the integrated Brier score for the KM estimator. The second part of this chapter discussed three modifications of the integrated Brier score. We developed these measures based on their advantage in interpretability. However, in practice, there

are still crucial issues with these modified integrated Brier scores. In the next chapter, we move to another type of performance measure in survival analysis, namely discrimination. In particular, we will develop an unbiased C-index that can be used to evaluate non-linear survival models that violate the PH assumption.

Chapter 4

Time-Dependent Uno's C-Index

In the previous chapter, we discussed the pitfalls of the integrated Brier score and the integrated Brier score for KM estimator as a solution to cope with such pitfalls. We moreover have proposed several novel calibration measures based on the integrated Brier score. In this chapter, we will discuss discrimination. We will first show that Uno's C-index is not a proper measure for evaluating the non-PH model when the PH assumption is violated in the data. Then, we will introduce time-dependent Uno's C-index to cope with the drawback of Uno's C-index. After showing the convergence of $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$, we will implement the measure through simulation studies and real-world examples.

4.1 Simulation 2: PH and Non-PH Data

We first recall Uno's C-index, $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$, as defined by (2.4.2) in Chapter 2. This simulation aims to assess the model performance using $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$ in two data types: PH and non-PH data. We used the PH data that have been generated in Simulation 1 (see Section 3.1.1 in Chapter 3). For the non-PH data, we modified slightly the data generation setting of Simulation 1. The PH assumption gets violated by changing the parameters of the event times in (3.1) in the following way. We set $\gamma = 0.001$ and made α dependent on covariates. We varied γ (the scale parameter) and α (the shape parameter) to obtain 1000 event times differently distributed from the generated event times in Simulation 1. In particular,

the formula of α in Simulation 2 is as follows

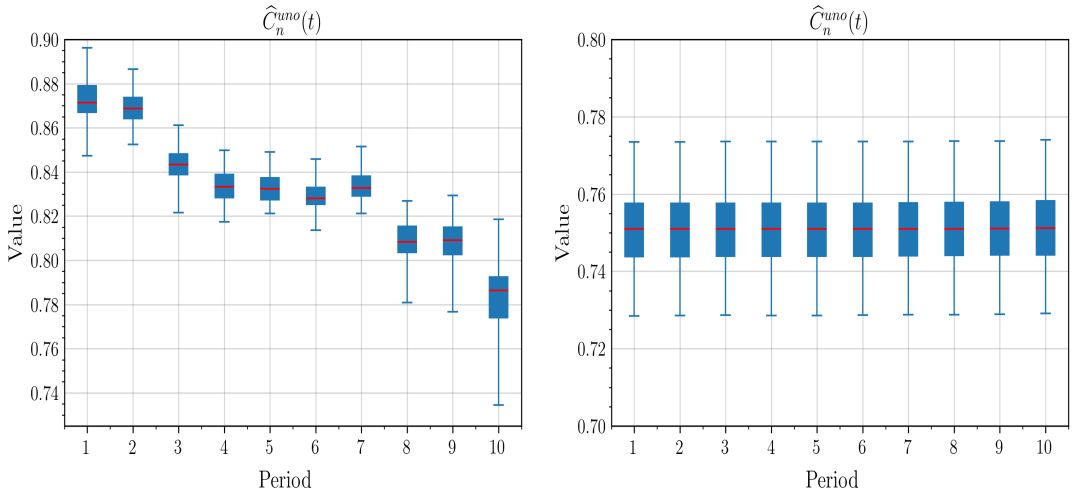
$$\alpha = \begin{cases} 0.1 & \text{if } Z_4 Z_5 \leq 0 \\ 0.4 & \text{otherwise,} \end{cases}$$

where we recall that the event time depends on five covariates $Z_1, Z_2, Z_3, Z_4 \sim \mathcal{N}(0,1)$ and $Z_5 \sim \mathcal{N}(0, 0.5)$). Because PH assumption is usually violated in the longer follow-up, we changed T^* from 70 in Simulation 1 to 150. For the other parameters, such as $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$, were kept to be the same as in Section 3.1.1. The observed times were discretised into 11 periods using $\mathcal{D}_{13}$={0, 5, 10, 15, 20, 25, 30, 40, 80, 90, 150, ∞} as the discretisation points.

We fitted the Nnet-survival with the architecture given in the left column of Table B.4. For the non-PH data, we used Nnet-survival with the architecture described in the right column of Table B.4. The architectures differ in the values of L^2 regularisation (λ), batch sizes, number of epochs, and learning rates. We can see from the table that the architecture for non-PH data has a smaller number of epochs, batch size, and λ, but it has a larger learning rate. As a result, the model might be more complex and not as good as the architecture for PH data. However, this unusual architecture has succeeded in exploring the non-PH behaviour of the fitted Nnet-survival.

There was a single train data ($n_{\text{train}} = 1000$) and 100 independent test data ($n_{\text{test}} = 1000$). We discretised those data using $\mathcal{D}_{13}$. The train and test data censoring rates are very close to 0%, so the effect of censoring can be minimised. Then, we computed $\widehat{\text{C}}_n^{\text{uno}}(t)$ for each $t \in \{1, \cdots, T_{max} - 1\}$ from the 100 test data. We present the model evaluation results of Nnet-survival over $\{1, \cdots, T_{max} - 1\}$ periods using $\widehat{\text{C}}_n^{\text{uno}}(t)$ in the generated non-PH data and PH data (Figure 4.1). Since $\widehat{\text{C}}_n^{\text{har}}(t)$ and $\widehat{\text{C}}_n^{\text{uno}}(t)$ are the same when the censoring rate is 0%, they will have the same properties, and hence we do not report $\widehat{\text{C}}_n^{\text{har}}(t)$ in this case study.

Figure 4.1 contains two panels, showing $\widehat{\text{C}}_n^{\text{uno}}(t)$ of Nnet-survival fitted to non-PH data (Panel (a)) and PH data (Panel (b)). Because $\widehat{\text{C}}_n^{\text{uno}}(t)$ assesses the model performance of two different data non-PH and PH data, we cannot compare them based on whether $\widehat{\text{C}}_n^{\text{uno}}(t)$ in Panel (a) is greater or smaller than in Panel (b). The differences between their variances happen because the discretisation setups are different. The non-PH data uses $\mathcal{D}_{13}$ meanwhile the PH data uses $\mathcal{D}_2$ defined in



(a) Non-PH Data

(b) PH Data

Figure 4.1: Boxplots of $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$ from non-PH data (a) and PH data (b) over $\{1, \cdots, T_{max} - 1\}$. They were computed on 100 independent test data (n_{test}=1000) from models fitted to a single fixed train data (n_{train}=1000).

Simulation 1. $\mathscr{D}_2$ results in the observed times of the PH data to be more evenly distributed than the non-PH data.

From Figure 4.1 , we mainly aim to see the variabilities (or the consistency) of the measures over the follow-up. As shown in Panel (a) of the figure, $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$ gives a volatile picture. The index changes significantly depending on which period is used to acquire the model prediction for survival probability. The oscillations happen because the rank reversion violates the PH assumption. Consequently, the reported model performance is inconsistent, leading to "C-hacking" where one may tend to report periods with good scores only (Sonabend *et al.*, 2022). On the other hand, in PH data (Panel(b) of Figure 4.1), $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$ at all periods are stable and mostly the same over time. Due to the proportionality behaviour in the datasets, the survival curves for each individual do not cross each other over the follow-up time. Thus, the order of any pairs of two different individuals was consistent at each period so that their contributions to $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$ were the same regardless the evaluated periods of interest.

Remark 4.1.1. Although $\widehat{\mathrm{C}}_n^{\mathrm{har}}(t)$ and $\widehat{\mathrm{C}}_n^{\mathrm{uno}}(t)$ may be easily employed by common standard survival analysis packages either in R or Python programming, we need to carefully check whether PH assumption holds in datasets. For instance, in the original paper of Nnet-survival by Gensheimer & Narasimhan (2019), the authors applied the continuous-time $\widehat{\mathrm{C}}_n^{\mathrm{har}}(t)$ for assessing model performance at one period (one-year prediction) on their real datasets even though the datasets violated the PH assumption.

4.2 Time-Dependent Uno's C-index

To derive the formula of time-dependent Uno's C-index, we apply a different approach to the original paper of time-dependent concordance by Antolini *et al.* (2005), which is defined as the weighted average of time-dependent AUC (Heagerty & Zheng, 2005). Consider the population probability of time-dependent concordance defined in (2.14) in Chapter 2 as follows

$$\mathrm{C} = \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}\big]. \tag{4.1}$$

As we have explained in Section 2.4.3, (4.1) is an overall measure instead of a time-specific measure as in (2.4).

We can rewrite the right-hand side of (4.1) as follows

$$\begin{aligned} &\frac{\mathbb{P}\big[I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}I_{\{T_i<T_j,T_i<T_{max}\}}\big]}{\mathbb{P}\big[I_{\{T_i<T_j,T_i<T_{max}\}}\big]} \\ &= \frac{\mathbb{E}\big[I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}I_{\{T_i<T_j,T_i<T_{max}\}}\big]}{\mathbb{E}\big[I_{\{T_i<T_j,T_i<T_{max}\}}\big]}. \end{aligned} \tag{4.2}$$

Then, the expected value of $\big(I_{\{T_i<D_i\}}/G(T_i)\big|T_i,\mathbb{Z}_i,T_j,\mathbb{Z}_j,D_j\big)$, where D_i is inde-

pendent of $T_i, \mathbb{Z}_i, T_j, \mathbb{Z}_j$, and D_j, is given by

$$\begin{aligned}
&\mathbb{E}\left[\frac{I_{\{T_i<D_i\}}}{G(T_i)}\Big|T_i, \mathbb{Z}_i, T_j, \mathbb{Z}_j, D_j\right] \\
&= \frac{1}{G(T_i)}\mathbb{E}\big[I_{\{T_i<D_i\}}\big|T_i, \mathbb{Z}_i, T_j, \mathbb{Z}_j, D_j\big] \\
&= \frac{1}{G(T_i)}\mathbb{E}\big[I_{\{T_i<D_i\}}\big|T_i\big] \\
&= \frac{1}{G(T_i)}\mathbb{E}\big[I_{\{T_i<D_i\}}\big|T_i\big] \\
&= \frac{1}{G(T_i)}G(T_i) \\
&= 1.
\end{aligned} \tag{4.3}$$

Since

$$\begin{aligned}
&\mathbb{E}\big[I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}I_{\{T_i<T_j,T_i<T_{max}\}}\big] \\
&\neq \mathbb{E}\big[I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}\big]\mathbb{E}\big[I_{\{T_i<T_j,T_i<T_{max}\}}\big],
\end{aligned}$$

and based on the results of (4.3), the numerator of the right-hand side of (4.2) can be rewritten as follows

$$\begin{aligned}
&\mathbb{E}\left[I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}I_{\{T_i<T_j,T_i<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{G(T_i)}\frac{I_{\{T_i<D_j\}}}{G(T_i)}\right] \\
&= \mathbb{E}\left[I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}I_{\{T_i<X_j,T_i<T_{max}\}}I_{\{T_i<D_i\}}G^{-2}(T_i)\right],
\end{aligned} \tag{4.4}$$

where $G^{-2}(T_i)$ can be also seen as the "penalty" term at T_i due to ignoring some pairs of individuals which are censored before we can observe the event of interest. By the same arguments, the denominator of the right-hand side of (4.2) can also be rewritten as follows

$$\begin{aligned}
&\mathbb{E}\left[I_{\{T_i<T_j,T_i<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{G(T_i)}\frac{I_{\{T_i<D_j\}}}{G(T_i)}\right] \\
&= \mathbb{E}\left[I_{\{T_i<X_j,T_i<T_{max}\}}I_{\{T_i<D_i\}}G^{-2}(T_i)\right].
\end{aligned} \tag{4.5}$$

Therefore, following (4.4) and (4.5), we can rewrite (4.1) as follows

$$\begin{aligned}
\mathrm{C} &= \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}\big] \\
&= \frac{\mathbb{E}\big[I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}I_{\{T_i<T_j,T_i<T_{max}\}}\big]}{\mathbb{E}\big[I_{\{T_i<T_j,T_i<T_{max}\}}\big]} \\
&= \frac{\mathbb{E}\big[I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}I_{\{T_i<X_j,T_i<T_{max}\}}I_{\{T_i<D_i\}}G^{-2}(T_i)\big]}{\mathbb{E}\big[I_{\{T_i<X_j,T_i<T_{max}\}}I_{\{T_i<D_i\}}G^{-2}(T_i)\big]}.
\end{aligned} \tag{4.6}$$

Following to the last equality of (4.6), we finally propose an estimator of C called time-dependent Uno's C-index as follows

$$\widehat{\mathrm{C}}_n^w = \frac{\sum_{i\neq j}^n I_{\{T_i<D_i\}} I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}} I_{\{T_i<X_j, T_i<T_{max}\}} \widehat{G}_n^{-2}(T_i)}{\sum_{i\neq j}^n I_{\{T_i<D_i\}} I_{\{T_i<X_j, T_i<T_{max}\}} \widehat{G}_n^{-2}(T_i)}, \tag{4.7}$$

where $\widehat{G}_n$ is an estimator of G, and the value is dependent on the sample size n. $\widehat{G}_n$ can be computed based on either train or test data with size n. However, throughout this thesis, we computed it based on independent test data, which is inline with most literatures (Gerds *et al.*, 2013; Uno *et al.*, 2011). Note that $\widehat{\mathrm{C}}_n^w$ can be also seen as the weighted version of the time-dependent concordance, $\widehat{\mathrm{C}}_n$.

4.3 Convergence of The Estimator

The uniform almost-sure convergence of U-processes theorem by Nolan & Pollard (1987) is essential to show the convergence of IPCW-based discrimination measure (Uno *et al.*, 2011). We apply the theorem and demonstrate that our proposed measure satisfies the assumptions of the theorem. We furthermore will show the convergence of $\widehat{\mathrm{C}}_n^w$ containing $\widehat{G}_n$ whose values depending on the size of n. In other words, throughout this thesis, we only consider that the values of $\widehat{G}_n$ can vary depending on n, and we do not discuss the case for which $\widehat{G}_n$ is fixed.

Before stating the convergence of $\widehat{\mathrm{C}}_n^w$ in Theorem, we first present definition, lemmas, and corollary that are relevant to the theorem. Note that we still consider Definition 2.4.1 and Theorem 2.4.3 discussed in Chapter 2 and use their notations in this Chapter.

Definition 4.3.1 (Envelope (Nolan & Pollard, 1987))**.** Consider class of symmetric functions $\mathscr{H}$. If $H(\cdot,\cdot) \geq |h(\cdot,\cdot)|$ for each $h \in \mathscr{H}$, then H is a positive envelope for $\mathscr{H}$.

Definition 4.3.2 (Covering Number (Nolan & Pollard, 1987))**.** Let H be a positive envelope of the class $\mathscr{H}$. For every $\delta > 0$, the covering number $N_p(\delta, Q, \mathscr{H}, H)$ with respect to measure Q such that $0 < Q(H^p) < \infty$ is defined as the smallest cardinality for a subclass $\mathscr{H}^*$ of $\mathscr{H}$ such that

$$\min_{h^*\in\mathscr{H}^*} Q\left|h - h^*\right|^p \leq \delta Q(H^p), \text{ for each } h \text{ in } \mathscr{H}.$$

Theorem 4.3.1 (Uniform Almost-Sure Convergence of U-processes (see Theorem 7 in page 787 in the paper by Nolan & Pollard (1987))). *Let $T_n(\cdot)$ be defined as*

$$T_n(h) = \sum_{1\le i\neq j\le n} \big(h(\boldsymbol{x}_{2i}, \boldsymbol{x}_{2j}) + h(\boldsymbol{x}_{2i}, \boldsymbol{x}_{2j-1}) + h(\boldsymbol{x}_{2i-1}, \boldsymbol{x}_{2j}) + h(\boldsymbol{x}_{2i-1}, \boldsymbol{x}_{2j-1})\big), \tag{4.8}$$

where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2n}$ are obtained by taking a double sample from a distribution F on $\mathscr{X}$, and h is a function in the symmetric class $\mathscr{H}$ with envelope H, respectively. For any h in $\mathscr{H}$, we define

$$F_n \otimes F(h) \;=\; n^{-1}\sum_{i=1}^{n} h(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

is the empirical measure over $\boldsymbol{x}_i$ for $i = 1, 2, \cdots, n$, and

$$F \otimes F(h) \;=\; \iint h(\boldsymbol{x}_i, \boldsymbol{x}_j)\, dF(\boldsymbol{x}_i)\, dF(\boldsymbol{x}_j)$$

is the expected value of $h(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with respect to $F \otimes F$.

If for each $\delta > 0$,

i. $\log N_1(\delta, T_n, H, H) = o_p(n)$

ii. $\log N_1(\delta, F_n \otimes F, \mathscr{H}, H) = o_p(n)$

iii. $N_1(\delta, F \otimes F, \mathscr{H}, H) < \infty$,

then

$$sup_{\mathscr{H}} \left| \left(\frac{\zeta_n}{n(n-1)} \right) - F \otimes F(h) \right| \overset{a.s.}{\to} 0,$$

as $n \to \infty$, where a.s. stands for convergence almost surely, and

$$\zeta_n(h) = \sum_{1\le i<j\le n} h(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

as defined in (2.16).

The sign "o_p" in Theorem 4.3.1 means the convergence in probability towards 0. For instance, when we have

$$\log N_1(\delta, T_n, H, H) = o_p(n)$$

in the theorem, it is the same as stating that

$$\lim_{n\to\infty}\left(\mathbb{P}\left[\left|\frac{\log N_1(\delta,T_n,H,H)}{n}\right|\right]\geq\epsilon\right)=0,$$

$\forall\epsilon>0$, or simply

$$\frac{\log N_1(\delta,T_n,H,H)}{n}\xrightarrow{p}0$$

as $n\to\infty$, where p stands for convergence in probability.

Next, we will demonstrate the convergence of the numerator of $\widehat{\mathrm{C}}_n^w$ by showing that all the assumptions of Theorem 4.3.1 are satisfied. In particular, the assumptions will be satisfied by the U-statistic with symmetrised kernel obtained from the numerator of $\widehat{\mathrm{C}}_n^w$.

Lemma 4.3.2 (Convergence of The Symmetrised Numerator of Time-dependent Uno's C-index)**.** *Define a class of functions as follows*

$$\mathcal{G}_\epsilon=\{g_n:\{1,\cdots,T_{max}-1\}\to[\epsilon,1]\}\tag{4.9}$$

for any n and $\forall\epsilon>0$. Assume that

$$g_n\xrightarrow{a.s.}G$$

as $n\to\infty$. Recall the terms inside the summation in the numerator of (4.7). For $i=1,2,\cdots$, we denote $\boldsymbol{x}_i=(\mathbb{Z}_i,T_i,D_i)$ so that $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots$ are independent samples from distribution F on $\mathscr{X}$. The terms can be rewritten as follows

$$h^g(\boldsymbol{x}_i,\boldsymbol{x}_j)=I_{\{T_i<D_i\}}I_{\{S(T_i;\mathbb{Z}_i)<S(T_i;\mathbb{Z}_j)\}}I_{\{T_i<X_j\}}g^{-2}(T_i),\tag{4.10}$$

where $g=\widehat{G}_n$ belongs to $\mathcal{G}_\epsilon$.

By regularity conditions (R1-R4) in section 2.1.3, for any $(\boldsymbol{x}_i,\boldsymbol{x}_j)$ in $\mathscr{X}\otimes\mathscr{X}$, we have

$$sup_{g\in\mathcal{G}_\epsilon}\left|\frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}\bar{h}^g(\boldsymbol{x}_i,\boldsymbol{x}_j)-F\otimes F(\bar{h}^G)\right|\xrightarrow{a.s.}0,$$

as $n\to\infty$, where $\bar{h}^g$ is the symmetrised version of h^g which belongs to

$$\mathscr{H}_\epsilon=\{h^g:g\in\mathcal{G}_\epsilon\},\tag{4.11}$$

and $\bar{h}^g$ belongs to

$$\bar{\mathscr{H}}_\epsilon=\{\bar{h}^g:(h^g+h^g)/2:h^g\in\mathscr{H}_\epsilon\}.\tag{4.12}$$

Proof of Lemma 4.3.2. The function g must be bounded below by $\epsilon > 0$ as given by the family $\mathcal{G}_\epsilon$ in 4.9. If it is not, h^g (and hence $\bar{h}^g$) will be undefined. Since $\bar{h}^g$ contains only some indicator functions except $g \in \mathcal{G}_\epsilon$, which is convergent almost surely to G, by Continuous Mapping Theorem (Shao, 2003) we have

$$\bar{h}^g \xrightarrow{a.s.} \bar{h}^G$$

as $n \to \infty$. Because

$$\frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \bar{h}^g(x_i, x_j)$$

is U-statistic with kernel $\bar{h}^g$ (see Definition 2.4.1) and $\bar{h}^G$ is the expected value of $\bar{h}^g$, we only need to show that $\bar{\mathscr{H}}_\epsilon$ with $F \otimes F$-integrable envelope $\bar{H}_\epsilon$ satisfies all the assumptions of Theorem 4.3.1 to prove Lemma 4.3.2.

First of all, we need to show that the first assumption of Theorem 4.3.1 is fulfilled. For any $\bar{h}^g \in \bar{\mathscr{H}}_\epsilon$,

$$\begin{aligned}
&T_n(\bar{h}^g) \\
&= \sum_{i \neq j}^{n} \left[\bar{h}^g(x_{2i}, x_{2j}) + \bar{h}^g(x_{2i}, x_{2j-1}) + \bar{h}^g(x_{2i-1}, x_{2j}) + \bar{h}^g(x_{2i-1}, x_{2j-1}) \right] \\
&= \sum_{i \neq j}^{n} \frac{1}{2} \Big[h^g(x_{2i}, x_{2j}) + h^g(x_{2j}, x_{2i}) + h^g(x_{2i}, x_{2j-1}) + h^g(x_{2j-1}, x_{2i}) \\
&\quad + h^g(x_{2i-1}, x_{2j}) + h^g(x_{2j}, x_{2i-1}) + h^g(x_{2i-1}, x_{2j-1}) + h^g(x_{2j-1}, x_{2i-1}) \Big].
\end{aligned}$$

As $\epsilon > 0$ is the minimum value of g, $\bar{H}^g = (1/\epsilon^2)$. Hence,

$$T_n(\bar{H}^g) = 4n(n-1)/\epsilon^2,$$

which is the value of T_n at the envelope $\bar{H}^g$.

To find the upper bound of $N_1(\delta, T_n, \bar{\mathscr{H}}_\epsilon, \bar{H}^g)$, we need to determine

$$\{\bar{h}^{g_1}, \bar{h}^{g_2}, \cdots, \bar{h}^{g_k}\} \subseteq \bar{\mathscr{H}}_\epsilon$$

such that $\forall \bar{h}^g \in \bar{\mathscr{H}}_\epsilon$ and every $\delta^* > 0$,

$$\begin{aligned}
\min_{1 \le k \le K} T_n\left(\left|\bar{h}^g - \bar{h}^{g_k}\right|\right) &\le \delta^* T_n(\bar{H}^g) \\
&= (\delta^*/\epsilon^2) 4n(n-1)
\end{aligned} \tag{4.13}$$

Meanwhile, for $1 \leq k \leq K$,

$$
\begin{aligned}
&T_n(\left|\bar{h}^g - \bar{h}^{g_k}\right|) \\
&= \sum_{i \neq j}^{n} \big[\left|\bar{h}^g\left(x_{2i}, x_{2j}\right) - \bar{h}^{g_k}\left(x_{2i}, x_{2j}\right)\right| \\
&\quad + \left|\bar{h}^g\left(x_{2i}, x_{2j-1}\right) - \bar{h}^{g_k}\left(x_{2i}, x_{2j-1}\right)\right| + | \, \bar{h}^g\left(x_{2i-1}, x_{2j}\right) \\
&\quad - \bar{h}^{g_k}\left(x_{2i-1}, x_{2j}\right) | + |\bar{h}^g\left(x_{2i-1}, x_{2j-1}\right) - \bar{h}^{g_k}\left(x_{2i-1}, x_{2j-1}\right) | \, \big] \\
&= \sum_{i \neq j}^{n} \frac{1}{2} \big[| \, \mathscr{K}\left(x_{2i}, x_{2j}\right) g^{-2}\left(x_{2i}\right) + \mathscr{K}\left(x_{2j}, x_{2i}\right) g^{-2}\left(x_{2j}\right) \\
&\quad - \mathscr{K}\left(x_{2i}, x_{2j}\right) g_k^{-2}\left(x_{2i}\right) - \mathscr{K}\left(x_{2j}, x_{2i}\right) g_k^{-2}\left(x_{2j}\right) | \\
&\quad + | \, \mathscr{K}\left(x_{2i}, x_{2j-1}\right) g^{-2}\left(x_{2i}\right) + \mathscr{K}\left(x_{2j-1}, x_{2i}\right) g^{-2}\left(x_{2j-1}\right) \\
&\quad - \mathscr{K}\left(x_{2i}, x_{2j-1}\right) g_k^{-2}\left(x_{2i}\right) - \mathscr{K}\left(x_{2j-1}, x_{2i}\right) g_k^{-2}\left(x_{2j-1}\right) | \\
&\quad + | \, \mathscr{K}\left(x_{2i-1}, x_{2j}\right) g^{-2}\left(x_{2i-1}\right) + \mathscr{K}\left(x_{2j}, x_{2i-1}\right) g^{-2}\left(x_{2j}\right) \\
&\quad - \mathscr{K}\left(x_{2i-1}, x_{2j}\right) g_k^{-2}\left(x_{2i-1}\right) - \mathscr{K}\left(x_{2j}, x_{2i-1}\right) g_k^{-2}\left(x_{2j}\right) | \\
&\quad + | \, \mathscr{K}\left(x_{2i-1}, x_{2j-1}\right) g^{-2}\left(x_{2i-1}\right) + \mathscr{K}\left(x_{2j-1}, x_{2i-1}\right) g^{-2}\left(x_{2j-1}\right) \\
&\quad - \mathscr{K}\left(x_{2i-1}, x_{2j-1}\right) g_k^{-2}\left(x_{2i-1}\right) - \mathscr{K}\left(x_{2j-1}, x_{2i-1}\right) g_k^{-2}\left(x_{2j-1}\right) | \big] \\
&\leq \sum_{i \neq j}^{n} \frac{1}{2} \big[\left|g^{-2}\left(x_{2i}\right) - g_k^{-2}\left(x_{2i}\right)\right| + \left|g^{-2}\left(x_{2j}\right) - g_k^{-2}\left(x_{2j}\right)\right| \\
&\quad + \left|g^{-2}\left(x_{2i}\right) - g_k^{-2}\left(x_{2i}\right)\right| + \left|g^{-2}\left(x_{2j-1}\right) - g_k^{-2}\left(x_{2j-1}\right)\right| \\
&\quad + \left|g^{-2}\left(x_{2i-1}\right) - g_k^{-2}\left(x_{2i-1}\right)\right| + \left|g^{-2}\left(x_{2j}\right) - g_k^{-2}\left(x_{2j}\right)\right| \\
&\quad + \left|g^{-2}\left(x_{2i-1}\right) - g_k^{-2}\left(x_{2i-1}\right)\right| + \left|g^{-2}\left(x_{2j-1}\right) - g_k^{-2}\left(x_{2j-1}\right)\right| \big] \\
&= \sum_{i \neq j}^{n} \big[\left|g^{-2}\left(x_{2i}\right) - g_k^{-2}\left(x_{2i}\right)\right| + \left|g^{-2}\left(x_{2i-1}\right) - g_k^{-2}\left(x_{2i-1}\right)\right| \big) \\
&\quad + \sum_{j=1}^{n} (n-1) \left|g^{-2}\left(x_{2j}\right) - g_k^{-2}\left(x_{2j}\right)\right| + \left|g^{-2}\left(x_{2j-1}\right) - g_k^{-2}\left(x_{2j-1}\right)\right| \\
&= \sum_{i=1}^{2n} (n-1) \left(\left|g^{-2}\left(x_i\right) - g_k^{-2}\left(x_i\right)\right| \right) + \sum_{j=1}^{2n} (n-1) \left(\left|g^{-2}\left(x_j\right) - g_k^{-2}\left(x_j\right)\right| \right],
\end{aligned}
$$

where the third inequality is due to triangle inequality and $0 \leq \mathscr{K}(.,.) \leq 1$.

Therefore,

$$\begin{aligned}
&\min_{1\le k\le K} T_n\left(\left|\bar{h}^{g}-\bar{h}^{g_k}\right|\right)\\
&\le \min_{1\le k\le K}\left(\sum_{i=1}^{2n}(n-1)\left(\left|g^{-2}(\boldsymbol{x}_i)-g_k^{-2}(\boldsymbol{x}_i)\right|\right)\right.\\
&\quad\left.+\sum_{j=1}^{2n}(n-1)\left(\left|g^{-2}(\boldsymbol{x}_j)-g_k^{-2}(\boldsymbol{x}_j)\right|\right)\right)\\
&\le (n-1)\left(\min_{1\le k\le K}\sum_{i=1}^{2n}\max_{1\le r\le 2n}\left(\left|g^{-2}(\boldsymbol{x}_r)-g_k^{-2}(\boldsymbol{x}_r)\right|\right)\right.\\
&\quad\left.+\min_{1\le k\le K}\sum_{j=1}^{2n}\max_{1\le r\le 2n}\left(\left|g^{-2}(\boldsymbol{x}_r)-g_k^{-2}(\boldsymbol{x}_r)\right|\right)\right)\\
&= 2n(n-1)\left(\min_{1\le k\le K}\max_{1\le r\le 2n}\left(\left|g^{-2}(\boldsymbol{x}_r)-g_k^{-2}(\boldsymbol{x}_r)\right|\right)\right.\\
&\quad\left.+\min_{1\le k\le K}\max_{1\le r\le 2n}\left(\left|g^{-2}(\boldsymbol{x}_r)-g_k^{-2}(\boldsymbol{x}_r)\right|\right)\right)\\
&= 4n(n-1)\min_{1\le k\le K}\max_{1\le r\le 2n}\left(\left|g^{-2}(\boldsymbol{x}_r)-g_k^{-2}(\boldsymbol{x}_r)\right|\right).
\end{aligned}$$

Thus, (4.13) holds if

$$\min_{1\le k\le K}\max_{1\le r\le 2n}\left(\left|g^{-2}(\boldsymbol{x}_r)-g_k^{-2}(\boldsymbol{x}_r)\right|\right)\le\delta, \tag{4.14}$$

where $\delta=(\delta^*/\epsilon^2)$ for every $\delta^*,\epsilon>0$.

Inequality (4.14) can be simply shown by finding g_k such that for each $g\in\mathcal{G}_\epsilon$, there exists k,

$$\max_{1\le r\le 2n}\left(\left|g^{-2}(\boldsymbol{x}_r)-g_k^{-2}(\boldsymbol{x}_r)\right|\right)\le\delta.$$

For simplification, let g^{-2} be denoted by w so that

$$\begin{aligned}
\mathscr{W} &= \left\{w: g^{-2}\text{ for } g\in\mathcal{G}_\epsilon\right\}\\
&= \left\{w:\{1,\cdots,T_{max}-1\}\to\left[1,\left(1/\epsilon^2\right)\right]: w(1)\ge 1,\right.\\
&\qquad \left. w\left(T_{max}-1\right)\le\left(1/\epsilon^2\right),\ w\text{ non-decreasing}\right\}.
\end{aligned}\tag{4.15}$$

Consider the r-th individual whose observed time equals $T_{max}-1$. So, maximum $T_{max}-1$ intervals with radius $\delta/2$ can be constructed such that the union of all

those intervals covers all possible values of $w(\boldsymbol{x}_r)$. As w is non-decreasing, the maximum possible outcomes of w over all $2n$ individuals is $(2n)^{T_{max}-1}$.

If we take a subset $\mathscr{W}^* = \{w_1, w_2, \cdots, w_k\}$ of $\mathscr{W}$ for $k \leq (2n)^{T_{max}-1}$, we can construct k intervals whose widths equal δ such that $\mathscr{W}^*$ are in the middle of the respective intervals and the union of the k intervals covers $\mathscr{W}$. In particular, for all $w \in \mathscr{W}$ there exists $w_k \in \mathscr{W}^*$ such that

$$\max_{1 \leq r \leq 2n} \left(\left| w(\boldsymbol{x}_r) - w_k(\boldsymbol{x}_r) \right| \right) \leq \delta,$$

which means that (4.13) holds. Hence,

$$N_1(\delta, T_n, \bar{\mathscr{H}_\epsilon}, \bar{H}^g) = k \leq (2n)^{T_{max}-1}. \tag{4.16}$$

Taking limit to infinity of log of the right hand side of (4.16) divided by n, we have

$$\begin{aligned} \lim_{n\to\infty} n^{-1} \log(2n)^{T_{max}-1} &= \lim_{n\to\infty} (T_{max} - 1) n^{-1} \log(2n) \\ &= (T_{max} - 1) \lim_{n\to\infty} (1/2n) \\ &= 0, \end{aligned} \tag{4.17}$$

where the last equality due to L'Hospital's rule.

From (4.16) and (4.17), we get the following inequality

$$\lim_{n\to\infty} n^{-1} \log N_1\big(\delta, T_n, \bar{\mathscr{H}_\epsilon}, \bar{H}^g\big) \leq \lim_{n\to\infty} n^{-1} \log(2n)^{T_{max}-1} = 0.$$

As the term $N_1\big(\delta, T_n, \bar{\mathscr{H}_\epsilon}, \bar{H}^g\big)$ is a non-negative integer,

$$\lim_{n\to\infty} n^{-1} \log N_1\big(\delta, T_n, \bar{\mathscr{H}_\epsilon}, \bar{H}^g\big) = 0$$

or

$$\log N_1\big(\delta, T_n, \bar{\mathscr{H}_\epsilon}, \bar{H}^g\big) = o_p(n),$$

which means that the first assumption of Theorem 4.3.1 is satisfied.

Next, we would show that the second assumption of Theorem 4.3.1 is satisfied. Note that all the expected values in this proof hold due to the regularity conditions (R1-R4) in section 2.1.3.

For $1 \leq k \leq K$, we have

$$\begin{aligned}
&F_n \otimes F\big(\,\big|\bar{h}^g - \bar{h}^{g_k}\big|\,\big) \\
&= F_n \otimes F\big(|\bar{h}^g(\boldsymbol{x}_i, \boldsymbol{x}_j) - \bar{h}^{g_k}(\boldsymbol{x}_i, \boldsymbol{x}_j)|\big) \\
&= F_n \otimes F\big(\frac{1}{2}|h^g(\boldsymbol{x}_i, \boldsymbol{x}_j) + h^g(\boldsymbol{x}_j, \boldsymbol{x}_i) - h^{g_k}(\boldsymbol{x}_i, \boldsymbol{x}_j) - h^{g_k}(\boldsymbol{x}_j, \boldsymbol{x}_i)|\big) \\
&= F_n \otimes F\big(\frac{1}{2}|\mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)g^{-2}(\boldsymbol{x}_i) - \mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)g_k^{-2}(\boldsymbol{x}_i) \\
&\quad + \mathscr{K}(\boldsymbol{x}_j, \boldsymbol{x}_i)g^{-2}(\boldsymbol{x}_j) - \mathscr{K}(\boldsymbol{x}_j, \boldsymbol{x}_i)g_k^{-2}(\boldsymbol{x}_j)|\big) \\
&\leq F_n \otimes F\big(\frac{1}{2}|\mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)g^{-2}(\boldsymbol{x}_i) - \mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)g_k^{-2}(\boldsymbol{x}_i)|\big) \\
&\quad + F_n \otimes F\big(\frac{1}{2}|\mathscr{K}(\boldsymbol{x}_j, \boldsymbol{x}_i)g^{-2}(\boldsymbol{x}_j) - \mathscr{K}(\boldsymbol{x}_j, \boldsymbol{x}_i)g_k^{-2}(\boldsymbol{x}_j)|\big) \\
&= F_n \otimes F\big(|V_1^g - V_1^{g_k}|\big) + F_n \otimes F\big(|V_2^g - V_2^{g_k}|\big),
\end{aligned} \tag{4.18}$$

where for any pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathscr{X} \otimes \mathscr{X}$,

$$V_1^g(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2}\mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)g^{-2}(\boldsymbol{x}_i),$$

and

$$V_2^g(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2}\mathscr{K}(\boldsymbol{x}_j, \boldsymbol{x}_i)g^{-2}(\boldsymbol{x}_j).$$

Therefore,

$$\begin{aligned}
&(V_1^g + V_2^g)(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
&= \frac{1}{2}\big(\mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)g^{-2}(\boldsymbol{x}_i) + \mathscr{K}(\boldsymbol{x}_j, \boldsymbol{x}_i)g^{-2}(\boldsymbol{x}_j)\big) \\
&= \bar{h}^g(\boldsymbol{x}_i, \boldsymbol{x}_j).
\end{aligned}$$

Moreover, we obtain

$$\begin{aligned}
&\mathscr{V}_1^g + \mathscr{V}_2^g \\
&= \big\{V_1^g + V_2^g : V_1^g \in \mathscr{V}_1^g, V_2^g \in \mathscr{V}_2^g\big\} \\
&= \big\{\bar{h}^g : g \in \mathcal{G}_\epsilon^{\mathcal{C}}\big\} \\
&= \bar{\mathscr{H}}_\epsilon.
\end{aligned}$$

We now would find the upper bound of the covering number

$$N_1(\delta_1, F_n \otimes F, \mathscr{V}_1^g, \bar{V}_1^g)$$

for each $\delta_1 > 0$, where $\bar{V}_1^g$ is the envelope of $\mathscr{V}_1^g$. Furthermore, for each $\delta_1^* > 0$, we need to determine

$$\{V_1^{g_1}, V_1^{g_2}, \cdots, V_1^{g_k}\} \subseteq \mathscr{V}_1^g$$

such that $\forall V_1^g \in \mathscr{V}_1^g$,

$$\min_{1\leq k\leq K} F_n \otimes F\left(\left|V_1^g - V_1^{g_k}\right|\right) \leq \delta_1^* F_n \otimes F(\bar{V}_1^g). \tag{4.19}$$

Since $F_n \otimes F(\bar{V}_1^g)$ equals $(1/2\epsilon^2)$, then (4.19) can be rewritten by

$$\min_{1\leq k\leq K} F_n \otimes F\left(\left|V_1^g - V_1^{g_k}\right|\right) \leq \delta_1, \tag{4.20}$$

for every $\delta_1 = \delta_1^*/(2\epsilon^2) > 0$ and $\epsilon > 0$.

By considering the following expectation

$$\begin{aligned}
&F_n \otimes F\big(|V_1^g - V_1^{g_k}|\big) \\
&= F_n \otimes F\left(\left|V_1^g(\boldsymbol{x}_i, \boldsymbol{x}_j) - V_1^{g_k}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right|\right) \\
&= \frac{1}{2n}\sum_{i=1}^{n}\left|\mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) g^{-2}(\boldsymbol{x}_i) - \mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) g_k^{-2}(\boldsymbol{x}_i)\right| \\
&\leq \frac{1}{2n}\sum_{i=1}^{n}\left|g^{-2}(\boldsymbol{x}_i) - g_k^{-2}(\boldsymbol{x}_i)\right|,
\end{aligned}$$

we obtain

$$\begin{aligned}
&\min_{1\leq k\leq K} F_n \otimes F\left(|V_1^g - V_1^{g_k}|\right) \\
&\leq \min_{1\leq k\leq K}\left(\frac{1}{2n}\sum_{i=1}^{n}\left|g^{-2}\left(x_i\right) - g_k^{-2}\left(x_i\right)\right|\right) \\
&\leq \frac{1}{2n}\min_{1\leq k\leq K}\left(\sum_{i=1}^{n}\max_{1\leq r\leq n}\left|g^{-2}\left(x_r\right) - g_k^{-2}\left(x_r\right)\right|\right) \\
&= \frac{1}{2}\min_{1\leq k\leq K}\max_{1\leq r\leq n}\left|g^{-2}\left(x_r\right) - g_k^{-2}\left(x_r\right)\right|
\end{aligned} \tag{4.21}$$

so that if

$$\frac{1}{2}\min_{1\leq k\leq K}\max_{1\leq r\leq n}\left|g^{-2}(x_r) - g_k^{-2}(x_r)\right| \leq \delta_1, \tag{4.22}$$

then (4.20) holds. This can simply be shown by finding a function g_k such that for each $g \in \mathcal{G}_\epsilon$, there exists k which satisfies

$$\max_{1\leq r\leq n}\left(\left|g^{-2}(x_r) - g_k^{-2}(x_r)\right|\right) \leq \delta_1.$$

To show this statement, we again consider a family $\mathscr{W}$ as defined in (4.15). Since w is non-decreasing, the maximum number of functions that are in the centre of the sub-intervals over $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ is $n^{T_{max}-1}$. As a consequence, for all $w \in \mathscr{W}$ there exists

$$w_k \in \mathscr{W}^* = \{w_1, w_2, \cdots, w_k\} \subseteq \mathscr{W}$$

such that

$$\max_{1 \le r \le n} \left| w(\boldsymbol{x}_r) - w_k(\boldsymbol{x}_r) \right| \le \delta_1.$$

Hence, (4.20) holds, which means that

$$N_1(\delta_1, F_n \otimes F, \mathscr{V}_1^g, \bar{V}_1^g) = k, \tag{4.23}$$

where $k \le n^{T_{max}-1}$.

After that, we would find the upper bound of the covering number

$$N_2(\delta_2, F_n \otimes F, \mathscr{V}_2^g, \bar{V}_2^g),$$

where $\bar{V}_2^g$ is the envelope of $\mathscr{V}_2^g$ for each $\delta_2 > 0$. We consider the second term of the last inequality of (4.18) so that for $1 \le k \le K$, we obtain

$$\begin{aligned}
& F_n \otimes F\left(\left|V_2^g - V_2^{g_k}\right|\right) \\
&= F_n \otimes F\left(\left|V_2^g(\boldsymbol{x}_i, \boldsymbol{x}_j) - V_2^{g_k}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right|\right) \\
&= F_n \otimes F\left(\frac{1}{2}\left|\mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) g^{-2}(\boldsymbol{x}_j) - \mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) g_k^{-2}(\boldsymbol{x}_j)\right|\right) \\
&\le F_n \otimes F\left(\frac{1}{2}\left|g^{-2}(\boldsymbol{x}_j) - g_k^{-2}(\boldsymbol{x}_j)\right|\right).
\end{aligned}$$

Therefore,

$$\min_{1 \le k \le K} F_n \otimes F\left(\left|V_2^g - V_2^{g_k}\right|\right) \le \min_{1 \le k \le K} F_n \otimes F\left(\frac{1}{2}\left|g^{-2}(\boldsymbol{x}_j) - g_k^{-2}(\boldsymbol{x}_j)\right|\right).$$

For each $\delta_2 > 0$, assuming that there exists k such that

$$\frac{1}{2} \min_{1 \le k \le K} F_n \otimes F\left(\frac{1}{2}\left|g^{-2}(\boldsymbol{x}_j) - g_k^{-2}(\boldsymbol{x}_j)\right|\right) \le \delta_2 F_n \otimes F(\bar{V}_2^g),$$

where $\bar{V}_2^g$ is the envelope of $\mathscr{V}_2^g$, and $\bar{V}_1^g + \bar{V}_2^g = \bar{H}^g$.

Recall the family $\mathscr{W}$ defined in (4.15). As w is non-decreasing, there exists

$$\mathscr{W}^* = \{w_1, w_2, \cdots, w_k\} \subseteq \mathscr{W}$$

such that

$$\bigcup_{r=1}^{k} \left[w_r - (\delta_2/2), w_r + (\delta_2/2)\right]$$

covers $\mathscr{W}$. In other words, there exists $w_k \in \mathscr{W}^*$ such that

$$\max_r \left|w(x_r) - w_k(x_r)\right| \leq \delta_2$$

for all $w \in \mathscr{W}$. Thus,

$$\min_{1\leq k\leq K} F_n \otimes F\big(|V_2^g - V_2^{g_k}|\big) \leq \delta_2 F_n \otimes F\big(\bar{V}_2^g\big),$$

or

$$N_1(\delta_2, F_n \otimes F, \mathscr{V}_2^g, \bar{V}_2^g) = k.$$

If k is infinite, there always exists a natural number $\tilde{n}$ such that

$$(k - \tilde{n}) = \infty$$

and

$$\bigcup_{r=1}^{k-\tilde{n}} \left[w_r - (\delta_2/2), w_r + (\delta_2/2)\right]$$

covers $\mathscr{W}$. This result contradicts the definition of covering number in which k must be the minimum number of sub-intervals with radius $(\delta_2/2)$ that cover $\mathscr{W}$ (see Definition (4.3.2)). As a consequence, k must be finite, that is

$$N_1(\delta_2, F_n \otimes F, \mathscr{V}_2^g, \bar{V}_2^g) = k, \tag{4.24}$$

where $k < \infty$.

By Lemma 16 in the paper by Nolan & Pollard (1987) and the results shown by (4.23) and (4.24), we have

$$\begin{aligned} & N_1(\delta, F_n \otimes F, \bar{\mathscr{H}_\epsilon}, \bar{H}^g) \\ &= N_1(2\delta_1 + 2\delta_2, F_n \otimes F, \mathscr{V}_1^g + \mathscr{V}_2^g, \bar{V}_1^g + \bar{V}_2^g) \\ &\leq N_1(\delta_1, F_n \otimes F, \mathscr{V}_1^g, \bar{V}_1^g) N_1(\delta_2, F_n \otimes F, \mathscr{V}_2^g, \bar{V}_2^g) \\ &\leq n^{T_{max}-1} k, \end{aligned}$$

for each $\delta = 2\delta_1 + 2\delta_2 > 0$ and $k < \infty$. Therefore,

$$\begin{aligned} \lim_{n\to\infty} n^{-1} \log(k n^{T_{max}-1}) &= k \lim_{n\to\infty} (T_{max} - 1) n^{-1} \log(n) \\ &= k(T_{max} - 1) \lim_{n\to\infty} (1/n) \\ &= 0, \end{aligned}$$

where in the last equality we use L'Hospital's rule. Thus,

$$\begin{aligned}\lim_{n\to\infty} n^{-1}\log N_1\big(\delta, F_n\otimes F, \bar{\mathscr{H}_\epsilon}, \bar{H}^g\big) &\leq \lim_{n\to\infty} n^{-1}\log(n^{T_{max}-1})\\ &= 0,\end{aligned}$$

or

$$\log N_1\big(\delta, F_n\otimes F, \bar{\mathscr{H}_\epsilon}, \bar{H}^g\big) = o_p(n)$$

so that the second assumption of Theorem 4.3.1 is satisfied.

Finally, we need to show that the third assumption of Theorem 4.3.1 is satisfied. For $1\leq k\leq K$,

$$\begin{aligned}&F\otimes F\Big(\big|\bar{h}^g-\bar{h}^{g_k}\big|\Big)\\ &= F\otimes F\Big(\big|\bar{h}^g(\boldsymbol{x}_i,\boldsymbol{x}_j)-\bar{h}^{g_k}(\boldsymbol{x}_i,\boldsymbol{x}_j)\big|\Big)\\ &= F\otimes F\Big(\frac{1}{2}\big|\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)g^{-2}(\boldsymbol{x}_i)+\mathscr{K}(\boldsymbol{x}_j,\boldsymbol{x}_i)g^{-2}(\boldsymbol{x}_j)\\ &\quad -\mathscr{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)g_k^{-2}(\boldsymbol{x}_i)-\mathscr{K}(\boldsymbol{x}_j,\boldsymbol{x}_i)g_k^{-2}(\boldsymbol{x}_j)\big|\Big)\\ &\leq F\otimes F\Big(\frac{1}{2}\big|g^{-2}(\boldsymbol{x}_i)+g^{-2}(\boldsymbol{x}_j)-g_k^{-2}(\boldsymbol{x}_i)-g_k^{-2}(\boldsymbol{x}_j)\big|\Big),\end{aligned}$$

where the last inequality is because of $0\leq\mathscr{K}(.,.)\leq 1$. Thus,

$$\begin{aligned}&\min_{1\leq k\leq K} F\otimes F\Big(\big|\bar{h}^g-\bar{h}^{g_k}\big|\Big)\\ &\leq \min_{1\leq k\leq K} F\otimes F\Big(\frac{1}{2}\big|g^{-2}(\boldsymbol{x}_i)+g^{-2}(\boldsymbol{x}_j)-g_k^{-2}(\boldsymbol{x}_i)-g_k^{-2}(\boldsymbol{x}_j)\big|\Big).\end{aligned}$$

Assuming there exists k such that

$$\min_{1\leq k\leq K} F\otimes F\Big(\frac{1}{2}\big|g^{-2}(\boldsymbol{x}_i)+g^{-2}(\boldsymbol{x}_j)-g_k^{-2}(\boldsymbol{x}_i)-g_k^{-2}(\boldsymbol{x}_j)\big|\Big)\leq \delta F\otimes F(\bar{H}^g),$$

and where $\bar{H}^g$ is the envelope of $\bar{\mathscr{H}_\epsilon}$. By the same arguments of showing (4.24), we have

$$N_1(\delta, F\otimes F, \bar{\mathscr{H}_\epsilon}, \bar{H}^g) = k<\infty,$$

meaning that the third assumption is fulfilled.

Because all assumptions of Theorem 4.3.1 are satisfied,

$$\sup_{g\in\mathcal{G}_\epsilon}\left|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}\bar{h}^g(\boldsymbol{x}_i,\boldsymbol{x}_j)-F\otimes F(\bar{h}^G)\right|\xrightarrow{a.s.}0 \tag{4.25}$$

as $n\to\infty$. □

We have demonstrated the convergence of the numerator of $\widehat{\mathrm{C}}_n^w$ as given by Lemma 4.3.2. By the same assumptions and arguments, the convergence of the denominator of $\widehat{\mathrm{C}}_n^w$ also follows the results of the lemma because their difference is only the indicator function $I_{\{S(T_i;Z_i)<S(T_i;Z_j)\}}$. We moreover state it formally in the following corollary.

Corollary 4.3.3 (Convergence of The Symmetrised Denominator of Time-dependent Uno's C-index)**.** *Denote $\boldsymbol{x}_i=(\mathbb{Z}_i,T_i,D_i)$ for $i=1,2,\cdots$, so that $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots$ are independent samples from distribution F on $\mathscr{X}$. Recall the terms inside the summation in the denominator of (4.7), and rewrite the terms as follows*

$$\psi^g(\boldsymbol{x}_i,\boldsymbol{x}_j)=I_{\{T_i<D_i\}}I_{\{T_i<X_j\}}g^{-2}(T_i), \tag{4.26}$$

where $g=\widehat{G}_n$ belongs to $\mathcal{G}_\epsilon=\{g_n:\{1,\cdots,T_{max}-1\}\to[\epsilon,1]\}$, $\forall\epsilon>0$, and we assume that $g_n\xrightarrow{a.s.}G$ as $n\to\infty$. By regularity conditions (R1-R4) in section 2.1.3, for any $(\boldsymbol{x}_i,\boldsymbol{x}_j)$ in $\mathscr{X}\otimes\mathscr{X}$, we have

$$sup_{g\in\mathcal{G}_\epsilon}\left|\frac{2}{n(n-1)}\sum_{1\le i<j\le n}\bar{\psi}^g(\boldsymbol{x}_i,\boldsymbol{x}_j)-F\otimes F(\bar{\psi}^G)\right|\xrightarrow{a.s.}0,$$

as $n\to\infty$, where $\bar{\psi}^g$ is the symmetrised version of ψ^g which belongs to

$$\Psi_\epsilon=\{\psi^g:g\in\mathcal{G}_\epsilon\}, \tag{4.27}$$

and $\bar{\psi}^g$ belongs to

$$\bar{\Psi}_\epsilon=\{\bar{\psi}^g:(\psi^g+\psi^g)/2:\psi^g\in\Psi_\epsilon\}. \tag{4.28}$$

In Lemma 4.3.2 and Corollary 4.3.3, we have demonstrated the convergence of U-statistics whose kernels are the symmetrised terms inside the summations in the numerator and denominator of $\widehat{\mathrm{C}}_n^w$, respectively. However, we need the convergence of the non-symmetrised terms inside the summations. Since the convergence of the symmetrised functions results in the convergence of the respective

non-symmetrised functions, the lemma and corollary guarantee the convergence of the quantities of interest. Moreover, the lemma and corollary results will be used in the following theorem.

Theorem 4.3.4 (Convergence of Time-dependent Uno's C-index)**.** *Define a family of functions as follows*

$$\mathcal{G}_\epsilon = \{g_n : \{1, \cdots, T_{max} - 1\} \to [\epsilon, 1]\},$$

$\forall \epsilon > 0$, *where* g_n *is any estimator of* G*, and its value is depending on sample size* n*. Assume that* $g_n \xrightarrow{a.s.} G$ *as* $n \to \infty$*. By regularity conditions (R1-R4) in section 2.1.3, we have*

$$\widehat{C}_n^w \xrightarrow{a.s.} C$$

as $n \to \infty$*.*

Proof of Theorem 4.3.4. We first denote $\boldsymbol{x}_i = (\mathbb{Z}_i, T_i, D_i)$ for $i = 1, 2, \cdots$, so that $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ are independent samples from distribution F on $\mathscr{X}$. Recall h^g as given in (4.10) as follows

$$h^g(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{T_i < D_i\}} I_{\{S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\}} I_{\{T_i < X_j, T_i < T_{max}\}} g^{-2}(T_i),$$

where $g = \widehat{G}_n \in \mathcal{G}_\epsilon$, and h^g belongs to $\mathscr{H}_\epsilon = \{h^g : g \in \mathcal{G}_\epsilon\}$ defined in (4.11), which is a class of functions on $\mathscr{X} \otimes \mathscr{X}$. According to Definition 2.4.1,

$$\frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \bar{h}^g(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

is U-statistic with kernel $\bar{h}^g$, where

$$\bar{h}^g\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) = \frac{1}{2}\big(h^g(\boldsymbol{x}_i, \boldsymbol{x}_j) + h^g(\boldsymbol{x}_j, \boldsymbol{x}_i)\big)$$

is the symmetrised version of h^g, and $\bar{h}^g \in \bar{\mathscr{H}}_\epsilon = \{\bar{h}^g : (h^g + h^g)/2 : h^g \in \mathscr{H}_\epsilon\}$ as defined in (4.12).

Denote

$$\bar{h}^g(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2}\big(\mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) g^{-2}(\boldsymbol{x}_i) + \mathscr{K}(\boldsymbol{x}_j, \boldsymbol{x}_i) g^{-2}(\boldsymbol{x}_j)\big),$$

where

$$\mathscr{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{T_i < D_i\}} I_{\{S(T_i;Z_i) < S(T_j,Z_j)\}} I_{\{T_i < X_j\}},$$

and $g^{-2}(\boldsymbol{x}_i) = g^{-2}(T_i)$. By regularity conditions (R1-R4) and Lemma 4.3.2, we have

$$\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{h}^g(\boldsymbol{x}_i, \boldsymbol{x}_j) \xrightarrow{a.s.} F \otimes F(\bar{h}^g)$$

as $n \to \infty$, where

$$\bar{h}^g(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2}\big(h^g(\boldsymbol{x}_i, \boldsymbol{x}_j) + h^g(\boldsymbol{x}_j, \boldsymbol{x}_i)\big)$$

is the symmetrised version of h^g.

The numerator of the statistic $\widehat{\mathrm{C}}_n^w$ equals

$$\sum_{i\neq j}^{n} h^{\widehat{G}_n}(x_i, x_j) = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{h}^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

We have

$$\begin{aligned}
&\left|\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{h}^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j) - F \otimes F(\bar{h}^G)\right| \\
&\le \left|\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{h}^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j) - F \otimes F(\bar{h}^{\widehat{G}_n})\right| + \left|F \otimes F(\bar{h}^{\widehat{G}_n}) - F \otimes F(\bar{h}^G)\right| \\
&\le \sup_{g\in\mathcal{G}_\epsilon} \left|\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{h}^g(\boldsymbol{x}_i, \boldsymbol{x}_j) - F \otimes F(\bar{h}^g)\right| + \left|F \otimes F(\bar{h}^{\widehat{G}_n}) - F \otimes F(\bar{h}^G)\right|.
\end{aligned}$$

By Lemma 4.3.2, the first term above converges a.s. to 0 as $n \to \infty$. For the convergence of the second term, we apply the dominated convergence theorem and almost sure uniform convergence of $\widehat{G}_n$ to G assumed in the statement of the theorem. In conclusion, we have

$$\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{h}^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j) \xrightarrow{a.s.} F \otimes F(\bar{h}^G) \tag{4.29}$$

as $n \to \infty$.

Recall ψ^g as given in (4.26) as follows

$$\psi^g(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{T_i<D_i\}} I_{\{T_i<X_j\}} g^{-2}(T_i). \tag{4.30}$$

By regularity conditions (R1-R4) and Corollary 4.3.3, we get

$$\sup_{g\in\mathcal{G}_\epsilon} \left|\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{\psi}^g(\boldsymbol{x}_i, \boldsymbol{x}_j) - F \otimes F(\bar{\psi}^G)\right| \xrightarrow{a.s.} 0$$

as $n \to \infty$, where $\bar{\psi}^g$ is the symmetrisation of $\psi^g(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{T_i<D_i\}} I_{\{T_i<X_j\}} g^{-2}(T_i)$.

Thus, we argue similarly for the denominator of $\widehat{\mathrm{C}}_n^w$ which is equal to

$$\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{\psi}^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

We have

$$\begin{aligned}
&\left| \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{\psi}^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j) - F \otimes F(\bar{\psi}^G) \right| \\
&\le \left| \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{\psi}^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j) - F \otimes F(\bar{\psi}^{\widehat{G}_n}) \right| \\
&\quad + \left| F \otimes F(\bar{\psi}^{\widehat{G}_n}) - F \otimes F(\bar{\psi}^G) \right| \\
&\le \sup_{g \in \mathcal{G}_\epsilon} \left| \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{\psi}^g(\boldsymbol{x}_i, \boldsymbol{x}_j) - F \otimes F(\bar{\psi}^g) \right| \\
&\quad + \left| F \otimes F(\bar{\psi}^{\widehat{G}_n}) - F \otimes F(\bar{\psi}^G) \right|
\end{aligned}$$

The first term on the right-hand side converges to 0 a.s. as $n \to \infty$ by Corollary 4.3.3. For the convergence of the second term, we use again the dominated convergence theorem.

Finally, we apply the Continuous Mapping Theorem (Shao, 2003) to conclude that

$$\frac{\sum_{1\le i<j\le n} h^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sum_{1\le i<j\le n} \psi^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j)} \xrightarrow{a.s.} \frac{F \otimes F(h^G)}{F \otimes F(\psi^G)}$$

as $n \to \infty$. It remains to note that the right-hand side equals

$$\begin{aligned}
&\frac{\mathbb{E}\big[I_{\{S(T_i;Z_i)<S(T_i;Z_j)\}} I_{\{T_i<T_j, T_i<T_{max}\}} \frac{I_{\{T_i<D_i\}}}{g(T_i)} \frac{I_{\{T_i<D_j\}}}{g(T_i)}\big]}{\mathbb{E}\big[I_{\{T_i<T_j, T_i<T_{max}\}} I_{\{T_i<D_i\}} \frac{I_{\{T_i<D_i\}}}{g(T_i)} \frac{I_{\{T_i<D_j\}}}{g(T_i)}\big]} \\
&= \frac{\mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j), T_i < T_j, T_i < T_{max}\big]}{\mathbb{P}\big[T_i < T_j, T_i < T_{max}\big]} \\
&= \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j) \big| T_i < T_j, T_i < T_{max}\big].
\end{aligned}$$

This completes the proof. □

4.4 Relationship between Time-dependent Uno's C-index and Time-dependent Concordance

Discussing the relationship between the IPCW (weighted) discrimination measures and the respective non-IPCW (unweighted) discrimination measures is interesting. If the relationship can be explained, we may quantify the difference between those two measures. Unfortunately, the relationship is not simple since it depends on more than one quantity, as discussed in this section.

To derive the mathematical formula of the relationship between C and C^{td} as the IPCW and non-IPCW measures, respectively, we can rewrite C as follows

$$\begin{aligned}
\mathrm{C} &= \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}\big] \\
&= \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}, T_i < D_i \wedge D_j\big] \\
&\qquad \mathbb{P}\big[T_i < D_i \wedge D_j | T_i < T_j, T_i < T_{max}\big] \\
&\quad + \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}, T_i \geq D_i \wedge D_j\big] \\
&\qquad \big(1 - \mathbb{P}\big[T_i < D_i \wedge D_j | T_i < T_j, T_i < T_{max}\big]\big).
\end{aligned} \tag{4.31}$$

Recall also that

$$\mathrm{C}^{td} = \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j)\big|T_i < T_j, T_i < T_{max}, T_i < D_i \wedge D_j\big].$$

Denote

$$\begin{aligned}
\theta &= \mathbb{P}\big[T_i < D_i \wedge D_j | T_i < T_j, T_i < T_{max}\big] \\
&= \frac{\mathbb{P}\big[T_i < D_i \wedge D_j, T_i < T_j, T_i < T_{max}\big]}{\mathbb{P}\big[T_i < T_j, T_i < T_{max}\big]} \\
&= \frac{\mathbb{E}\big[I_{\{T_i < D_i \wedge D_j\}} I_{\{T_i < X_j, T_i < T_{max}\}} I_{\{T_i < D_i\}}\big]}{\mathbb{E}\left[I_{\{T_i < X_j, T_i < T_{max}\}} I_{\{T_i < D_i\}} \frac{I_{\{T_i < D_i\}}}{G(T_i)} \frac{I_{\{T_i < D_j\}}}{G(T_i)}\right]} \\
&= \frac{\mathbb{E}\big[G^2(T_i) I_{\{T_i < X_j, T_i < T_{max}\}} I_{\{T_i < D_i\}}\big]}{\mathbb{E}\left[I_{\{T_i < X_j, T_i < T_{max}\}} I_{\{T_i < D_i\}} \frac{I_{\{T_i < D_i\}}}{G(T_i)} \frac{I_{\{T_i < D_j\}}}{G(T_i)}\right]},
\end{aligned} \tag{4.32}$$

where the last equality because D_i and D_j are assumed to have the same distribution, and to be independent from each other and from T_i, T_j, X_i, X_j. Hence, in the numerator of the last equality we obtain

$$\mathbb{E}\big[I_{\{T_i < D_i \wedge D_j\}}\big] = \mathbb{P}\big[D_i > T_i\big]\mathbb{P}\big[D_j > T_i\big] = G^2(T_i).$$

Therefore, based on the last equality of (4.32), we propose

$$\widehat{\theta}_n = \frac{\sum_{i\neq j}^{n} I_{\{T_i<D_i\}} I_{\{T_i<X_j, T_i<T_{max}\}} \widehat{G}_n^2(T_i)}{\sum_{i\neq j}^{n} I_{\{T_i<D_i\}} I_{\{T_i<X_j, T_i<T_{max}\}} \widehat{G}_n^{-2}(T_i)}.$$

as the estimator of θ. However, we will not discuss $\widehat{\theta}_n$ in more details for the rest of this thesis because it is beyond the thesis' aims.

Note that (4.31) can be rewritten as

$$\mathrm{C} = \theta \mathrm{C}^{td} + (1-\theta)\tilde{\mathrm{C}}, \tag{4.33}$$

where

$$\tilde{\mathrm{C}} = \mathbb{P}\big[S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_j) \big| T_i < T_j, T_i < T_{max}, T_i \geq D_i \wedge D_j\big]. \tag{4.34}$$

As we can see in (4.33), the relationship depends not only on θ but also on $\tilde{\mathrm{C}}$. $\tilde{\mathrm{C}}$ is the concordance probability from some "unusable" pairs of individuals that are ignored when we estimate C and C^{td}. Our proposed estimator $\widehat{\mathrm{C}}_n^w$ does not estimate $\tilde{\mathrm{C}}$ since it only evaluates "usable" pairs following the rule proposed by Harrell's C-index (Harrell *et al.*, 1982). However, the influence of $\tilde{\mathrm{C}}$ to C can be quantified from (4.33), especially the value of θ representing the probability that the survival time of individual i is uncensored given that $T_i < T_j$ and $T_i < T_{max}$.

Based on all possible values of θ, the effects of $\tilde{\mathrm{C}}$ on C can be explained as follows:

i. $\theta = 1$, which means that all individuals are uncensored. This situation is where $\widehat{\mathrm{C}}_n^w$ is the same as $\widehat{\mathrm{C}}_n$. Their values furthermore represent the true values of both discrimination measures. In this case, we predict that $\tilde{\mathrm{C}}$ does not have any contribution to C as $(1-\theta) = 0$.

ii. $\theta \in (0, 1)$. This situation is the most common case in survival analysis since survival data usually contains censored observations. In practice, $\widehat{\mathrm{C}}_n^w$ has different values to $\widehat{\mathrm{C}}_n$, where $\widehat{\mathrm{C}}_n^w$ is closer to their true values. In this case, our prediction shows that the contribution of $\tilde{\mathrm{C}}$ is high when θ is close to zero. Otherwise, its contribution to C is minimal.

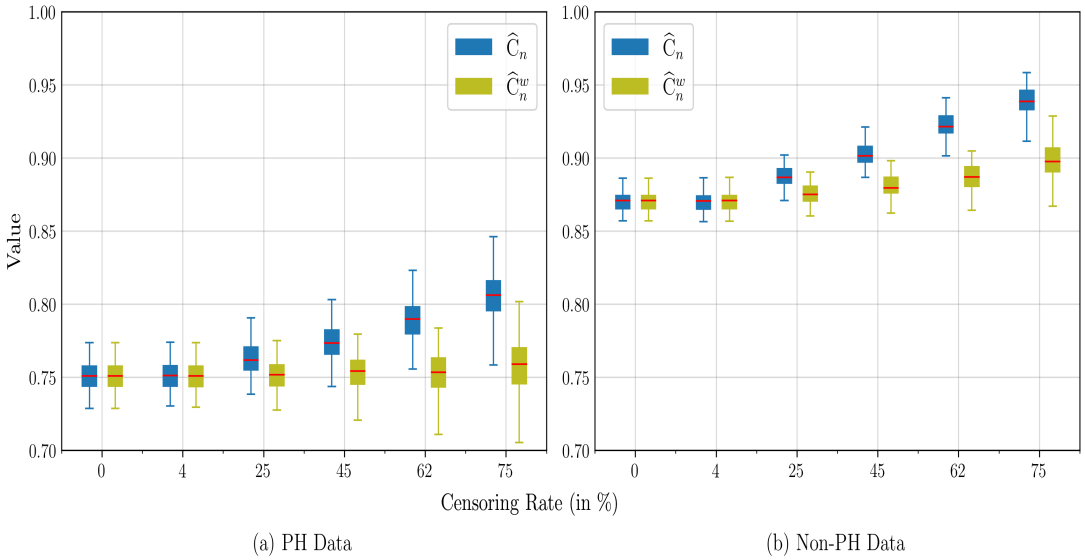
iii. $\theta = 0$. This case is the most extreme in survival analysis because all survival data are censored. Most survival models cannot be fitted to those datasets since they rely on at least one uncensored observation. In this case, C is the same with $\tilde{\mathrm{C}}$, which means that we cannot estimate the value of C in practice. In other words, $\widehat{\mathrm{C}}_n^w$ is undefined.

4.5 Case Studies

In the previous sections, we have shown the theoretical behaviour of time-dependent Uno's C-index, such as the convergence of the proposed estimator and the relationship to time-dependent concordance. In this section, we will implement $\widehat{\mathrm{C}}_n^w$ and $\widehat{\mathrm{C}}_n$ through simulation studies and real-world examples to investigate the practical behaviour of the measures in more detail. We will present two simulation scenarios where we apply the measures to PH and non-PH data. For the real-world examples, we will use the TCGA data discussed in the previous chapter and a new data, namely Heart Failure data.

4.5.1 Simulation 3: PH and Non-PH Data with Various Censoring Rates

The main aim of this simulation is to show how time-dependent Uno's C-index and time-dependent concordance behave when evaluating a fixed model but the censoring rate in the test data varies. To achieve the simulation objective, we employed two types of data, namely PH and non-PH data. The PH data were obtained from Simulation 1 in Section 3.1.1 (Chapter 3) while the non-PH data from Simulation 2 in Section 4.1. We used $\mathscr{D}_{13}$ and $\mathscr{D}_{14}$=$\{0, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, \infty\}$ to discretise the non-PH and PH data, respectively. We fitted the PH data architecture in Table B.5 (Appendix B.2) to a fixed PH train data (n_{train}=1000) with a very small censoring rate (close to 0%) and evaluated the model performance on 100 independent PH test data (n_{test}=1000) with varying censoring rates, i.e. 0%, 4%, 25%, 45%, 62%, and 75%. We applied the same scenario for non-PH data but we fitted the non-PH data architecture in Table B.5 (Appendix B.2).



(a) PH Data (b) Non-PH Data

Figure 4.2: $\widehat{C}_n^w$ and $\widehat{C}_n$ over six different censoring rates in test data in PH data (a) and non-PH data (b). They were estimated based on 100 independent test data sets (n_{test}=1000) from a fixed almost fully uncensored train data (n_{train}=1000).

With regards to censoring induced bias (i.e. the deviation of the measure from its true value), panel (a) and panel (b) in Figure 4.2 show the performance measures evaluated on test data with varying percentages of censoring in the test data for PH and non-PH data, respectively. Note that these refer to one Nnet-survival trained on a sample data with almost no censored observations. As we can see from the figure, the medians and the standard deviations of the boxplots get larger as the censoring rate in the test data increases. In other words, the bias of the measures gets more significant as the censoring increases, where the true value is approximated by the value of the estimator computed from complete data (0% censoring rate). However, generally, it is very clear from the figure that $\widehat{C}_n^w$ is much more stable than $\widehat{C}_n$. For all censoring rates, moreover, $\widehat{C}_n^w$ are much closer to the results for 0% censoring rate than $\widehat{C}_n$. Since the number of ignored pairs increases as censoring rates increase, we need higher weight values to compensate for that loss. Due to the misspecification of $\widehat{G}_n$, the errors of the weights are also significant. As a result, $\widehat{C}_n^w$ is slightly biased and has large standard deviations when the censoring rates are large (e.g. 62% or 75% censoring rates).

4.5.2 Simulation 4: Downward Bias of Time-Dependent Concordance

The aim of this simulation is to demonstrate that the bias due to censoring is not always positive (meaning an improved measure). There are situations where increasing censoring causes the performance indices to deteriorate. A different setting was applied for generating the PH data in Simulation 1. This was particularly with respect to censoring times as well as the relationship between covariates and event times. We generated independent event times based on (3.1) with $\alpha = 0.0005$ and $\gamma = 0.3$. We also assumed that the event times depend on the same covariates and the same coefficient values as in Simulation 1. The follow-up truncation time $T^* = 35$ such that all individuals with $T_i \geq T^*$ were administratively censored at T^*. By changing T^* from 70 (Simulation 1) to 35, we made the data distribution at each period obtained from the discretisation process to be more sensitive to the change of discretisation points. Therefore, we expected that the fitted model would induce downward bias. In particular, we applied $\mathscr{D}_{14}$=$\{0, 4, 7, 9.5, 11.5, 13, 14, 16, 17, 19, 21, 23, 25, 28, 35, \infty\}$ as the discretisation setup. For the censoring distribution, we independently generated 1000 discrete censored times using several predetermined probabilities at each period (p_t) such that the desired censoring rates were achieved. As usually, the observed times were the minimum between survival and censored times.

Next, we conditionally randomised the observed times and their covariates for a fraction of individuals by randomly perturbing covariates of individuals whose observed times are less than or equal z. Meanhile, we preserved the relationships for the generated observed times that are greater than z. To tune the appropriate value for z, we did cross-validation, such that $\widehat{\mathrm{C}}_n$ was downward biased while keeping $\widehat{\mathrm{C}}_n^w$ less biased. This procedure was conducted by "trial and error" because we initially did not know the minimum proportion of individuals whose covariates should be perturbed, so the time-dependent concordance would be downward biased. In other words, we fine-tuned the model that could satisfy this simulation objective. Based on our experiments, we found 10 as the value for z.

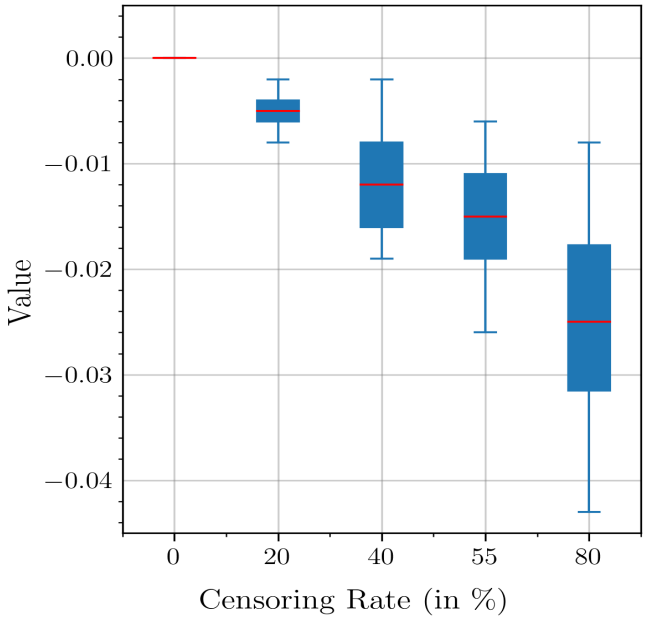


Figure 4.3: Boxplots of $(\widehat{\mathrm{C}}_n - \widehat{\mathrm{C}}_n^w)$ on five different censoring rates. They were estimated from 100 independent test data sets (n_{test}=1000) trained on a fixed almost fully uncensored data (n_{train}=1000).

We fitted Sim.4 Nnet-survival architecture in Table B.5 (Appendix B.2) to a single train data with close to 0% censoring rate. Then, we evaluated the model performance on 100 independent test data with varying censoring rates (i.e. 0%, 20%, 40%, 55%, and 80%) by using $\widehat{\mathrm{C}}_n^w$ and $\widehat{\mathrm{C}}_n$. Figure 4.3 shows $(\widehat{\mathrm{C}}_n - \widehat{\mathrm{C}}_n^w)$ over the five censoring rates. We can see that the difference gets bigger as censoring rate increases and the values are negative. This results show that the downward (negative) bias of time dependence concordance is bigger than time-dependent Uno's C-index.

To sum up, in this section, we have applied $\widehat{\mathrm{C}}_n^w$ and $\widehat{\mathrm{C}}_n$ in HF data, where the results complement the common insight in the literature showing that non-IPCW estimators of discrimination measures are upward biased (see the works by Gerds *et al.* (2013)). On the other hand, from our numerical implementations, $\widehat{\mathrm{C}}_n$ can be downward biased, depending on the fitted models and the data behaviour. In the next section, we will also apply $\widehat{\mathrm{C}}_n^w$ and $\widehat{\mathrm{C}}_n$ to TCGA and investigate their behaviour in the data.

4.6 Real-world Examples: Heart Failure Data and TCGA Data

In this section, we will provide real-world examples of implementing the time-dependent Uno's C-index and time-dependent concordance in two real data: heart failure (HF) data and TCGA data. We choose HF data as an example because it shows that time-dependent concordance can have a downward bias.

Heart failure (HF) data was initially collected by Ahmad *et al.* (2017) and was elaborated in more detail in Chicco & Jurman (2020). The data contains the time to death due to heart failure of 299 patients and their baseline clinical information. The follow-up time is 289 days, where 96 (32%) patients were uncensored, and 203 (68%) patients were censored. The response variable is the time until the death of patients (in days), and eleven available clinical features were used as covariates (i.e. six continuous variables and five categorical variables)

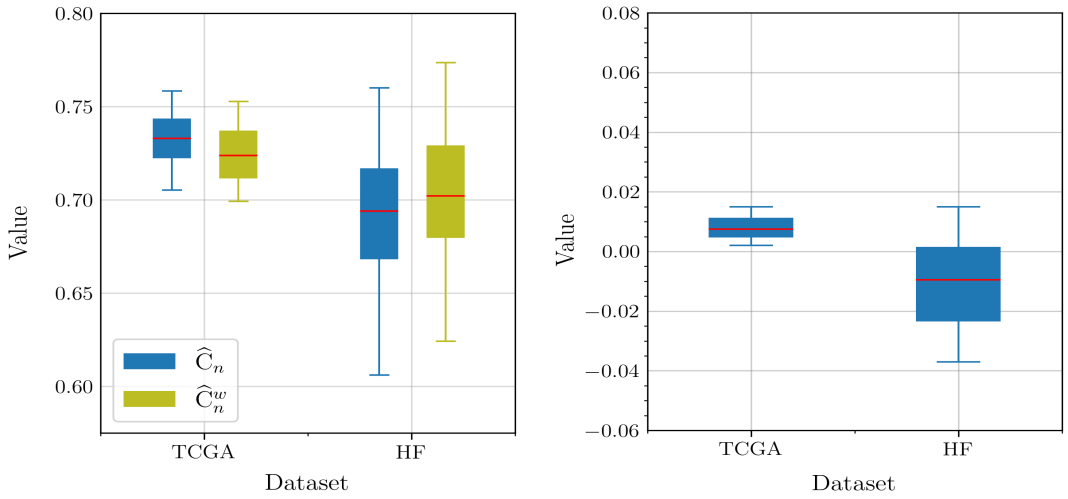
The TCGA data used in this section was obtained from Section 3.1.2 in Chapter 3. To obtain the train and test data from the TCGA and HF data, we randomly divided the TCGA data into 70% train data and 30% test data. We repeated this procedure 100 times to have 100 pairs of the train and test data. We discretised the training and test data in TCGA data using $\mathscr{D}_9$, while in HF data we used $\mathscr{D}_{15}$={ 0, 14.25, 28.5, 42.75, 57, 71.25, 85.5, 99.75, 114, 128.25, 142.5, 156.75, 171, 185.25, 199.5, ∞}. For the TCGA data, the good model architecture in Table B.2 (Appendix B.1) was fitted to each train data of TCGA data. Meanwhile, we fitted the architecture in Table B.5 (Appendix B.2) to each train data of HF data. Then, we evaluated the models' performance on their respective 100 independent test data.

Figure 4.4 shows the results of the numerical implementation of this section. Panel (a) presents the exact values of $\widehat{\mathrm{C}}_n^w$ and $\widehat{\mathrm{C}}_n$ for the two data, where blue and yellow boxplots represent the estimated $\widehat{\mathrm{C}}_n^w$ and $\widehat{\mathrm{C}}_n$ from 100 test data, respectively. As we can see from the panel (a), $\widehat{\mathrm{C}}_n$ is upward (positive) biased in TCGA data. In contrast, $\widehat{\mathrm{C}}_n$ is downward (negative) biased in HF data. To make the presentation in panel (a) clearer, we also draw the boxplots of the differences between time-dependent concordance and time-dependent Uno's C-index ($\widehat{\mathrm{C}}_n - \widehat{\mathrm{C}}_n^w$) for each test data as given in panel (b). We can see from panel (b) that the plots

of $(\widehat{\mathrm{C}}_n - \widehat{\mathrm{C}}_n^w)$ in TCGA data are always positive, which means that $\widehat{\mathrm{C}}_n$ is upward (positive) biased. On the other hand, most plots of $(\widehat{\mathrm{C}}_n - \widehat{\mathrm{C}}_n^w)$ in HF data are negative, which means that $\widehat{\mathrm{C}}_n$ is downward (negative) biased.

In summary, from the results obtained in this section, we have shown that the unweighted discrimination measures, namely time-dependent concordance, can be either upward or downward bias, depending on the fitted models and the data characteristics. Along with Simulation 4 in Section 4.5.2, the results from HF data complement some available works in the literature (Gerds *et al.*, 2013; Gönen & Heller, 2005) regarding the bias direction of the non-IPCW discrimination measure. In particular, we have shown via simulation study and real-world implementation that the non-IPCW discrimination measures can be downward biased.

By the end of this chapter, we have shown through a simulation study that $\widehat{\mathrm{C}}_n^{\mathrm{uno}}$ (and hence $\widehat{\mathrm{C}}_n^{\mathrm{har}}$) were not proper to evaluate the model performance of non-linear models where the PH assumption was violated in the data. These results have motivated us to propose the time-dependent Uno's C-index as the unbiased version of time-dependent concordance to cope with such an issue. In more detail, we have also shown the convergence of time-dependent Uno's C-index to its population probability. In addition, by a simulation study and real-world examples, we have demonstrated the behaviour of the time-dependent Uno's C-index when censoring rate increases in the test data. Although in practice, $\widehat{\mathrm{C}}_n^w$ still has a large standard deviation when the censoring rate is large, its bias is smaller than $\widehat{\mathrm{C}}_n$. In the next chapter, we will introduce a novel measure called pair calibration, which can be seen as a measure of calibration composed of discrimination quantities.



(a) The Exact Values

(b) The Differences

Figure 4.4: Boxplots of $\widehat{C}_n$ and $\widehat{C}_n^w$ (a) and $(\widehat{C}_n - \widehat{C}_n^w)$ (b) for TCGA data and HF data. They were estimated on 100 test data (n_{test}=30% of the data) from models fitted to 100 train data (n_{train}=70% of the data).

Chapter 5

Pair Calibration

In Chapter 3, we have delved into the significance of calibration measures, specifically the integrated Brier score. Moving on to Chapter 4, we introduced the time-dependent Uno's C-index as a discrimination measure, crucial for handling the violation of PH assumption when we fitted non-linear models. Calibration, as we know, focuses on assessing the difference between the model outputs and the outcomes. Conversely, discrimination is a tool that evaluates how effectively a model distinguishes the risks of two different individuals. It is important to note that we can have a model with good calibration but poor discrimination, or vice versa (see Chapter 6). One should have good model discrimination capability when the study objective is to distinguish the risks amongst individuals in two or more groups under study. Meanwhile, one may prefer good calibration capability when the study objective is to minimise the difference between the quantities estimated from the sample and their respective population parameters. These two types of measures are devoted to different tasks. However, we sometimes aim for models that exhibit good calibration and discrimination, although this is not always the case in practice.

In this chapter, we will propose a novel performance measure called pair calibration, which is a measure that calibrates a discrimination quantity. To elaborate, pair calibration is a measure that calibrates the predicted probabilities of the order of two individuals' event times with their outcomes. The concordance between the order of the outputs and the respective outcomes can increase

model discrimination capability, potentially minimising the differences between the predicted probability of the order and the respective outcomes.

We will explain the primary motivation behind pair calibration by using an example case as follows. Consider we have a population containing two groups of patients characterised by risk factors to have a heart attack within the follow-up. The first group has a high risk of having a heart attack. Meanwhile, the second group has a low risk. Hence, the two groups' survival curves should be significantly different. Then, we fit a model to a population sample such that the predicted survival curves of all individuals in the sample are very close to each other. In this situation, the standard discrimination measures, such as Harrell's C-index and time-dependent concordance, may show a good discrimination value because they only identify the concordance between the order of the predicted survival curves and the respective outcomes. On the other hand, the proposed measure in this chapter, i.e. pair calibration, will show a bad model performance because pair calibration is based on how close the probability of the order of the predicted survival curves is to the respective outcomes. In other words, pair calibration is better than the standard discrimination measures in showing the characteristic of the population, i.e. the risks of the two groups are highly different. We can also state that pair calibration is an extension of the standard discrimination measures.

This chapter initiates with the formulation of pair calibration and its estimators. We then proceed to conduct simulation studies and present real-world examples to elaborate on the behaviour and properties of these estimators. This practical insight has led us to propose several measures that can be used as the reference value of pair calibration. Moreover, we introduce the truncated PC, a practical and effective solution for those interested in prediction at sub-intervals of the follow-up time. In the final section of this chapter, we demonstrate the convergence of the estimators.

5.1 Formulation of Pair Calibration

To derive the measure, we first define the following expectation with respect to T_i and T_j, for all pairs $(i \neq j)(i, j = 1, \cdots, n)$

$$\text{PC} = \mathbb{E}\left[\left(I_{\{T_i \leq T_j\}} - \widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i < T_{max}\right]\right)^2 \middle| T_i < T_{max}\right] \tag{5.1}$$

as our proposed measure called pair calibration, where all individuals within T_{max} are administratively censored. The probability of interest in (5.1) is computed by

$$\begin{aligned}
&\widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i < T_{max}\right] \\
&= \frac{\widehat{\mathbb{P}}\left[T_i \leq T_j, T_i < T_{max}\right]}{\widehat{\mathbb{P}}\left[T_i < T_{max}\right]} \\
&= \frac{\sum_{t=1}^{T_{max}-1} \widehat{\mathbb{P}}\left[T_i = t\right] \widehat{\mathbb{P}}[T_j > t-1]}{1 - S(T_{max} - 1; \mathbb{Z}_i)} \\
&= \frac{\sum_{t=1}^{T_{max}-1} \left(S(t-1; \mathbb{Z}_i) - S(t; \mathbb{Z}_i)\right) S(t-1; \mathbb{Z}_j)}{1 - S(T_{max} - 1; \mathbb{Z}_i)},
\end{aligned} \tag{5.2}$$

for $t \in \{1, \cdots, T_{max} - 1\}$. Throughout the rest of this chapter, we denote

$$\pi_{ij}^{T_{max}} = \widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i < T_{max}\right].$$

The possible values of pair calibration are within zero and one, where the closer the pair calibration to zero, the better the model performance. However, pair calibration will only be zero if the probability of interest $\pi_{ij}^{T_{max}}$ is either zero or one, which is the property of the population and never seen in practice. Although pair calibration and Brier score are the expectation of squared error, they completely calibrate different quantities. Brier score calibrates a single predicted survival curve with its respective outcome; pair calibration evaluates the difference between the predicted probability of the order of two individuals' survival times and its respective outcome. Pair calibration is an overall performance measure, while the Brier score assesses the model's predictive performance at a fixed period $t \in \mathcal{T}$.

In the two subsequent sections, we will propose two estimators of pair calibration. The first estimator is directly derived from the formula (5.1), while

the second estimator is derived by slightly modifying a condition in the first estimator. The modification allows us to employ more usable pairs that the first estimator ignored. Thus, the second estimator is computed based on more information than the first. By proposing the two estimators of (5.1), we will show that the difference between them is only due to the weight error.

5.1.1 First Estimator of Pair Calibration

To obtain the first estimator of pair calibration, we first rewrite (5.1) as follows

$$\mathrm{PC} = \frac{\mathbb{E}\left[\left(I_{\{T_i \le T_j\}} - \pi_{ij}^{T_{max}}\right)^2 I_{\{T_i < T_{max}\}}\right]}{\mathbb{E}\left[I_{\{T_i < T_{max}\}}\right]}, \tag{5.3}$$

where the expectation is with respect to T_i and T_j. Then, the right-hand side of (5.3) can be transformed as follows

$$\frac{\mathbb{E}\left[\left(I_{\{T_i \le T_j\}} - \pi_{ij}^{T_{max}}\right)^2 \frac{I_{\{T_i < D_i\}}}{G(T_i)} \frac{I_{\{T_j < D_j\}}}{G(T_j)} I_{\{T_i < T_{max}\}}\right]}{\mathbb{E}\left[I_{\{T_i < T_{max}\}} \frac{I_{\{T_i < D_i\}}}{G(T_i)}\right]} \tag{5.4}$$

where the expectations of $I_{\{T_i < D_i\}}/G(T_i)$ and $I_{\{T_j < D_j\}}/G(T_j)$ follow (4.3). Thus, based on (5.4), we propose

$$\widehat{\mathrm{PC}}_n^1 = \frac{\sum_{i \ne j}^n \left[\left(I_{\{T_i \le T_j\}} - \pi_{ij}^{T_{max}}\right)^2 I_{\{T_i \le T_j, T_i < T_{max}\}} \frac{I_{\{T_i < D_i\}}}{\widehat{G}_n(T_i)} \frac{I_{\{T_j < D_j\}}}{\widehat{G}_n(T_j)}\right]}{\sum_{i=1}^n \left[I_{\{T_i < T_{max}\}} \frac{I_{\{T_i < D_i\}}}{\widehat{G}_n(T_i)}\right]} \tag{5.5}$$

as the first estimator of (5.1). The convergence of $\widehat{\mathrm{PC}}_n^1$ to PC is stated in the following theorem, where the proof will be discussed in Section 5.7.

Theorem 5.1.1 (Convergence of The First Estimator of Pair Calibration)**.** *Define a family of functions as follows*

$$\mathcal{G}_\epsilon = \{g_n : \{1, \cdots, T_{max} - 1\} \to [\epsilon, 1]\},$$

where g_n is any estimator of G, and its value is depending on sample size n. Assume that $g_n \xrightarrow{a.s.} G$ as $n \to \infty$. Suppose the regularity conditions (R1-R4) in Section 2.1.3 hold. Then,

$$\widehat{PC}_n^1 \xrightarrow{a.s.} PC, \tag{5.6}$$

as $n \to \infty$.

5.1.2 Second Estimator of Pair Calibration

To propose the second estimator of pair calibration, we first decompose the right-hand side of (5.1) into two terms as follows

$$\mathbb{E}\left[\left(1-\pi_{ij}^{T_{max}}\right)^2 I_{\{T_i\leq T_j\}}\middle|T_i<T_{max}\right]+\mathbb{E}\left[\left(0-\pi_{ij}^{T_{max}}\right)^2 I_{\{T_i>T_j\}}\middle|T_i<T_{max}\right], \quad (5.7)$$

where the first and second terms are based on $I_{\{T_i\leq T_j\}}$ and $I_{\{T_i>T_j\}}$ cases, respectively. Then, (5.7) is rewritten as follows

$$\begin{aligned}&\mathbb{E}\left[\left(1-\pi_{ij}^{T_{max}}\right)^2 I_{\{T_i\leq T_j\}}\frac{I_{\{T_i<D_i\}}}{G(T_i)}\frac{I_{\{T_j<D_j\}}}{G(T_j)}I_{\{T_i<T_{max}\}}\middle|T_i<T_{max}\right]\\&\quad+\mathbb{E}\left[\left(0-\pi_{ij}^{T_{max}}\right)^2 I_{\{T_j<T_i\}}\frac{I_{\{T_j<D_j\}}}{G(T_j)}\frac{I_{\{T_i<D_i\}}}{G(T_i)}I_{\{T_i<T_{max}\}}\middle|T_i<T_{max}\right].\end{aligned} \quad (5.8)$$

We may include more information from the potential ignored pairs (i,j) in (5.8) by replacing $I_{\{T_j<D_j\}}/G(T_j)$ with $I_{\{T_i\leq D_j\}}/G(T_i^-)$ in the first term of (5.8) so that we obtain

$$\begin{aligned}&\mathbb{E}\left[\left(1-\pi_{ij}^{T_{max}}\right)^2 I_{\{T_i\leq T_j\}}\frac{I_{\{T_i<D_i\}}}{G(T_i)}\frac{I_{\{T_i\leq D_j\}}}{G(T_i^-)}I_{\{T_i<T_{max}\}}\middle|T_i<T_{max}\right]\\&\quad+\mathbb{E}\left[\left(0-\pi_{ij}^{T_{max}}\right)^2 I_{\{T_j<T_i\}}\frac{I_{\{T_j<D_j\}}}{G(T_j)}\frac{I_{\{T_i<D_i\}}}{G(T_i)}I_{\{T_i<T_{max}\}}\middle|T_i<T_{max}\right],\end{aligned} \quad (5.9)$$

where $G(T_i^-)$ means the left-hand limit, and for discretely recorded time it is $G(T_i^-)=G(T_i-1)$. The replacement is valid because the expected value of $\left(I_{\{T_i\leq D_j\}}/G(T_i-1)\middle|T_i,\mathbb{Z}_i,T_j,\mathbb{Z}_j,D_i\right)$, where D_j is independent of $T_i,\mathbb{Z}_i,T_j,\mathbb{Z}_j$,

and D_i, satisfies

$$\begin{aligned}
&\mathbb{E}\left[\frac{I_{\{T_i\leq D_j\}}}{G(T_i-1)}\middle| T_i,\mathbb{Z}_i,T_j,\mathbb{Z}_j,D_i\right]\\
&=\frac{1}{G(T_i-1)}\mathbb{E}\left[I_{\{T_i<D_j\}}\cup I_{\{T_i=D_j\}}\middle|T_i,\mathbb{Z}_i,T_j,\mathbb{Z}_j,D_i\right]\\
&=\frac{1}{G(T_i-1)}\mathbb{E}\left[I_{\{T_i<D_j\}}\cup I_{\{T_i=D_j\}}\middle|T_i\right]\\
&=\frac{1}{G(T_i-1)}\left(\mathbb{E}\left[I_{\{T_i<D_j\}}\middle|T_i\right]+\mathbb{E}\left[I_{\{T_i=D_j\}}\middle|T_i\right]\right)\\
&=\frac{1}{G(T_i-1)}\left(\mathbb{P}[D_j>T_i]+\mathbb{P}[D_j=T_i]\right)\\
&=\frac{1}{G(T_i-1)}\mathbb{P}[D_j>T_i-1]\\
&=\frac{1}{G(T_i-1)}G(T_i-1)\\
&=1.
\end{aligned}\tag{5.10}$$

The second equality of (5.10) because $I_{\{T_i<D_j\}}$ and $I_{\{T_i=D_j\}}$ are disjoint events, and the fifth equality because in discrete time the probability of tied event times may not equal zero. However, we cannot apply the indicator functions $I_{\{T_i<D_i\}}I_{\{T_i\leq D_j\}}$ to the second term of (5.8) because it will not allow us to evaluate (i,j) when $(T_i>T_j)$.

Based on (5.9), we finally propose the second estimator of pair calibration as follows

$$\begin{aligned}
\widehat{\mathrm{PC}}_n^2=&\left[\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_i<D_i\}}\widehat{G}_n^{-1}(T_i)\right]\right]^{-1}\\
&\sum_{i\neq j}^{n}\left[\left(1-\pi_{ij}^{T_{max}}\right)^2 I_{\{T_i\leq T_j,T_i<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_i\leq D_j\}}}{\widehat{G}_n(T_i-1)}\right.\\
&\left.+\left(0-\pi_{ij}^{T_{max}}\right)^2 I_{\{T_j<T_i,T_i<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_j<D_j\}}}{\widehat{G}_n(T_j)}\right],
\end{aligned}\tag{5.11}$$

where $\widehat{G}_n$ is an estimator of G, and it is dependent on the sample size n. The convergence of $\widehat{\mathrm{PC}}_n^2$ to PC is stated in the following theorem, where the proof will be discussed in Section 5.7.

Theorem 5.1.2 (Convergence of The Second Pair Calibration Estimator)**.** *Define a family of functions as follows*

$$\mathcal{G}_\epsilon = \{g_n : \{1, \cdots, T_{max} - 1\} \to [\epsilon, 1]\},$$

where g_n is any estimator of G, and its value is depending on sample size n. Assume that $g_n \xrightarrow{a.s.} G$ as $n \to \infty$. Suppose the regularity conditions (R1-R4) in Section 2.1.3 hold. Then,

$$\widehat{PC}_n^2 \xrightarrow{a.s.} PC, \tag{5.12}$$

as $n \to \infty$.

5.2 Truncated Pair Calibration

Pair calibration (5.1) evaluates model performance over the observation periods $\{1, \cdots, T_{max} - 1\}$. This section introduces the more general pair calibration evaluating model performance over the sub-intervals of the follow-up time, namely $\{\tau_1, \cdots, \tau_2 - 1\}$ for $1 \leq \tau_1 < \tau_2 \leq T_{max}$, which is called as truncated pair calibration. An example of the similar truncation applied to a different performance measure was discussed by Song *et al.* (2012). They proposed the truncated version of Uno's C-index that can be used to evaluate the model performance within $[a, b] \subset [0, T_{max}]$.

The truncated pair calibration over $\{\tau_1, \cdots, \tau_2 - 1\}$ is defined as follows

$$\begin{aligned} &\mathrm{PC}^{\tau_1, \tau_2} \\ &= \mathbb{E}\left[\left(I_{\{T_i \leq T_j\}} - \widehat{\mathbb{P}}\left[T_i \leq T_j | T_i, T_j \in [\tau_1, \tau_2)\right]\right)^2 \middle| T_i, T_j \in [\tau_1, \tau_2)\right], \end{aligned} \tag{5.13}$$

where

$$
\begin{aligned}
&\widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i, T_j \in [\tau_1, \tau_2)\right] \\
&= \frac{\widehat{\mathbb{P}}\left[T_i \leq T_j, \tau_1 \leq T_i < \tau_2, \tau_1 \leq T_j < \tau_2\right]}{\widehat{\mathbb{P}}\left[\tau_1 \leq T_i < \tau_2, \tau_1 \leq T_j < \tau_2\right]} \\
&= \frac{1}{\widehat{\mathbb{P}}\left[\tau_1 \leq T_i < \tau_2, \tau_1 \leq T_j < \tau_2\right]} \\
&\quad \sum_{t=\tau_1}^{\tau_2-1} \widehat{\mathbb{P}}\left[T_i = t\right] \, \widehat{\mathbb{P}}\left[T_j \geq t, t < \tau_2, T_j < \tau_2\right] \\
&= \frac{1}{\widehat{\mathbb{P}}\left[\tau_1 \leq T_i < \tau_2, \tau_1 \leq T_j < \tau_2\right]} \\
&\quad \sum_{t=\tau_1}^{\tau_2-1} \widehat{\mathbb{P}}\left[T_i = t\right] \, \widehat{\mathbb{P}}\left[t \leq T_j < \tau_2\right] \\
&= \frac{1}{\widehat{\mathbb{P}}\left[\tau_1 \leq T_i < \tau_2\right] \widehat{\mathbb{P}}\left[\tau_1 \leq T_j < \tau_2\right]} \\
&\quad \sum_{t=\tau_1}^{\tau_2-1} \left(S(t-1;\mathbb{Z}_i) - S(t;\mathbb{Z}_i)\right) \left(S(t-1;\mathbb{Z}_j) - S(\tau_2-1;\mathbb{Z}_j)\right) \\
&= \frac{1}{\left(\widehat{\mathbb{P}}\left[T_i \geq \tau_1\right] - \widehat{\mathbb{P}}\left[T_i \geq \tau_2\right]\right) \left(\widehat{\mathbb{P}}\left[T_j \geq \tau_1 - \widehat{\mathbb{P}}\left[T_j \geq \tau_2\right]\right]\right)} \\
&\quad \sum_{t=\tau_1}^{\tau_2-1} \left(S(t-1;\mathbb{Z}_i) - S(t;\mathbb{Z}_i)\right) \left(S(t-1;\mathbb{Z}_j) - S(\tau_2-1;\mathbb{Z}_j)\right) \\
&= \frac{1}{\left(S(\tau_1-1;\mathbb{Z}_i)) - S(\tau_2-1;\mathbb{Z}_i)\right) \left(S(\tau_1-1;\mathbb{Z}_j) - S(\tau_2-1;\mathbb{Z}_j)\right)} \\
&\quad \sum_{t=\tau_1}^{\tau_2-1} \left(S(t-1;\mathbb{Z}_i) - S(t;\mathbb{Z}_i)\right) \left(S(t-1;\mathbb{Z}_j) - S(\tau_2-1;\mathbb{Z}_j)\right).
\end{aligned}
$$

We furthermore transform (5.13) as follows

$$\begin{aligned}
&\mathbb{E}\left[\left(I_{\{T_i\leq T_j\}}-\widehat{\mathbb{P}}\left[T_i\leq T_j\middle|T_i,T_j\in[\tau_1,\tau_2)\right]\right)^2\middle|T_i,T_j\in[\tau_1,\tau_2)\right]\\
&=\frac{1}{\mathbb{E}\left[I_{\{T_i<D_i\}}G^{-1}(T_i)I_{\{T_j<D_j\}}G^{-1}(T_j)I_{\{\tau_1\leq T_i<\tau_2\}}I_{\{\tau_1\leq T_j<\tau_2\}}\right]}\\
&\quad\mathbb{E}\Bigg[\left(I_{\{T_i\leq T_j\}}-\widehat{\mathbb{P}}\left[T_i\leq T_j\middle|T_i,T_j\in[\tau_1,\tau_2)\right]\right)^2\\
&\quad\frac{I_{\{T_i<D_i\}}}{G(T_i)}\frac{I_{\{T_j<D_j\}}}{G(T_j)}I_{\{\tau_1\leq T_i<\tau_2\}}I_{\{\tau_1\leq T_j<\tau_2\}}\Bigg].
\end{aligned}$$

Then, based on the right-hand side of this equality, we propose the estimator of (5.13) as follows

$$\begin{aligned}
&\widehat{\mathrm{PC}}_n^{\tau_1,\tau_2}\\
&=\left[\sum_{i\neq j}^{n}\left[I_{\{T_i<D_i\}}\widehat{G}_n^{-1}(T_i)I_{\{T_j<D_j\}}\widehat{G}_n^{-1}(T_j)I_{\{\tau_1\leq T_i<\tau_2\}}I_{\{\tau_1\leq T_j<\tau_2\}}\right]\right]^{-1}\\
&\quad\sum_{i\neq j}^{n}\Bigg[\left(I_{\{T_i\leq T_j\}}-\widehat{\mathbb{P}}\left[T_i\leq T_j\middle|T_i,T_j\in[\tau_1,\tau_2)\right]\right)^2\\
&\quad\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_j<D_j\}}}{\widehat{G}_n(T_j)}I_{\{\tau_1\leq T_i<\tau_2\}}I_{\{\tau_1\leq T_j<\tau_2\}}\Bigg].
\end{aligned}\tag{5.14}$$

By employing $\mathrm{PC}^{\tau_1,\tau_2}$, we may break down the follow-up time into several mutually exclusive groups of periods and report each group's model performance, respectively. This type of pair calibration might be more beneficial in medical studies. For instance, a doctor may be more interested in cancer patients' survival curves from the third to the sixth month after surgery. The convergence of $\widehat{\mathrm{PC}}_n^{\tau_1,\tau_2}$ is stated in the following theorem, where the proof will be discussed in Section 5.7.

Theorem 5.2.1 (Convergence of The Truncated Pair Calibration Estimator)**.** *Define a family of functions as follows*

$$\mathcal{G}_\epsilon=\{g_n:\{1,\cdots,T_{max}-1\}\to[\epsilon,1]\},$$

where g_n is any estimator of G, and its value is depending on sample size n. Assume that $g_n\overset{a.e.}{\to}G$ as $n\to\infty$. Suppose the regularity conditions (R1-R4) in Section 2.1.3 hold. Then,

$$\widehat{PC}_n^{\tau_1,\tau_2}\overset{a.s.}{\to}PC^{\tau_1,\tau_2}.\tag{5.15}$$

We are also interested in a special case of $\mathrm{PC}^{\tau_1,\tau_2}$, where $\tau_1 = 1$ and $\tau_2 = T_{max}$. The follow-up periods of this pair calibration are the same to PC, namely $\{1, \cdots, T_{max}-1\}$, but they only evaluate the event times that are less than T_{max}. The advantage is that its estimator will not affected by the quality of weights $\widehat{G}_n$ as in $\widehat{\mathrm{PC}}_n^1$ or $\widehat{\mathrm{PC}}_n^2$. However, suppose the number of censoring in the data is massive. In that case, the truncated pair calibration will only use a smaller sample size and underestimate the effect of the ignored pairs. In particular, this type of truncated pair calibration can be defined as follows

$$\begin{aligned}
&\mathrm{PC}^{1,T_{max}} \\
&= \mathbb{E}\left[\left(I_{\{T_i \le T_j\}} - \widehat{\mathbb{P}}\left[T_i \le T_j \middle| T_i \vee T_j < T_{max}\right]\right)^2 \middle| T_i \vee T_j < T_{max}\right],
\end{aligned} \tag{5.16}$$

where $T_i \vee T_j = \max(T_i, T_j)$, and

$$\begin{aligned}
&\widehat{\mathbb{P}}\left[T_i \le T_j \middle| T_i \vee T_j < T_{max}\right] \\
&= \frac{1}{\widehat{\mathbb{P}}\left[T_i < T_{max}, T_j < T_{max}\right]} \widehat{\mathbb{P}}\left[T_i \le T_j, T_i < T_{max}, T_j < T_{max}\right] \\
&= \frac{1}{\widehat{\mathbb{P}}\left[T_i < T_{max}, T_j < T_{max}\right]} \\
&\quad \sum_{t=1}^{T_{max}-1} \widehat{\mathbb{P}}\left[T_i = t\right] \widehat{\mathbb{P}}\left[T_j \ge t, t < T_{max}, T_j < T_{max}\right] \\
&= \frac{1}{\widehat{\mathbb{P}}\left[T_i < T_{max}, T_j < T_{max}\right]} \sum_{t=1}^{T_{max}-1} \widehat{\mathbb{P}}\left[T_i = t\right] \widehat{\mathbb{P}}\left[t \le T_j < T_{max}\right] \\
&= \frac{1}{(1 - S(T_{max}-1; \mathbb{Z}_i))(1 - S(T_{max}-1; \mathbb{Z}_j))} \\
&\quad \sum_{t=1}^{T_{max}-1} \left(S(t-1; \mathbb{Z}_i) - S(t; \mathbb{Z}_i)\right)\left(S(t-1; \mathbb{Z}_j) - S(T_{max}-1; \mathbb{Z}_j)\right),
\end{aligned} \tag{5.17}$$

where the second equality due to independent events. To obtain the estimator of (5.16), we first transform it as follows

$$\begin{aligned}
&\mathbb{E}\left[\left(I_{\{T_i \le T_j\}} - \widehat{\mathbb{P}}\left[T_i \le T_j \middle| T_i \vee T_j < T_{max}\right]\right)^2 \middle| T_i \vee T_j < T_{max}\right] \\
&= \frac{1}{\mathbb{E}\left[I_{\{T_i < D_i\}} G^{-1}(T_i) I_{\{T_j < D_j\}} G^{-1}(T_j) I_{\{T_i \vee T_j < T_{max}\}}\right]} \\
&\quad \mathbb{E}\left[\left(I_{\{T_i \le T_j\}} - \widehat{\mathbb{P}}\left[T_i \le T_j \middle| T_i \vee T_j < T_{max}\right]\right)^2 \frac{I_{\{T_i < D_i\}}}{G(T_i)} \frac{I_{\{T_j < D_j\}}}{G(T_j)} I_{\{T_i \vee T_j < T_{max}\}}\right].
\end{aligned}$$

Then, based on the right-hand side of this equality, the estimator of (5.16) is defined as follows

$$\widehat{\text{PC}}_n^{1,T_{max}} = \frac{1}{\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_j<T_{max}\}}I_{\{T_i<D_i\}}\widehat{G}_n^{-1}(T_i)I_{\{T_j<D_j\}}\widehat{G}_n^{-1}(T_j)\right]} \sum_{i\neq j}^{n}\left[\left(I_{\{T_i\leq T_j\}} - \widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i \vee T_j < T_{max}\right]\right)^2 \frac{I_{\{T_i<D_i\}}I_{\{T_j<D_j\}}}{\widehat{G}_n(T_i)\widehat{G}_n(T_j)} I_{\{T_i<T_{max}\}}I_{\{T_j<T_{max}\}}\right]. \tag{5.18}$$

Remark 5.2.1. When we compute the probabilities of interests, such as $\pi_{ij}^{T_{max}}$, $\widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i, T_j \in [\tau_1, \tau_2)\right]$, and $\widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i \vee T_j < T_{max}\right]$, we need to make sure that their denominators are not equal to zero to prevent the probabilities from undefined values, even though it may only occur in a few numbers of pairs ($i \neq j$). To cope with such potential problem, throughout this thesis, if the predicted survival curves in the denominators equal one, we will subtract them by a very small number, i.e. $1E-10$. This approach minimises the underestimation issue and still preserves the monotonic behaviour of the predicted survival curve.

5.3 Case Studies

Three case studies will be conducted in this section. The first case is based on the PH data generated in Simulation 1 (Section 3.1). In addition to the PH data, the other case studies also employ the non-PH data obtained from Simulation 2 (Section 4.1). We also apply pair calibration to assess the predictive performance of PH and non-PH data that had been generated in previous chapters.

5.3.1 Simulation 1: PH Data

This simulation aims to see how $\widehat{\text{PC}}_n^{1}$, $\widehat{\text{PC}}_n^{2}$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ behave in evaluating the model performance of Nnet-survival fitted to the generated PH data. These estimators evaluate the same length of follow-up, namely $\{1, \cdots, T_{max}-1\}$, although $\widehat{\text{PC}}_n^{1,T_{max}}$ evaluates a different probability of interest. The main objectives of this simulation are the same as Simulation 1 in Section 3.1.1 (Chapter 3) depending

on its scenarios. In the first scenario, $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ will be applied to assess the predictive performance of a fixed model on the test data discretised by three discretisation setups. In the second scenario, we apply those estimators in three different models.

Scenario 1: Performance Evaluation of A Fixed Model

To achieve the goal of this first scenario, we fitted the good and overfitted Nnet-survival architectures in Table B.1 (Appendix B). Moreover, we fitted the architectures to a single fixed train data ($n_{\text{train}} = 1000$) discretised by $\mathscr{D}_1$. Then, the models' predictive performance was evaluated on 100 independent test data ($n_{\text{test}} = 1000$), where $\mathscr{D}_1$, $\mathscr{D}_2$, and $\mathscr{D}_3$ were applied to discretise the test data.

The main results of the good and overfitted models are given in Figure 5.1 and Figure 5.2, respectively. As expected, the values of all estimators in the overfitted models are much higher than in the good models. These results can be explained by the boxplots of their predicted probabilities of interest as presented in Figure 5.3 and Figure 5.4. In general, the figures show that the standard deviations of the probabilities in the good models (Figure 5.3) are smaller than in the overfitted models (Figure 5.4). Smaller standard deviations indicate that the predictions from good models are more consistent as opposed to the overfitted models, whose outputs are more spread out. Moreover, the good models' medians of outputs (the horizontal line within the boxplots) are much closer to the outcomes than the overfitted models. For instance, for all discretisation setups, the means of the outputs of $\widehat{\text{PC}}_n^1$ in the good models (panels (a), (d), and (g) in Figure 5.3) are very close to their outcomes, especially when the outcomes are $I_{\{T_i \leq T_j\}} = 1$. On the contrary, in the overfitted models (panels (a), (d), and (g) in Figure 5.4), we have very different results.

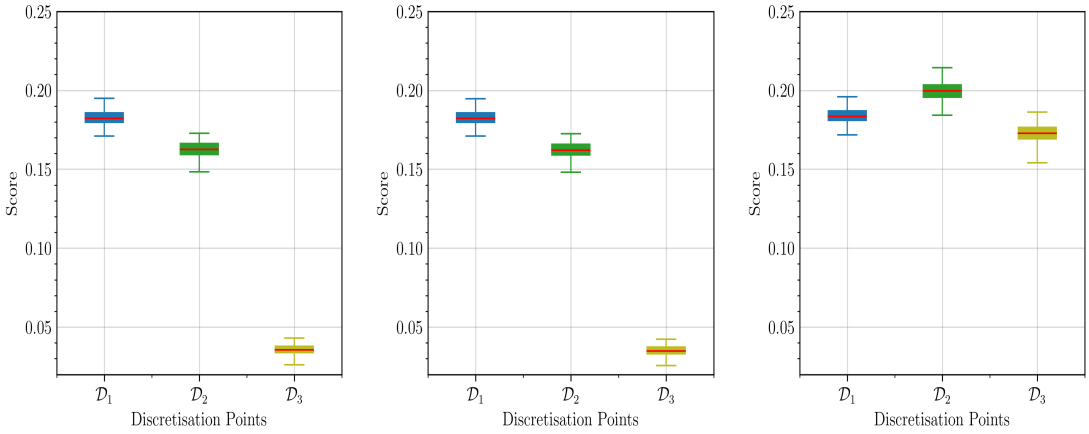
We can see from Figure 5.1 and Figure 5.2 that $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ oscillate as we changed the discretisation setups in the test data. The change of discretisation setup from $\mathscr{D}_1$ to $\mathscr{D}_2$ and $\mathscr{D}_3$ have increased the censoring rates of the test data, where the additional censored individuals are within the truncation period (i.e. the respective T_{max} in $\mathscr{D}_2$ and $\mathscr{D}_3$). In other words, the additional censored individuals are due to administrative censoring so that the given weights in $\mathscr{D}_2$

and $\mathscr{D}_3$ are still the same as $\mathscr{D}_1$. On the other hand, the changes of discretisation setup in the test data mainly affect $\pi_{ij}^{T_{max}}$ as shown in panel (b) and panel (h) in Figure 5.3, where from panel (h) the outputs for $I_{\{T_i \leq T_j\}} = 1$ are much closer than panel (b). Since $\pi_{ij}^{T_{max}}$ are computed based on smaller T_{max} in $\mathscr{D}_2$ and $\mathscr{D}_3$, the denominator of $\pi_{ij}^{T_{max}}$ will be smaller resulting in higher values of $\pi_{ij}^{T_{max}}$. As a result, for the outcomes $I_{\{T_i \leq T_j\}} = 1$, the contributions of the pairs will be smaller and decrease $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$.

The numerical experiment results also show that $\widehat{\text{PC}}_n^{1,T_{max}}$ is not affected too much by the change of the (administrative) censoring rates in the test data (see panel (c) of Figure 5.1 and in panel (c) of Figure 5.2). As we can see in panels (c),(f), and (g) of Figure 5.3 and Figure 5.4), $\widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i \vee T_j < T_{max}\right]$ slightly increase from $\mathscr{D}_1$ to $\mathscr{D}_2$ and $\mathscr{D}_3$. Although we have used smaller T_{max} in $\mathscr{D}_2$ and $\mathscr{D}_3$, the number of usable pairs in $\widehat{\text{PC}}_n^{1,T_{max}}$ is much smaller than in $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ so that some higher probabilities from the ignored pairs are not included in $\widehat{\text{PC}}_n^{1,T_{max}}$.

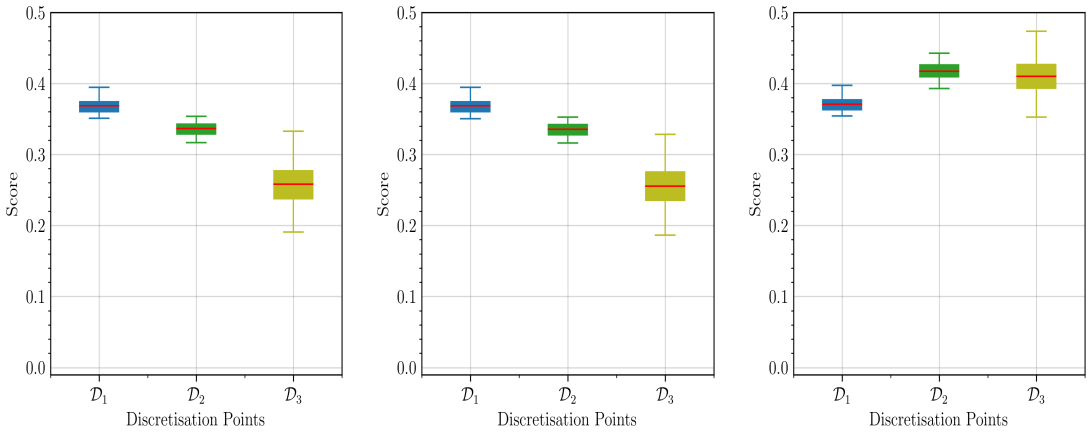
Figure 5.1 and Figure 5.2 also demonstrate that for the initial discretisation setup $\mathscr{D}_1$, the values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ are almost the same because the censoring rate in the test data is close to 0% so that the probability of interest of $\widehat{\text{PC}}_n^{1,T_{max}}$ in (5.17) is almost identical to (5.2) as shown by the upper panels of Figure 5.3 and Figure 5.4. However, when the discretisation setup in the test data change to $\mathscr{D}_2$ and $\mathscr{D}_3$, the censoring rates of the test data also increase resulting in more significant difference between (5.17) and (5.2). Hence, $\widehat{\text{PC}}_n^1$and $\widehat{\text{PC}}_n^2$ significantly differ to $\widehat{\text{PC}}_n^{1,T_{max}}$ when we use $\mathscr{D}_2$ and $\mathscr{D}_3$.

To sum up, the model's predictive performance using $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ may oscillate depending on the discretisation setup used in the test data. The discretisation setups have increased the censoring rates in the test data, affecting the values of the predicted probabilities of interest in pair calibration and the number of usable pairs. However, in this simulation, $\widehat{\text{PC}}_n^{1,T_{max}}$ is more stable as we change the discretisation setup in the test data because it ignores lots of usable pairs $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ that possibly have smaller contributions to the measures. In the next section, we will move on to Scenario 2 of Simulation 1, where we apply $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ to assess the prediction performance of different models.



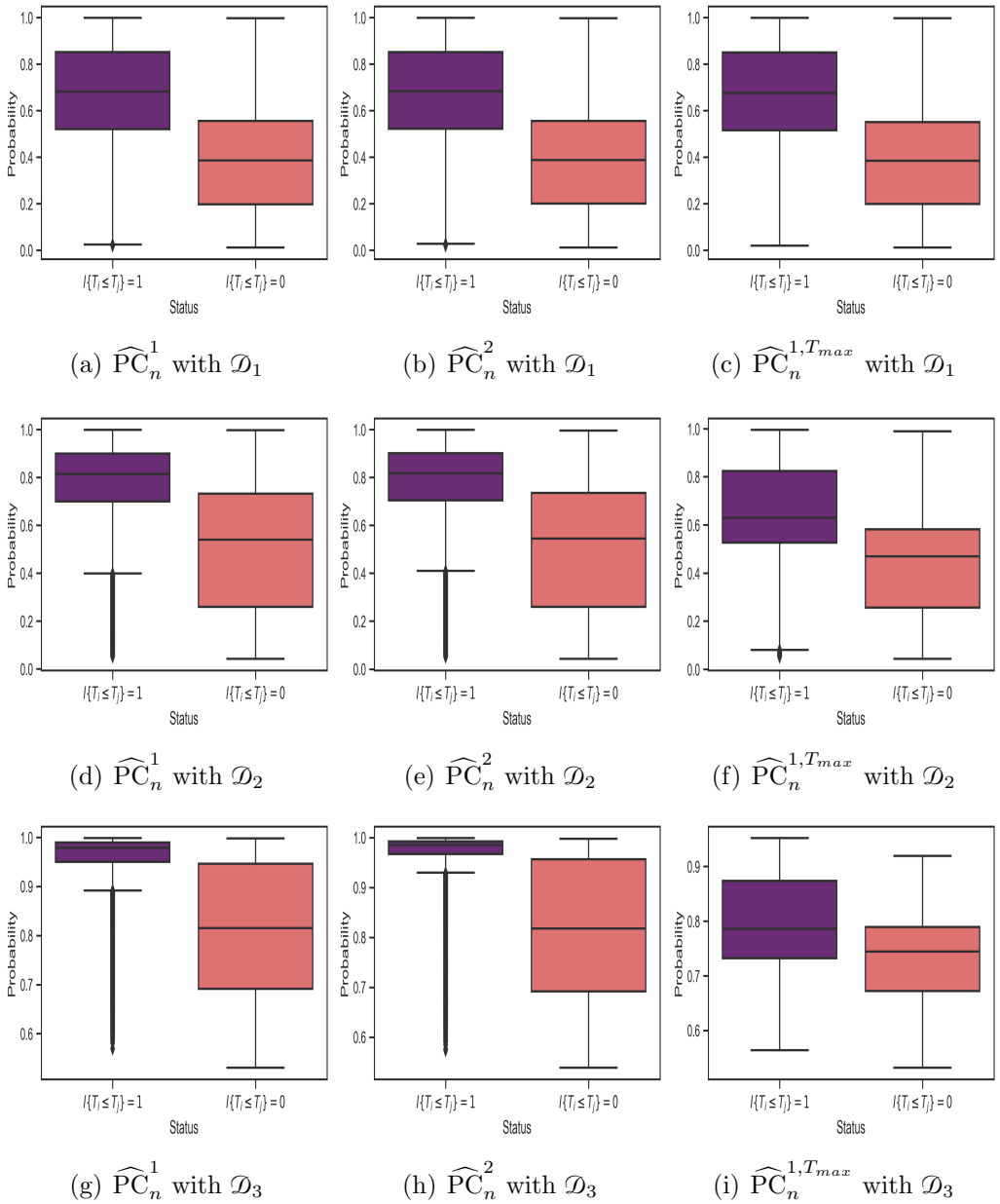
(a) $\widehat{\text{PC}}_n^1$ (b) $\widehat{\text{PC}}_n^2$ (c) $\widehat{\text{PC}}_n^{1,T_{max}}$

Figure 5.1: Boxplots of (a) $\widehat{\text{PC}}_n^1$, (b) $\widehat{\text{PC}}_n^2$, and (c) $\widehat{\text{PC}}_n^{1,T_{max}}$ from Scenario 1 Simulation 1 over three sets of discretisation points in test data. They were estimated on 100 independent test data (n_{test}=1000) from the good Nnet-survival architecture fitted to a train data (n_{train}=1000) discretised by $\mathscr{D}_1$.



(a) $\widehat{\text{PC}}_n^1$ (b) $\widehat{\text{PC}}_n^2$ (c) $\widehat{\text{PC}}_n^{1,T_{max}}$

Figure 5.2: Boxplots of (a) $\widehat{\text{PC}}_n^1$, (b) $\widehat{\text{PC}}_n^2$, and (c) $\widehat{\text{PC}}_n^{1,T_{max}}$ from Scenario 1 Simulation 1 over three sets of discretisation points in test data. They were estimated on 100 independent test data (n_{test}=1000) from the overfitted Nnet-survival architecture fitted to a train data (n_{train}=1000) discretised by $\mathscr{D}_1$.

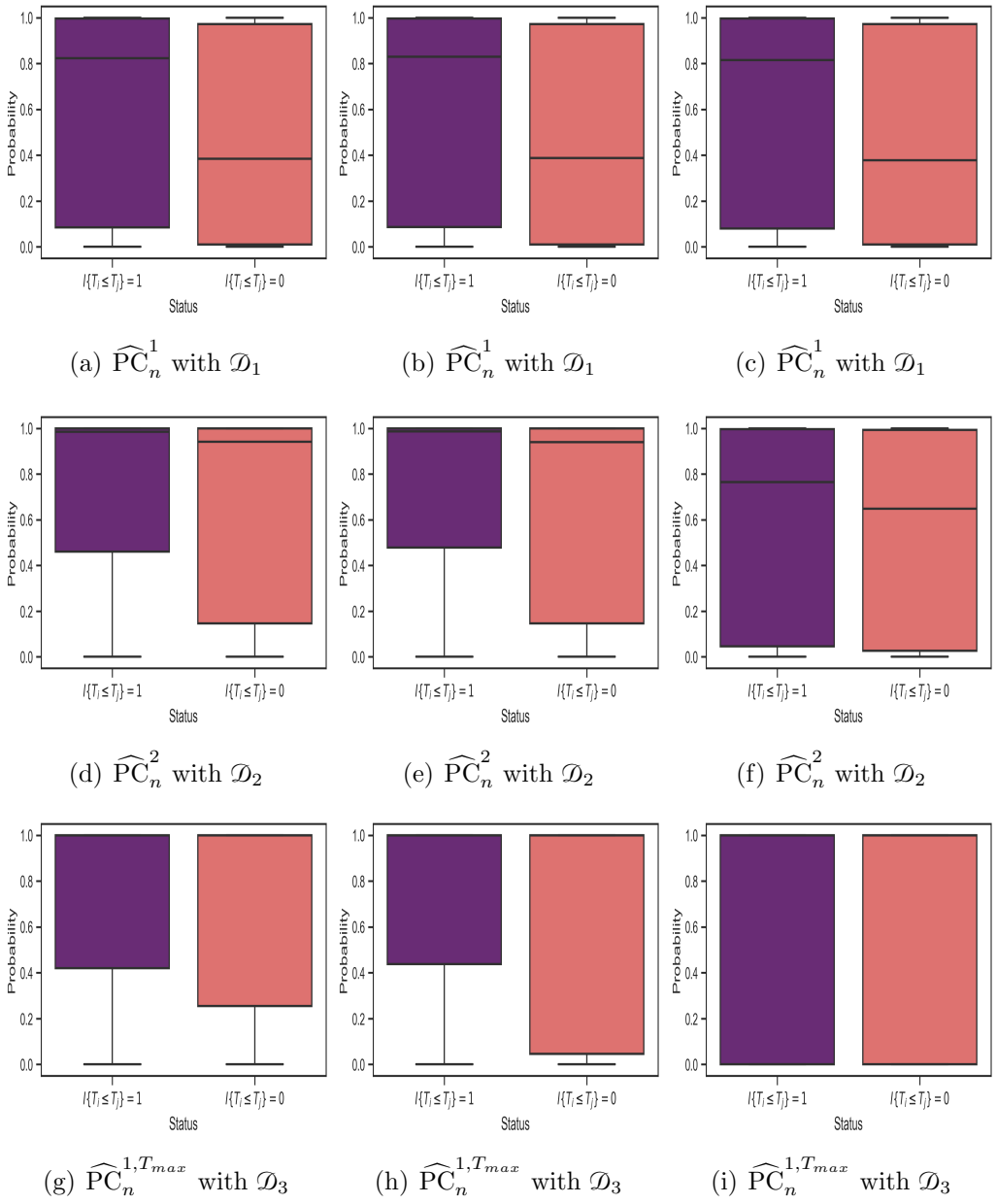


(a) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_1$ (b) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_1$ (c) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_1$

(d) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_2$ (e) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_2$ (f) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_2$

(g) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_3$ (h) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_3$ (i) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_3$

Figure 5.3: The predicted probabilities of interest of $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ from Scenario 1 of Simulation 1 for 20% randomly selected pairs $i \neq j$ grouped by $I_{\{T_i \leq T_j\}}$ over each discretisation setup. They were obtained from the good Nnet-survival fitted to a fixed train data ($n_{\text{train}} = 1000$) and evaluated on the first test data ($n_{\text{test}} = 1000$) and the train data.



(a) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_1$ (b) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_1$ (c) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_1$

(d) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_2$ (e) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_2$ (f) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_2$

(g) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_3$ (h) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_3$ (i) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_3$

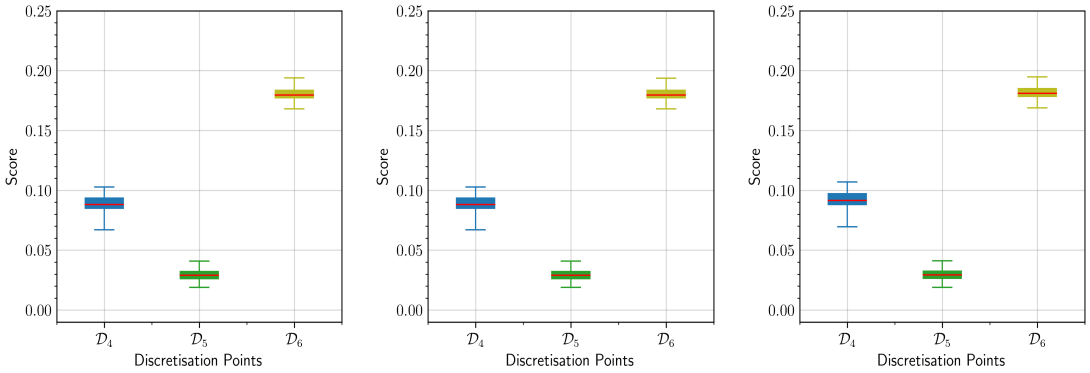
Figure 5.4: The predicted probabilities of interest of $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ from Scenario 1 of Simulation 1 for 20% randomly selected pairs $i \neq j$ grouped by $I_{\{T_i \leq T_j\}}$ over each discretisation setup. They were obtained from the overfitted Nnet-survival fitted to a fixed train data ($n_{\text{train}} = 1000$) and evaluated on the first test data ($n_{\text{test}} = 1000$) and the train data.

Scenario 2: Performance Evaluation of Different Models

This scenario uses the same setting and data generated in Scenario 2 of Simulation 1 in Section 3.1.1. However, the aim is to investigate how $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ behave in evaluating the predictive performance of different models fitted to PH data. To obtain various models, $\mathscr{D}_4, \mathscr{D}_5$, and $\mathscr{D}_6$ were employed to discretise the train and test data. The event time distributions of the train data from each discretisation setup can be seen in Figure 3.6. Note that the distributions of the event times in the test data were more or less the same to the train data since they were generated based on the same setting. Then, we fitted the good and overfitted Nnet-survival architectures in Table B.1 (Appendix B) to a train data ($n_{\text{train}} = 1000$) discretised by the three discretisation setups. Then, $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ were computed on 100 independent test data ($n_{\text{test}} = 1000$) that were discretised by the same discretisation setup.

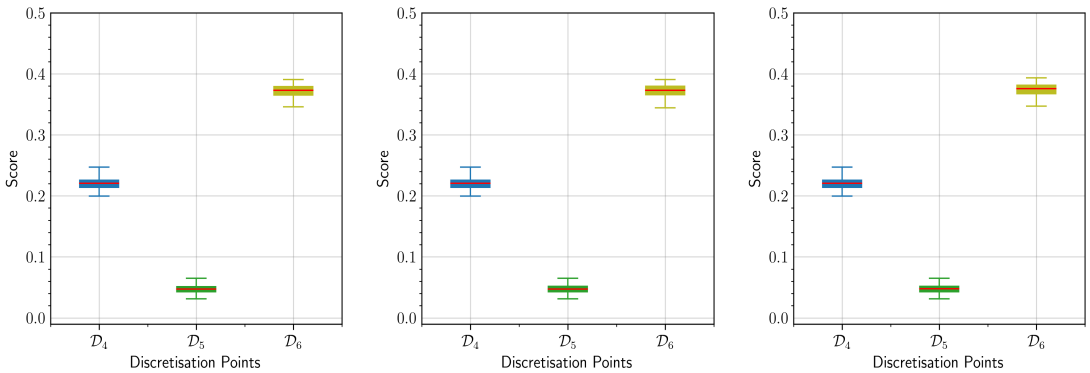
The implementation results from the good and overfitted models are given in Figure 5.5 and Figure 5.6, respectively. As we can see from the figures, as expected, $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ in the overfitted models are much higher than in the good models when the discretisation setups are $\mathscr{D}_4$ and $\mathscr{D}_6$. However, there is almost no difference between $\widehat{\text{PC}}_n^2$ in the good and overfitted models when we use $\mathscr{D}_5$. Even though we apply the overfitted models, we still have pair calibrations close to zero. By employing $\mathscr{D}_5$, $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ could not differentiate the good and the overfitted models. As we know from Section 3.1.1, most of the event times are in the first period resulting in most predicted survival curves are close to zero. Moreover, in Figure A.4, the predicted probabilities of interest of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ in the overfitted models are mostly very close to one regardless of the outcomes. Since most the outcomes are $I_{\{T_i \leq T_j\}} = 1$, particularly $I_{\{T_i = T_j\}} = 1$ (see panel (b) of Figure 3.6), then the values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ are very close to zero.

In summary, in this second scenario, we have elaborated that pair calibration might be affected by the test data structure. In the next section, we will see how $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ behave when the censoring rates in the test data increase but the discretisation setup in the test data is fixed.



(a) $\widehat{\text{PC}}_n^1$ (b) $\widehat{\text{PC}}_n^2$ (c) $\widehat{\text{PC}}_n^{1,T_{max}}$

Figure 5.5: Boxplots of (a) $\widehat{\text{PC}}_n^1$, (b) $\widehat{\text{PC}}_n^2$, and (c) $\widehat{\text{PC}}_n^{1,T_{max}}$ from Scenario 2 Simulation 1 over three sets of discretisation points in test data. They were estimated on 100 independent test data (n_{test}=1000) from the good Nnet-survival architecture fitted to a train data (n_{train}=1000) discretised by $\mathscr{D}_1, \mathscr{D}_2$, and $\mathscr{D}_3$.



(a) $\widehat{\text{PC}}_n^1$ (b) $\widehat{\text{PC}}_n^2$ (c) $\widehat{\text{PC}}_n^{1,T_{max}}$

Figure 5.6: Boxplots of (a) $\widehat{\text{PC}}_n^1$, (b) $\widehat{\text{PC}}_n^2$, and (c) $\widehat{\text{PC}}_n^{1,T_{max}}$ from Scenario 2 Simulation 1 over three sets of discretisation points in test data. They were estimated on 100 independent test data (n_{test}=1000) from the overfitted Nnet-survival architecture fitted to a train data (n_{train}=1000) discretised by $\mathscr{D}_1, \mathscr{D}_2$, and $\mathscr{D}_3$

5.3.2 Simulation 3: PH and Non-PH Data with Varying Censoring Rates and Fixed Discretisation Setup

This simulation investigates the behaviour of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ in both PH and non-PH data when censoring rates in the test data vary but the discretisation setup is fixed. Note that increasing the censoring rates in this simulation differs from the approach in Simulation 1. In Simulation 3, the additional censored individuals are mainly not due to administrative censoring. Theoretically, $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ converge to their respective expected values that do not depend on the censoring distribution by applying the IPCW approach to the estimators. However, as we will show in this simulation, it is unlikely to happen in numerical implementations.

We apply the scenario settings and the data generation mechanism in Simulation 1 (Section 3.1) and in Simulation 2 (Section 4.1) for the PH and non-PH data, respectively. For both types of data, we first fitted Nnet-survival architectures given in Table A.3, where the left column is for the non-PH data and the right column is for the PH data, to a fixed train data (n_{train}=1000), and then evaluated the model performance on 100 independent test data using $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$. The train and test data in non-PH data were discretised by $\mathscr{D}_{13}$={0, 5, 10, 15, 20, 25, 30, 40, 80, 90, 150, ∞} , while $\mathscr{D}_{14}$={0, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, ∞} is for the non-PH data as defined in Chapter 4.

Figure 5.7 reports the results of these numerical experiments. The figure shows that all the estimators behave almost the same as the censoring increases regardless of the PH or non-PH data. When the censoring rates in the test data are not too massive, i.e. up to 45% censoring rate, the values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ are stable and close to their values when the censoring rate in the test data is 0% assumed as the "true" values of the estimators. Then, starting from a 62% censoring rate, their values significantly move away from their true values with higher standard deviations. When censoring rate increases, the bias due to censoring distribution must be alleviated by the given weights $\widehat{G}_n^{-1}$ and $\widehat{G}_n^{-2}$. $\widehat{G}_n$ will never be completely specified when censoring presences in the data, that is $\widehat{G}_n \neq G$. Furthermore, weights get larger since we have to penalise larger ignored

pairs when the censoring rates are large (i.e. at 65% and 75%). The higher the censoring rates, the higher the bias of the estimators.

We can see also more significant differences between $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ when censoring rates are 75% even though theoretically they are unbiased estimators of (5.1) (Figure 5.7). We know that the only difference between $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ is the used indicator functions when $I_{\{T_i \leq T_j, T_i < T_{max}\}}$. Moreover, $\widehat{\text{PC}}_n^1$ contains

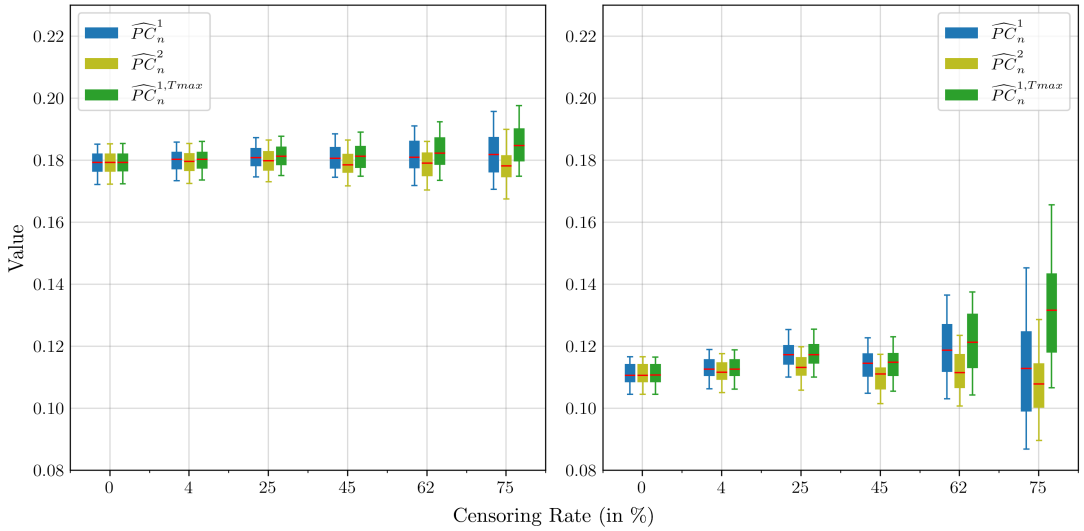
$$\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)} \frac{I_{\{T_j<D_j\}}}{\widehat{G}_n(T_j)}, \tag{5.19}$$

meanwhile $\widehat{\text{PC}}_n^2$ consists of

$$\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)} \frac{I_{\{T_i\leq D_j\}}}{\widehat{G}_n(T_i-1)}, \tag{5.20}$$

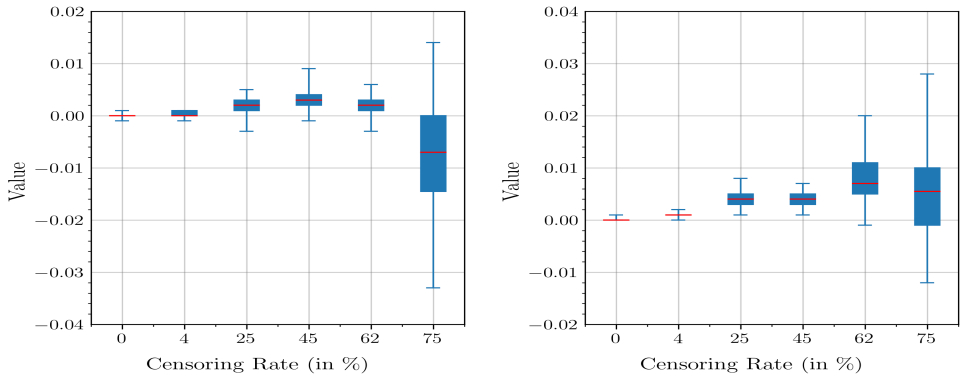
when $I_{\{T_i \leq T_j, T_i < T_{max}\}}$. Meanwhile, the rest of the terms in $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ are the same.

By the indicator functions contained in (5.19) and (5.20), the number of pairs $i \neq j$ used in $\widehat{\text{PC}}_n^2$ is greater or equal to that in $\widehat{\text{PC}}_n^1$. This is because (5.20) does not restrict that $T_j < D_j$. Although we may fit the most sophisticated model to estimate G, there will always be a misspecification error due to $\widehat{G}_n \neq G$ when we compute $\widehat{G}_n$ from the data containing censored individuals. Therefore, when the difference between the number of pairs in $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ increases due to censoring, the errors from their given weights also increase, resulting in the increase of $(\widehat{\text{PC}}_n^1 - \widehat{\text{PC}}_n^2)$ as the censoring rate increases (see Figure 5.8). When censoring rate is 75%, in the figure, generally, $\widehat{\text{PC}}_n^1 < \widehat{\text{PC}}_n^2$ in the PH data (panel(a)) and $\widehat{\text{PC}}_n^1 > \widehat{\text{PC}}_n^2$ in the non-PH data (panel(b)), depending on the values of their given weights.



(a) PH Data (b) Non-PH Data

Figure 5.7: Boxplots of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ in assesing the model performance fitted to PH data (a) and non-PH data (b) over six censoring rates in test data. They were computed on 100 independent test data (n_{test}=1000) from a fixed model fitted to a train data (n_{train}=1000). The censoring rates of the train data are very close to 0%. $\mathscr{D}_{13}$ and $\mathscr{D}_{14}$ were used to discretise the non-PH and PH data, respectively.



(a) PH Data (b) Non-PH Data

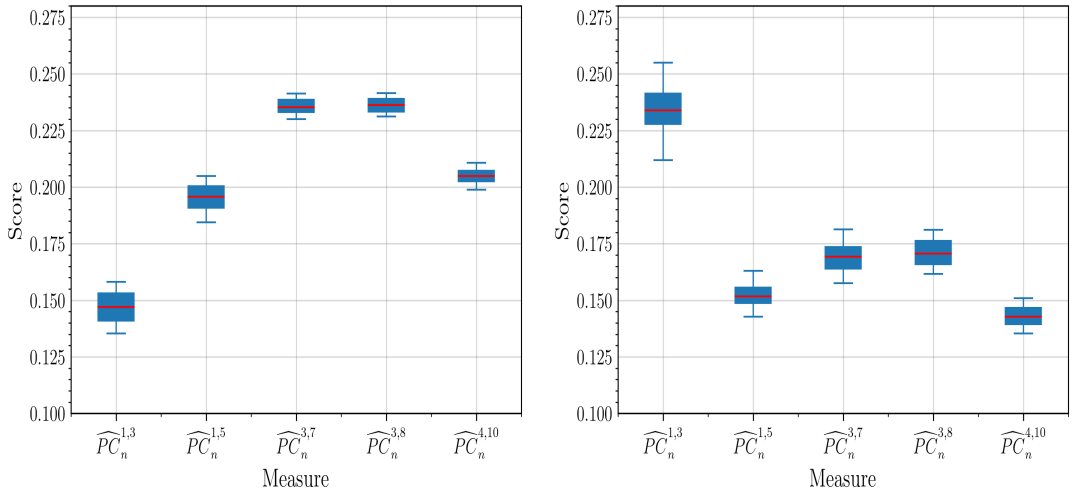
Figure 5.8: Boxplots of $(\widehat{\text{PC}}_n^1 - \widehat{\text{PC}}_n^2)$ in the PH data (a) and the non-PH data over six different censoring rates. They were estimated from 100 independent test data sets (n_{test}=1000) trained on a fixed almost fully uncensored data (n_{train}=1000).

In this case study, we have employed $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ in assessing a fixed model's predictive performance in PH and non-PH data. As the censoring rate increases, they are stable and relatively close to their 'true' value. However, we will see significant bias in the measures when the censoring rate is large. We also have found that the difference between $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ will be more apparent as the censoring rate increases. However, we cannot decide which one is the better estimator between $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ because the differences between them are due to misspecification of $\widehat{G}_n$ and the number of usable pairs. In the next section, we will apply the truncated pair calibration to assess the model's predictive performance in the PH and non-PH data.

5.3.3 Simulation 5: PH and Non-PH Data with Fixed Censoring Rate

This Simulation aims to apply the truncated pair calibration (5.14) to assess the predictive performance of the models fitted to the PH and non-PH data in Simulation 3. However, we do not vary the censoring rates in the train and test data. Moreover, we only computed the measures on test data with almost 0% censoring rate.

We evaluated the model performance using the truncated pair calibration on five different sub-intervals, namely [1,3), [1,5), [3,7), [3,8), and [4,10). The results are given in Figure 5.9, where panel (a) is for the PH data and panel (b) shows the results from non-PH data. In the PH data, for sub-interval locating on the left-tail of the follow-up time, i.e. [1,3], the model performance is relatively good, indicating that the models predict the probability of interest well. Nevertheless, as the sub-intervals move to the right-hand side of the follow-up, the prediction accuracy gets worse, indicated by higher values of the truncated pair calibration in other sub-intervals. On the other hand, in the non-PH data (panel (b) of Figure 5.9), the subinterval [1,3) has the highest values of the truncated pair calibration before it decreases significantly in other sub-intervals.



(a) PH Data

(b) Non-PH Data

Figure 5.9: Boxplots of the truncated pair calibration in the PH data (a) and the non-PH data over five different sub-intervals of the follow-up time. They were estimated from 100 independent test data sets (n_{test}=1000) trained on a fixed almost fully uncensored data (n_{train}=1000).

In summary, the values of the truncated pair calibration oscillate depending on the quality of the outputs at each sub-interval. The event time distributions obtained from the chosen discretisation setup significantly affect the model performance at each sub-interval. As expected, pair calibration, including its truncated version, does not depend on the proportionality assumption. It behaves the same regardless of whether the assumption holds in the data. From these results, the truncated pair calibration provides reasonable alternative solutions for anyone not interested in the model performance over the whole follow-up time. In the next section, we will discuss the behaviour of the proposed pair calibration in two real-world examples, namely TCGA data and breast cancer data.

5.4 Real-World Examples

This section can be seen as the continuation of the real-world examples in Section 3.1.2 (Chapter 3). We will use the same settings, the same data (i.e. TCGA data and breast cancer data) including the used covariates and discretisation setups, and the same objectives of the real-world examples in Chapter 3. The only difference is that we now assess the model performance on the test data using $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$.

5.4.1 TCGA Data

This data has been described in detail in Section 3.1.2. We still used the same implementation setting and employed the obtained 100 pairs of train and test data (see Figure 3.11 for the distribution of the observed times). After we fitted the good and overfitted architectures in Table B.2 (Appendix B.1) to the 100 train data (70% of the original data) discretised by $\mathscr{D}_7$, we computed $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ on their respective 100 independent test data (30% of the original data) discretised by $\mathscr{D}_7$, $\mathscr{D}_8$, and $\mathscr{D}_9$ as defined in Chapter 3.

In Figure 5.10, panel (a) presents the results from the good models, and panel (b) shows the results from the overfitted models. In both panels of the figure, $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ decrease as we vary the discretisation setups in the test data. These results are in line with their respective predicted probabilities of interest shown in Figure A.5 and Figure A.6 (Appendix A). The outputs for $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ are getting closer to one as we change the discretisation setups in the test data from $\mathscr{D}_7$ to $\mathscr{D}_8$ and $\mathscr{D}_9$. Since $\mathscr{D}_8$ and $\mathscr{D}_9$ are the subsets of $\mathscr{D}_7$ by ignoring some points in the right tail of $\mathscr{D}_7$, some of individuals in the test data discretised by $\mathscr{D}_8$ and $\mathscr{D}_9$ were administratively censored due to the truncation of the test data with $\mathscr{D}_7$. As a consequence, the probability of interest increases while most of the outcomes are $I_{\{T_i \leq T_j\}}$ resulting in smaller $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ in $\mathscr{D}_8$ and $\mathscr{D}_9$. Due to the increase of administratively censored individuals in $\mathscr{D}_8$ and $\mathscr{D}_9$, the probability (5.2) also increased so that in such situtation $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ were smaller.

In the good models, $\widehat{\text{PC}}_n^{1,T_{max}}$ is stable as we vary the discretisation setup (panel (a) of Figure 5.10). These results are consistent with Figure A.5, where

the outputs of $\widehat{\text{PC}}_n^{1,T_{max}}$ are almost the same regardless of the applied discretisation setups in the test data (see panels (c),(f), and (i) of the figure). Since $\widehat{\text{PC}}_n^{1,T_{max}}$ only considers the event times within the sub-interval of interest, increasing censoring in the data does not significantly affect $\widehat{\text{PC}}_n^{1,T_{max}}$. For the overfitted models (panel (b) of Figure 5.10), $\widehat{\text{PC}}_n^{1}$ and $\widehat{\text{PC}}_n^{2}$ are highly affected by the change of discretisation setup. Although at a glance, $\widehat{\text{PC}}_n^{1,T_{max}}$ is also affected by the discretisation setup change, actually, the values are in the range 0.325-0.4. In other words, the change of $\widehat{\text{PC}}_n^{1,T_{max}}$ does not support our justification of the models because, from the values, we still think that the models are bad. These results are also in line with the outputs of $\widehat{\text{PC}}_n^{1,T_{max}}$ displayed in Figure A.5 (Appendix A), where the change of the outputs of $\widehat{\text{PC}}_n^{1,T_{max}}$ over the discretisation setups is minimal.

To sum up, changing the discretisation setup in the test data have significantly affected the values of $\widehat{\text{PC}}_n^{1}$ and $\widehat{\text{PC}}_n^{2}$. One may be confused to justify the predictive performance of a fixed model performance using $\widehat{\text{PC}}_n^{1}$ and $\widehat{\text{PC}}_n^{2}$ when the test data structure changes. However, $\widehat{\text{PC}}_n^{1,T_{max}}$ is more stable as the discretisation setup changes, which is consistent to the results in Scenario 1 of Simulation 1. In the next subsection, we will give another real-world example for the implementation of $\widehat{\text{PC}}_n^{1}, \widehat{\text{PC}}_n^{2}$ and $\widehat{\text{PC}}_n^{1,T_{max}}$.

5.4.2 Breast Cancer Data

In this second real-world example, we used the same experiment settings, and the same train and test data in Section 3.1.2. We will investigate how $\widehat{\text{PC}}_n^{1}, \widehat{\text{PC}}_n^{2}$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ behave in various different models. To obtain the models, we combined three different discretisation setups, namely $\mathscr{D}_{10}$, $\mathscr{D}_{11}$, $\mathscr{D}_{12}$ as defined in Section 3.1.2 (Chapter 3), and three values of L^2-regularisation (λ) so that we have nine models from the combination. As explained in Section 3.1.2, we have 100 pairs of train and test data, where we fitted Nnet-survival architecture in Table B.3 to the train data, and then computed $\widehat{\text{PC}}_n^{1}, \widehat{\text{PC}}_n^{2}$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ on the test data. Note that the train and the respective test data are discretised using the same discretisation setup. We repeated the model fitting and prediction scenario

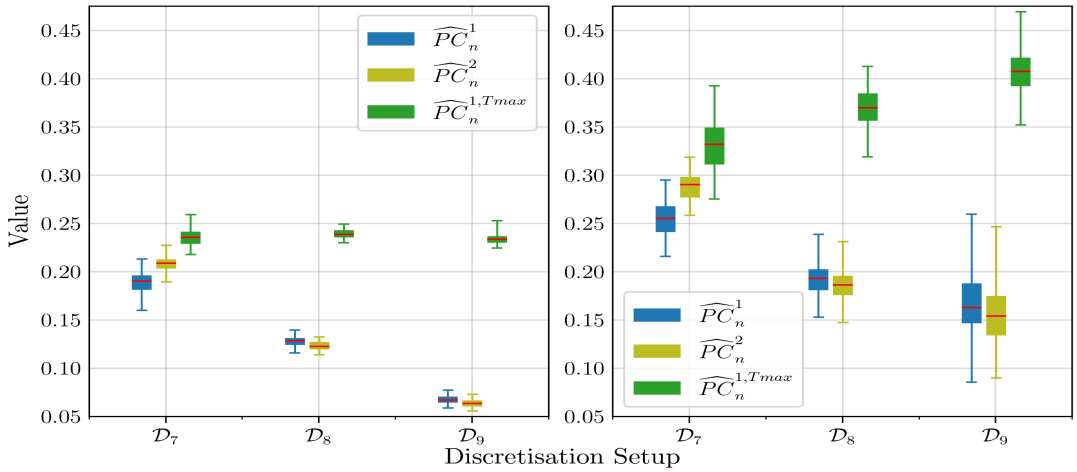


Figure 5.10: Boxplots of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ in (a) the good models and (b) the overfitted models obtained from 100 test data of TCGA data over three different discretisation setups

100 times following the number of the obtained 100 pairs of the train and test data.

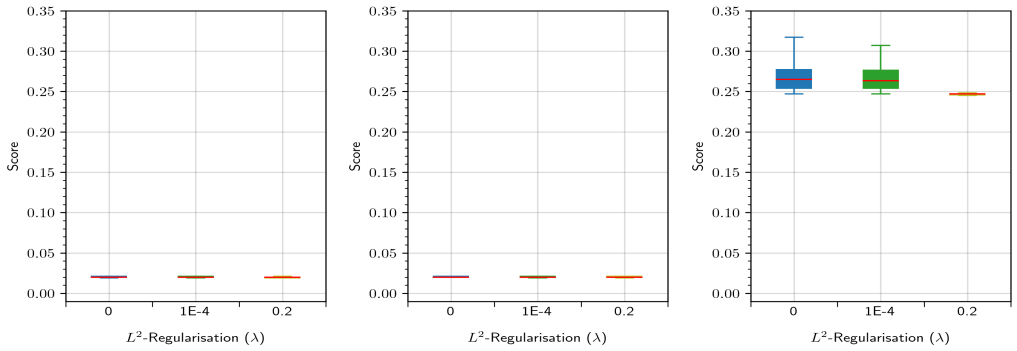
The main results of this numerical implementation are given in Figure 5.11. As we can see from the figure that, $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ are not affected by the change of λ. Meanwhile, $\widehat{\text{PC}}_n^{1,T_{max}}$ goes up as we increase the value of λ, where we obtain the best $\widehat{\text{PC}}_n^{1,T_{max}}$ at $\lambda = 0.2$ (panels (c),(f), and (i) of Figure 5.11). $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ always report close values to zero regardless of the fitted models. From the results of $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$, one will say that all models are always good. This justification may not appropriate because neural networks usually are not good when $\lambda = 0$. Due to the breast cancer data structure (see Figure 3.15), where most all individuals are administratively censored (around 95.5%), the predicted probability of interest (5.1) are always close to one. For example, panels (a)-(f) of Figure 5.12 show that the values of $\pi_{ij}^{T_{max}}$ from 20% randomly selected usable pairs in $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ are mostly close to one. As the λ increase from 0 to 0.2, their values have smaller standard deviation and even closer to one. Since most of the outcomes are $I_{\{T_i \leq T_j\}} = 1$, $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ will be even close to zero.

$\widehat{\text{PC}}_n^{1,T_{max}}$ ignores the censored individuals within the follow-up time. In partic-

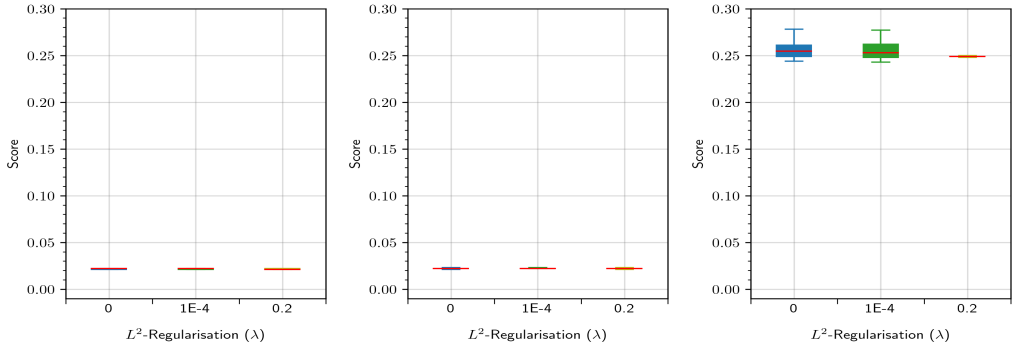
ular, it is computed based on $\widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i \vee T_j < T_{max}\right]$ from pairs of uncensored individuals ($i \neq j$) with $T_i \vee T_j < T_{max}$ only. Since the predicted probability of interest $\widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i \vee T_j < T_{max}\right]$ from such pairs are mostly between [0.55-0.6] (see panels (g)-(i) of Figure 5.12 as an example obtained from $\mathscr{D}_{10}$), $\widehat{\text{PC}}_n^{1,T_{max}}$ mostly lie within the range [0.18-0.25] for all combinations of discretisation setup and λ (see panels (c),(f), and (i) in Figure 5.11). Although the values of $\widehat{\text{PC}}_n^{1,T_{max}}$ are significantly far from zero $\widehat{\text{PC}}_n^{1}$ and $\widehat{\text{PC}}_n^{2}$, they still have almost uniform values regardless of the fitted models.

In summary, the values of $\widehat{\text{PC}}_n^{1}$ and $\widehat{\text{PC}}_n^{2}$ are always good when the censoring rate in the data is extremely large regardless the quality of the fitted models. On the other hand, $\widehat{\text{PC}}_n^{1,T_{max}}$ may differ to $\widehat{\text{PC}}_n^{1}$ and $\widehat{\text{PC}}_n^{2}$ because $\widehat{\text{PC}}_n^{1,T_{max}}$ only evaluates the predicted probability of interest from smaller number of pairs satisfying some conditions, namely pairs of uncensored ($i \neq j$) whose event times less than T_{max}. Although $\widehat{\text{PC}}_n^{1,T_{max}}$ may report the more make sense model performance than $\widehat{\text{PC}}_n^{1}$ and $\widehat{\text{PC}}_n^{2}$, ignoring too many censored individuals may underestimate the 'true' model performance.

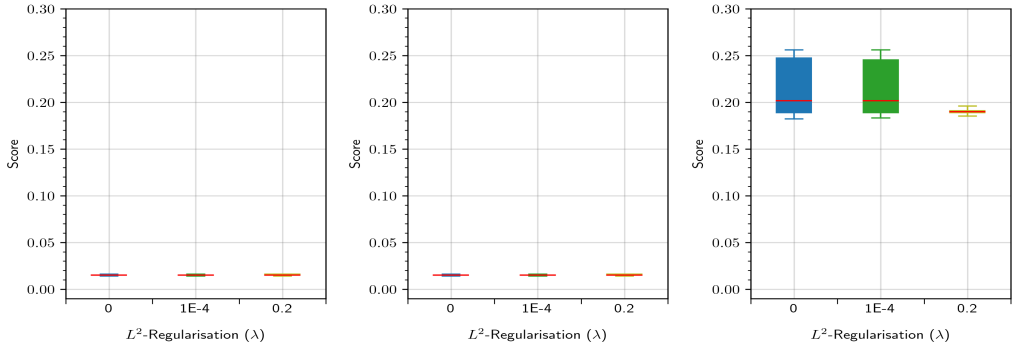
This section discussed some case studies to investigate the behaviour of some estimators of pair calibration. We have found that pair calibration sometimes may not well justify the model predictive performance. Furthermore, it is not sensitive the model change when (administrative) censoring rate is extremely large. In the next section, we will introduce some reference values as the complements of for pair calibration to have a better justification when evaluating the model performance.



(a) $\widehat{\mathrm{PC}}_n^1$ with $\mathcal{D}_{10}$ (b) $\widehat{\mathrm{PC}}_n^2$ with $\mathcal{D}_{10}$ (c) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathcal{D}_{10}$

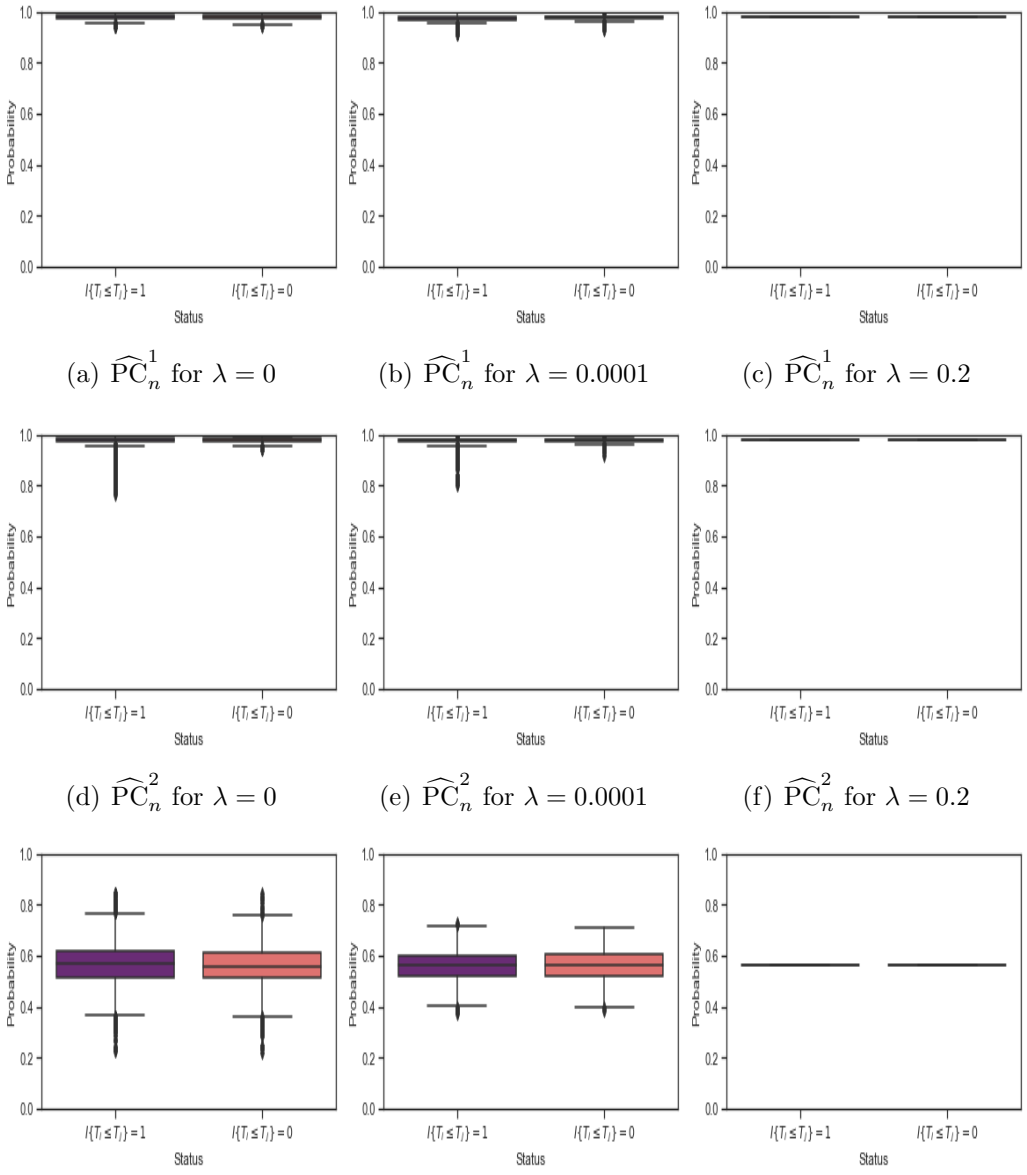


(d) $\widehat{\mathrm{PC}}_n^1$ with $\mathcal{D}_{11}$ (e) $\widehat{\mathrm{PC}}_n^2$ with $\mathcal{D}_{11}$ (f) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathcal{D}_{11}$



(g) $\widehat{\mathrm{PC}}_n^1$ with $\mathcal{D}_{12}$ (h) $\widehat{\mathrm{PC}}_n^2$ with $\mathcal{D}_{12}$ (i) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathcal{D}_{12}$

Figure 5.11: $\widehat{\mathrm{PC}}_n^1, \widehat{\mathrm{PC}}_n^2$ and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ over different values of L^2- regularisation (λ) for three splitting points, i.e. $\mathcal{D}_{10}$, $\mathcal{D}_{11}$, and $\mathcal{D}_{12}$. They were estimated on 100 breast cancer test data from the models fitted to 100 breast cancer train data.



(a) $\widehat{\mathrm{PC}}_n^1$ for $\lambda = 0$ (b) $\widehat{\mathrm{PC}}_n^1$ for $\lambda = 0.0001$ (c) $\widehat{\mathrm{PC}}_n^1$ for $\lambda = 0.2$

(d) $\widehat{\mathrm{PC}}_n^2$ for $\lambda = 0$ (e) $\widehat{\mathrm{PC}}_n^2$ for $\lambda = 0.0001$ (f) $\widehat{\mathrm{PC}}_n^2$ for $\lambda = 0.2$

(g) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for $\lambda = 0$ (h) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for $\lambda = 0.0001$ (i) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for $\lambda = 0.2$

Figure 5.12: The predicted probabilities of interest of $\widehat{\mathrm{PC}}_n^1$, $\widehat{\mathrm{PC}}_n^2$, and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ in breast cancer data for 20% randomly selected pairs $i \neq j$ grouped by $I_{\{T_i \leq T_j\}}$ overL^2-regularisation (λ). They were obtained from the net-survival fitted to a fixed train data ($n_{\text{train}} = 1000$) and evaluated on the first test data ($n_{\text{test}} = 1000$) and the train data. The train and test data were discretised by $\mathscr{D}_{10}$.

5.5 Reference Values based on The Worst Prediction

One approach to developing the reference value for a measure is to employ the measure in evaluating the performance of the worst model fitted to the data. In particular, one of the worst models may happen when the model outputs the same values for all individuals regardless of the specific characteristics of each individual. If a model always outputs either 1 or 0 for our predicted probabilities of interests, e.g. (5.1), then the model cannot learn from the data. Based on this idea, in this section, we propose the values of pair calibration obtained from models outputting 1 or 0 for each predicted probability of interest as the reference values. Therefore, a good model should have pair calibration much less than the reference value. The model's predictive performance is poor if the pair calibration equals the reference values. If the pair calibration is much higher than the reference values, we have really bad model performance.

5.5.1 Formulation of The Reference Values

Denote $\chi_1 = 1$ and $\chi_2 = 0$. The formulas of the first type of reference values for $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ for χ_1 are defined as follows

$$\begin{aligned}
&\text{ref}_1^{\chi_1} \\
&= \left(1 \Big/ \frac{1}{n}\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_i<D_i\}}\widehat{G}_n^{-1}(T_i)\right]\right) \\
&\quad \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\left[\left(I_{\{T_i\leq T_j\}}-\chi_1\right)^2\frac{I_{\{T_i<D_i\}}I_{\{T_j<D_j\}}}{\widehat{G}_n(T_i)\widehat{G}_n(T_j)}I_{\{T_i<T_{max}\}}\right],
\end{aligned} \tag{5.21}$$

$$\begin{aligned}
&\text{ref}_2^{\chi_1} \\
&= \left(1 \Big/ \frac{1}{n}\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_i<D_i\}}\widehat{G}_n^{-1}(T_i)\right]\right) \\
&\quad \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\left[\left(1-\chi_1\right)^2\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_i\leq T_j\}}}{\widehat{G}_n(T_i-1)}I_{\{T_i<T_{max}\}}\right. \\
&\quad \left.+\left(0-\chi_1\right)^2 I_{\{T_j<T_i\}}\frac{I_{\{T_j<D_j\}}}{\widehat{G}_n(T_j)}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}I_{\{T_i<T_{max}\}}\right],
\end{aligned} \tag{5.22}$$

and

$$\begin{aligned}&\mathrm{ref}_3^{\chi_1}\\&=\frac{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\left[\left(I_{\{T_i\leq T_j\}}-\chi_1\right)^2\frac{I_{\{T_i<D_i\}}I_{\{T_j<D_j\}}}{\widehat{G}_n(T_i)\widehat{G}_n(T_j)}I_{\{T_i<T_{max}\}}I_{\{T_j<T_{max}\}}\right]}{\frac{1}{n}\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_j<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_j<D_j\}}}{\widehat{G}_n(T_j)}\right]},\end{aligned}\tag{5.23}$$

respectively. The conditions of the reference values are: (1) if the estimators of pair calibration are much less than their respective reference values, the model performance is excellent, (2) if the estimators are close to their respective reference values, the model performance is poor, and (3) if the estimators much greater than their respective reference values, we have very bad model performance.

For χ_2 , we can replace χ_1 with χ_2 in (5.21), (5.22), and (5.23) so that we will easily obtain the formulas of $\mathrm{ref}_1^{\chi_2}$, $\mathrm{ref}_2^{\chi_2}$, and $\mathrm{ref}_3^{\chi_2}$, respectively. Since $\chi_1+\chi_2=1$ while other values are the same, then we obtain relationships as follows

$$\mathrm{ref}_r^{\chi_1}=1-\mathrm{ref}_r^{\chi_2}\tag{5.24}$$

for $r=1,2,3$. In the paragraph, hlwe will show only (5.24) for $r=1$ since (5.24) for $r=2$ and $r=3$ follow the results.

Denote $\chi=\{\chi_1,\chi_2\}$, and let p_i be the joint probability mass function of $\left(I_{\{T_i\leq T_j\}}-\chi\right)^2$ and $(T_i<T_{max})$ for $I_{\{T_i\leq T_j\}}=1$, and consequently $(1-p_i)$ be the joint probability mass function of $\left(I_{\{T_i\leq T_j\}}-\chi\right)^2$ and $(T_i<T_{max})$ for $I_{\{T_i\leq T_j\}}=0$. When $\chi=\chi_1$,

$$\begin{aligned}&\mathbb{E}\left[\left(I_{\{T_i\leq T_j\}}-\chi_1\right)^2\middle|T_i<T_{max}\right]\\&=\mathbb{E}\left[(1-1)^2\middle|T_i<T_{max}\right]p_i+\mathbb{E}\left[(0-1)^2\middle|T_i<T_{max}\right](1-p_i)\\&=0+\mathbb{E}\left[1\middle|T_i<T_{max}\right](1-p_i)\\&=(1-p_i)\end{aligned}\tag{5.25}$$

Meanwhile, when $\chi=\chi_2$,

$$\begin{aligned}&\mathbb{E}\left[\left(I_{\{T_i\leq T_j\}}-\chi_2\right)^2\middle|T_i<T_{max}\right]\\&=\mathbb{E}\left[(1-\chi_2)^2\middle|T_i<T_{max}\right]p_i+\mathbb{E}\left[(0-\chi_2)^2\middle|T_i<T_{max}\right](1-p_i)\\&=\mathbb{E}\left[(1-0)^2\middle|T_i<T_{max}\right]p_i+\mathbb{E}\left[(0-0)^2\middle|T_i<T_{max}\right](1-p_i)\\&=\mathbb{E}\left[1\middle|T_i<T_{max}\right]p_i+0\\&=p_i.\end{aligned}\tag{5.26}$$

By the last equalities in (5.25) and (5.26) , we obtain

$$\mathbb{E}\left[\left(I_{\{T_i\leq T_j\}}-\chi_1\right)^2\Big|T_i<T_{max}\right]+\mathbb{E}\left[\left(I_{\{T_i\leq T_j\}}-\chi_2\right)^2\Big|T_i<T_{max}\right]=1. \quad (5.27)$$

Since $\text{ref}_1^{\chi_1}$ and $\text{ref}_1^{\chi_2}$ are the estimators of $\mathbb{E}\left[\left(I_{\{T_i\leq T_j\}}-\chi_1\right)^2\Big|T_i<T_{max}\right]$ and $\mathbb{E}\left[\left(I_{\{T_i\leq T_j\}}-\chi_2\right)^2\Big|T_i<T_{max}\right]$, respectively, then (5.24) for $r=1$ is valid.

5.5.2 Pitfall of The Reference Values

Recall the formulas of $\text{ref}_r^{\chi_1}$ and $\text{ref}_r^{\chi_2}$ for $r=1,2,3$. Although these reference values are easily implemented and interpreted, we will show their crucial pitfall via numerical experiments. We will implement $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ along with $\text{ref}_r^{\chi_1}$ and $\text{ref}_r^{\chi_2}$ $(r=1,2,3)$ to assess the model performance in Simulation 1. We employ the train and test PH data generated in Simulation 1 (Chapter 3) with two discretisation setups in test data. Moreover, we first fitted the overfitted Nnet-survival architecture in Table B.1 to a train data discretised by $\mathscr{D}_1$, and then evaluated the model performance on the first five test data discretised by $\mathscr{D}_2$ and $\mathscr{D}_3$.

The results of the numerical experiments are given in Table 5.1. We have known from the results shown in Figure 3.3 that the models overfitted the data for all discretisation setups $\mathscr{D}_1, \mathscr{D}_2$, and $\mathscr{D}_3$. However, as we can see in Table 5.1, by using $\text{ref}_r^{\chi_1}$ and $\text{ref}_r^{\chi_2}$ $(r=1,2,3)$ altogether as the reference values, we can inconsistently justify the model performance. For instance, consider the results for $\mathscr{D}_2$ in Table 5.1. We can see that $\text{ref}_1^{\chi_1} < \widehat{\text{PC}}_n^1 < \text{ref}_1^{\chi_2}$ confusing us to conclude the model performance. On the other hand, $\widehat{\text{PC}}_n^{1,T_{max}}$ is not in line with $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$, where $\widehat{\text{PC}}_n^{1,T_{max}}$ is always better than the reference values, although we know that the models overfitted the data. The more extreme case is shown by the results of $\mathscr{D}_3$, where all the estimators $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ are much worse than their respective reference values $\text{ref}_r^{\chi_1}$ $(r=1,2,3)$, but they are much better than their respective $\text{ref}_r^{\chi_2}$ $(r=1,2,3)$.

$\mathscr{D}_2$								
$\widehat{\text{PC}}_n^1$	$\text{ref}_1^{\chi_1}$	$\text{ref}_1^{\chi_2}$	$\widehat{\text{PC}}_n^2$	$\text{ref}_2^{\chi_1}$	$\text{ref}_2^{\chi_2}$	$\widehat{\text{PC}}_n^{1,T_{max}}$	$\text{ref}_3^{\chi_1}$	$\text{ref}_3^{\chi_2}$
0.321	0.299	0.701	0.319	0.299	0.701	0.410	0.447	0.553
0.330	0.298	0.702	0.330	0.298	0.702	0.433	0.448	0.552
0.345	0.313	0.687	0.344	0.313	0.687	0.428	0.446	0.554
0.342	0.311	0.689	0.341	0.311	0.689	0.421	0.444	0.556
0.337	0.322	0.678	0.336	0.322	0.678	0.415	0.447	0.553
$\mathscr{D}_3$								
$\widehat{\text{PC}}_n^1$	$\text{ref}_1^{\chi_1}$	$\text{ref}_1^{\chi_2}$	$\widehat{\text{PC}}_n^2$	$\text{ref}_2^{\chi_1}$	$\text{ref}_2^{\chi_2}$	$\widehat{\text{PC}}_n^{1,T_{max}}$	$\text{ref}_3^{\chi_1}$	$\text{ref}_3^{\chi_2}$
0.256	0.041	0.959	0.251	0.041	0.959	0.421	0.235	0.765
0.286	0.041	0.959	0.281	0.041	0.959	0.423	0.244	0.756
0.244	0.047	0.953	0.243	0.047	0.953	0.367	0.248	0.752
0.265	0.043	0.957	0.262	0.043	0.957	0.409	0.224	0.776
0.231	0.047	0.953	0.226	0.047	0.953	0.389	0.250	0.750

Table 5.1: The reference values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ for the first five test data discretised by $\mathscr{D}_2$ and $\mathscr{D}_3$. These results were obtained from the overfitted Nnet-survival architecture in Simulation 1.

The inconsistent justification, as we have shown in the numerical experiment, because the used discretisation setups have made the difference between the reference values $\text{ref}_r^{\chi_1}$ and $\text{ref}_r^{\chi_2}$ $(r = 1, 2, 3)$ large. Since in the overfitted models we have large values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$, then they tend to lie between $\text{ref}_r^{\chi_1}$ and $\text{ref}_r^{\chi_2}$ $(r = 1, 2, 3)$, resulting in confusing model performance justification. This is because the model's outcomes $I_{\{T_i \leq T_j\}}$ tend to one of their possible values, 1 or 0. Therefore, a solution to such an issue is to check first the data distribution. If most of the outcomes are 1, then we should use $\text{ref}_r^{\chi_1}$ $(r = 1, 2, 3)$. Otherwise, we should use $\text{ref}_r^{\chi_2}$ $(r = 1, 2, 3)$. However, this solution is impractical since we have to check the distribution of outcomes before deciding which reference values to use.

To sum up, we have developed new reference values for pair calibration based on one of the worst possible models we may have. The reference values evaluate the predicted probabilities of interest with only two possible values, namely 1 or

0. We have shown in our experiments that the reference values may have pitfalls when pair calibration lies between $\text{ref}_r^{\chi_1}$ and $\text{ref}_r^{\chi_2}$. Although we can cope with this issue, the proposed solution is impractical. Therefore, in the next section, we will propose another type of reference value for pair calibration that does not have such weakness.

5.6 Reference Values based on The Outcomes Proportion

This section aims to propose another type of the reference values for pair calibration based on proportion of the outcomes in the test data. The reference values are computed empirically from the test data regardless the training process using the same formulas as for their respective pair calibration estimators. Since those reference values provide only one value for each pair calibration estimator, they will not be suffered by the same issues in the previous reference values.

5.6.1 Formulation of The Reference Values

To derive the reference value, we first recall the probabilities of interest in $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$. Suppose these quantities are computed as the proportion of usable pairs in the numerator over the number of usable pairs in the denominator of those estimators. In particular, we first define

$$\tilde{p}_1 = \frac{\frac{1}{n(n-1)} \sum_{i \neq j}^{n} \left[I_{\{T_i \leq T_j\}} \frac{I_{\{T_i < D_i\}} I_{\{T_j < D_j\}}}{\widehat{G}_n(T_i) \widehat{G}_n(T_j)} I_{\{T_i < T_{max}\}} \right]}{\frac{1}{n} \sum_{i=1}^{n} \left[I_{\{T_i < T_{max}\}} \frac{I_{\{T_i < D_i\}}}{\widehat{G}_n(T_i)} \right]},$$

$$\begin{aligned} &\tilde{p}_2 \\ &= \left(1 \Big/ \frac{1}{n} \sum_{i=1}^{n} \left[I_{\{T_i < T_{max}\}} I_{\{T_i < D_i\}} \widehat{G}_n^{-1}(T_i) \right] \right) \\ &\quad \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \left[\frac{I_{\{T_i < D_i\}}}{\widehat{G}_n(T_i)} \frac{I_{\{T_i \leq T_j\}}}{\widehat{G}_n(T_i - 1)} I_{\{T_i < T_{max}\}} \right] \end{aligned}$$

as the estimators of $\pi_{ij}^{T_{max}}$, and

$$\tilde{p}_3 = \frac{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\left[I_{\{T_i\leq T_j\}}\frac{I_{\{T_i<D_i\}}I_{\{T_j<D_j\}}}{\widehat{G}_n(T_i)\widehat{G}_n(T_j)}I_{\{T_i<T_{max}\}}I_{\{T_j<T_{max}\}}\right]}{\frac{1}{n}\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_j<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_j<D_j\}}}{\widehat{G}_n(T_j)}\right]}$$

as the estimator of $\mathbb{P}\left[T_i \leq T_j \middle| T_i \vee T_j < T_{max}\right]$, where $\tilde{p}_1, \tilde{p}_2$, and $\tilde{p}_3$ are computed from the test data. We can see that they are obtained by taking out the squared difference terms from their respective pair calibration estimators.

We then replace $\pi_{ij}^{T_{max}}$ in $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ by $\tilde{p}_1$ and $\tilde{p}_2$, respectively. Meanwhile, $\widehat{\mathbb{P}}\left[T_i \leq T_j \middle| T_i \vee T_j < T_{max}\right]$ in $\widehat{\text{PC}}_n^{1,T_{max}}$ is replaced by $\tilde{p}_3$. Hence, we propose

$$\begin{aligned}
&\text{ref}_1 \\
&= \frac{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\left[\left(I_{\{T_i\leq T_j\}}-\tilde{p}_1\right)^2\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_i\leq T_j\}}}{\widehat{G}_n(T_i-1)}I_{\{T_i<T_{max}\}}\right]}{\frac{1}{n}\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\right]},
\end{aligned} \tag{5.28}$$

$$\begin{aligned}
&\text{ref}_2 \\
&= \left(1\Big/\frac{1}{n}\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_i<D_i\}}\widehat{G}_n^{-1}(T_i)\right]\right) \\
&\quad \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\left[\left(1-\tilde{p}_2\right)^2\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_i\leq T_j\}}}{\widehat{G}_n(T_i-1)}I_{\{T_i<T_{max}\}}\right. \\
&\quad \left.+\left(0-\tilde{p}_2\right)^2 I_{\{T_j<T_i\}}\frac{I_{\{T_j<D_j\}}}{\widehat{G}_n(T_j)}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}I_{\{T_i<T_{max}\}}\right],
\end{aligned} \tag{5.29}$$

and

$$\begin{aligned}
&\text{ref}_3 \\
&= \frac{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\left[\left(I_{\{T_i\leq T_j\}}-\tilde{p}_3\right)^2\frac{I_{\{T_i<D_i\}}I_{\{T_j<D_j\}}}{\widehat{G}_n(T_i)\widehat{G}_n(T_j)}I_{\{T_i<T_{max}\}}I_{\{T_j<T_{max}\}}\right]}{\frac{1}{n}\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_j<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{\widehat{G}_n(T_i)}\frac{I_{\{T_j<D_j\}}}{\widehat{G}_n(T_j)}\right]}.
\end{aligned} \tag{5.30}$$

as the second types of references values for $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$, respectively. We then defined three conditions for the reference values: (1) if the estimators are much less than their respective reference values, we have excellent model performance, (2) if the estimators are close to their respective reference values, we have poor model performance, and (3) if the estimators much greater than their respective reference values, we have very bad model performance.

The reference values developed in this section have only one possible value as opposed to the first reference values, which have two possible cases based on χ_1 and χ_2. Therefore, we will not have two opposite conclusions when evaluating the predictive performance of a model. Moreover, if $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ are much less than the reference values, then we have an excellent model performance. The model performance is poor if the estimators equal the reference values. We have very bad model performance if they are much higher than the reference values. In the next section, we will see how these reference values behave in several case studies.

5.6.2 Case Studies

This section is devoted to applying the second reference values for $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ to the numerical experiments that have been implemented in several sections of this chapter. We moreover complement the proposed reference values to the Simulation 1 (Section 5.3.1) and the Real-World Examples (Section 5.4).

Simulation 1

Simulation 1 comprises two scenarios: Scenario 1 and Scenario 2. The results from the good and overfitted models in Scenario 1 are given in Table 5.2 and Table 5.3, respectively. As we know from Section 5.3.1, $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ may change significantly as the test data structure changed due to the raise of (administrative) censoring rates. Consequently, by changing the test data structure, where in our setting we varied the discretisation setup, one may manipulate the data structure to achieve the desired values of $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$. On the other hand, with the same setting, $\widehat{\text{PC}}_n^{1,T_{max}}$ was more stable in such situation because it only evaluated pairs of uncensored individuals ($i \neq j$) for any $T_i, T_j < T_{max}$.

Table 5.2 displays $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ from the first five test data accompanied with their respective reference values for three discretisation setups (i. e. $\mathscr{D}_1$, $\mathscr{D}_2$, and $\mathscr{D}_3$) in the good models. If we only use $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ solely, we can see the significant difference in their values over the three discretisation setups. For instance, the values of $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_1$ are within the range [0.176,0.186], but

then they fall dramatically to the range [0.033,0.041] in $\mathscr{D}_3$. After accompanying them with their respective reference values, we can see that $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ for all discretisation setups are less than their reference values, meaning that the model performance is good regardless of the discretisation setups used in the test data. The results of the first five test data from the overfitted models for each $\mathscr{D}_1$, $\mathscr{D}_2$, and $\mathscr{D}_3$ are given in Table 5.3. For all discretisation setups, we can see from the table that $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ are much greater than ref_1, ref_2, and ref_3, respectively. Hence, we can conclude that the models' performance are really bad, consistent with the results shown in Figure 3.3.

The results of the first five test data for the good models in Scenario 2 are given in Table 5.4. From the table, $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ are always smaller than ref_1, ref_2, and ref_3, respectively, indicating that the model performance is good regardless of the used discretisation setups in the test data. On the other hand, in the bad models, $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ are always greater than their respective reference values as given by Table C.1.

To sum up, by accompanying $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ with their reference values ref_1, ref_2, and ref_3, respectively, we can cope with some issues discussed in Simulation 1. By comparing the estimators to the reference values, we moreover can be more convinced in justifying the predictive performance of the fitted models. In the next section, we will investigate how the reference values behave as the complementary measures for pair calibration in TCGA data.

$\mathscr{D}_1$					
$\widehat{\mathrm{PC}}_n^1$	ref_1	$\widehat{\mathrm{PC}}_n^2$	ref_2	$\widehat{\mathrm{PC}}_n^{1,T_{max}}$	ref_3
0.176	0.248	0.176	0.248	0.177	0.249
0.180	0.249	0.180	0.249	0.181	0.249
0.186	0.249	0.186	0.249	0.187	0.249
0.186	0.248	0.186	0.248	0.188	0.249
0.181	0.248	0.180	0.248	0.182	0.249
$\mathscr{D}_2$					
$\widehat{\mathrm{PC}}_n^1$	ref_1	$\widehat{\mathrm{PC}}_n^2$	ref_2	$\widehat{\mathrm{PC}}_n^{1,T_{max}}$	ref_3
0.151	0.209	0.151	0.209	0.191	0.247
0.159	0.209	0.159	0.209	0.203	0.247
0.168	0.313	0.215	0.313	0.215	0.247
0.168	0.311	0.214	0.311	0.214	0.247
0.164	0.322	0.218	0.322	0.218	0.247
$\mathscr{D}_3$					
$\widehat{\mathrm{PC}}_n^1$	ref_1	$\widehat{\mathrm{PC}}_n^2$	ref_2	$\widehat{\mathrm{PC}}_n^{1,T_{max}}$	ref_3
0.033	0.039	0.033	0.039	0.168	0.180
0.035	0.040	0.034	0.040	0.174	0.184
0.041	0.045	0.040	0.045	0.180	0.187
0.037	0.041	0.036	0.041	0.166	0.174
0.041	0.045	0.041	0.045	0.186	0.187

Table 5.2: The reference values of $\widehat{\mathrm{PC}}_n^1, \widehat{\mathrm{PC}}_n^2$ and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for the first five test data discretised by $\mathscr{D}_1, \mathscr{D}_2$, and $\mathscr{D}_3$. These results were obtained from the good Nnet-survival architecture in Scenario 1 of Simulation 1.

$\mathscr{D}_1$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.354	0.248	0.354	0.248	0.357	0.249
0.364	0.249	0.365	0.249	0.366	0.249
0.371	0.249	0.371	0.249	0.374	0.249
0.377	0.248	0.376	0.248	0.379	0.249
0.358	0.248	0.358	0.248	0.360	0.249
$\mathscr{D}_2$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.321	0.209	0.319	0.209	0.410	0.247
0.330	0.209	0.330	0.209	0.433	0.247
0.345	0.215	0.344	0.215	0.428	0.247
0.342	0.214	0.341	0.214	0.421	0.247
0.337	0.218	0.336	0.218	0.415	0.247
$\mathscr{D}_3$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.256	0.039	0.251	0.039	0.421	0.180
0.286	0.040	0.281	0.040	0.423	0.184
0.244	0.045	0.243	0.045	0.367	0.187
0.265	0.041	0.262	0.041	0.409	0.174
0.231	0.045	0.226	0.045	0.389	0.187

Table 5.3: The reference values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ for the first five test data discretised by $\mathscr{D}_1, \mathscr{D}_2$, and $\mathscr{D}_3$. These results were obtained from the over-fitted Nnet-survival architecture in Scenario 1 of Simulation 1.

$\mathscr{D}_4$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.076	0.127	0.076	0.127	0.080	0.132
0.084	0.126	0.084	0.126	0.088	0.131
0.097	0.137	0.097	0.137	0.100	0.141
0.095	0.135	0.095	0.135	0.099	0.140
0.097	0.135	0.097	0.135	0.100	0.138
$\mathscr{D}_5$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.033	0.036	0.033	0.036	0.033	0.036
0.037	0.043	0.037	0.043	0.037	0.043
0.026	0.029	0.026	0.029	0.026	0.029
0.027	0.032	0.028	0.032	0.028	0.032
0.023	0.023	0.023	0.023	0.023	0.023
$\mathscr{D}_6$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.173	0.242	0.172	0.242	0.174	0.243
0.179	0.242	0.179	0.242	0.180	0.243
0.184	0.242	0.184	0.242	0.185	0.242
0.184	0.241	0.184	0.241	0.186	0.242
0.175	0.241	0.175	0.241	0.177	0.242

Table 5.4: The reference values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ for the first five test data discretised by $\mathscr{D}_4, \mathscr{D}_5$, and $\mathscr{D}_6$. These results were obtained from the good Nnet-survival architecture in Scenario 2 of Simulation 1.

TCGA Data

This section is a continuation of the implementation of pair calibration to TCGA data in Section 5.4. In the experiment, we found that $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ varied as the administrative censoring increases due to the change of discretisation setup in the test data while the model is fixed. For the same setting, $\widehat{\text{PC}}_n^{1,T_{max}}$ was more stable for all discretisation setups in the same test data. We will accompany the three estimators of pair calibration with their respective reference values, namely

$\text{ref}_1, \text{ref}_2$, and ref_3.

In this implementation, we still use the same setting, Nnet-survival architectures, the train and test data used in Section 5.4. The results for three discretisation setups, such as $\mathscr{D}_7, \mathscr{D}_8$, and $\mathscr{D}_9$, in the test data are given completely in Table 5.5 and Table C.2 (Appendix C.1). Note that we only report the results from the first five test data .

Table 5.5 shows the results obtained from the good models using the good Nnet-survival architecture in Table B.2. We can see from the table that $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ are less than their respective reference values $\text{ref}_1, \text{ref}_2$, and ref_3 for all discretisation setups. These results means that they are consistent as opposed to when we do not use the reference values. Moreover, we can conclude that the models' performance is good. Meanwhile, Table C.2 (Appendix C.1) shows the results obtained from the overfitted Nnet-survival architecture. From the table, $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ are much greater than their respective reference values, meaning that we have really bad predictive performance in the overfitted models. In the next section, we will present the implementation to the breast cancer data.

$\mathscr{D}_7$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.197	0.205	0.214	0.479	0.241	0.249
0.193	0.203	0.207	0.441	0.237	0.249
0.188	0.207	0.203	0.446	0.222	0.249
0.199	0.210	0.211	0.452	0.231	0.249
0.194	0.203	0.213	0.388	0.237	0.249
$\mathscr{D}_8$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.135	0.141	0.130	0.166	0.242	0.248
0.126	0.132	0.123	0.154	0.244	0.248
0.119	0.129	0.114	0.150	0.230	0.247
0.125	0.132	0.119	0.153	0.241	0.248
0.122	0.127	0.119	0.150	0.243	0.248
$\mathscr{D}_9$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.069	0.069	0.065	0.074	0.231	0.240
0.065	0.065	0.063	0.070	0.241	0.240
0.060	0.063	0.058	0.068	0.232	0.241
0.067	0.066	0.062	0.071	0.238	0.240
0.065	0.062	0.061	0.067	0.237	0.240

Table 5.5: The reference values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ for the first five test data discretised by $\mathscr{D}_7, \mathscr{D}_8$, and $\mathscr{D}_9$. These results were obtained from the good Nnet-survival architecture in TCGA data.

Breast Cancer Data

In this section, we accompany $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ from the implementation results in Section 5.4.2 with $\text{ref}_1, \text{ref}_2$, and ref_3, respectively. We still use the same setting, Nnet-survival architecture, the train and test data used in the section.

$\widehat{\text{PC}}_n^1$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_1
0.021	0.021	0.021	0.026
0.02	0.02	0.02	0.025
0.02	0.019	0.02	0.025
0.02	0.02	0.02	0.025
0.02	0.02	0.02	0.025
$\widehat{\text{PC}}_n^2$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_2
0.021	0.021	0.021	0.026
0.021	0.021	0.021	0.025
0.02	0.02	0.02	0.025
0.02	0.02	0.02	0.025
0.027	0.02	0.02	0.025
$\widehat{\text{PC}}_n^{1,T_{max}}$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_3
0.249	0.246	0.247	0.247
0.246	0.247	0.247	0.247
0.246	0.243	0.246	0.246
0.253	0.252	0.247	0.247
0.256	0.247	0.247	0.247

Table 5.6: The reference values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ for the first five test data discretised by $\mathscr{D}_{10}$. These results were obtained from the implementation to breast cancer data.

The results for the first five test data discretised by $\mathscr{D}_{10}, \mathscr{D}_{11}$, and $\mathscr{D}_{12}$ are given in Table 5.6, Table C.3, and Table C.4, respectively. We can see from the tables that all values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ are almost the same as their respective reference values regardless of the used discretisation setup and λ. As a result, we are convinced that all fitted models have poor prediction performance regardless of the fitted models. In the next section, we will show the proof of the convergence of the proposed pair calibration estimators.

5.7 The Convergence Proofs

In this section, we only prove the convergence of $\widehat{\mathrm{PC}}_n^1$ given by Theorem 5.1.1. The convergences of $\widehat{\mathrm{PC}}_n^2$ (Theorem 5.1.2) and $\widehat{\mathrm{PC}}_n^{\tau_1,\tau_2}$ (Theorem 5.2.1) follow the results.

Proof of Theorem 5.1.1. For $i = 1, 2, \cdots$, we denote $\boldsymbol{x}_i = (\mathbb{Z}_i, T_i, D_i)$ so that $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ are independent samples from distribution F on $\mathscr{X}$. Recall the terms inside the summation in the numerator of $\widehat{\mathrm{PC}}_n^1$ and rewrite the terms as follows

$$\phi^g(\boldsymbol{x}_i, \boldsymbol{x}_j) = I_{\{T_i<D_i\}} I_{\{T_j<D_j\}} I_{\{T_i<T_{max}\}} \left(I_{\{T_i\le T_j\}} - \pi_{ij}\right)^2 g^{-1}(T_i) g^{-1}(T_j),$$

where $\pi_{ij} = \widehat{\mathbb{P}}(T_i \le T_j | T_i < T_{max})$, $g = \widehat{G}_n$ belongs to

$$\mathcal{G}_\epsilon = \{g_n : \{1, \cdots, T_{max} - 1\} \to [\epsilon, 1]\},$$

$g_n \xrightarrow{a.s.} G$ as $n \to \infty$, and ϕ^g belongs to a family $\varphi_\epsilon = \{\phi^g : g \in \mathcal{G}_\epsilon\}$.

Since $g^{-1} \in [1, \infty)$ and

$$0 \le I_{\{T_i<D_i\}} I_{\{T_i<D_j\}} I_{\{T_i<T_{max}\}} \left(I_{\{T_i<T_j\}} - \pi_{ij}\right)^2 \le 1,$$

then $\phi^g(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is bounded. Thus, the properties of ϕ^g are the same to h^g defined in (4.10). By adapting Lemma 4.3.2 and using the same arguments for obtaining (4.29), we obtain

$$\frac{2}{n(n-1)} \sum_{1\le i<j\le n} \bar{\phi}^{\widehat{G}_n}(\boldsymbol{x}_i, \boldsymbol{x}_j) \xrightarrow{a.s.} F \otimes F(\bar{\phi}^G) \tag{5.31}$$

as $n \to \infty$.

We now recall the denominator of $\widehat{\mathrm{PC}}_n^1$ and rewrite the terms as follows

$$\frac{1}{n} \sum_{i}^{n} \left[I_{\{T_i<T_{max}\}} I_{\{T_i<D_i\}} g^{-1}(T_i)\right], \tag{5.32}$$

where $g \in \mathcal{G}_\epsilon = \{g_n : \{1, \cdots, T_{max} - 1\} \to [\epsilon, 1]\}$, and $g_n \xrightarrow{a.s.} G$ as $n \to \infty$. By the regularity conditions R1-R4 and Kolmogorov strong law of large numbers (Loève & Loève, 1977), we get

$$\frac{1}{n} \sum_{i=1}^{n} \left[I_{\{T_i<T_{max}\}} I_{\{T_i<D_i\}} g^{-1}(T_i)\right] \xrightarrow{a.s.} \mathbb{E}\left[I_{\{T_i<T_{max}\}} \frac{I_{\{T_i<D_i\}}}{G(T_i)}\right]. \tag{5.33}$$

Applying the Continuous Mapping Theorem (Shao, 2003), we have

$$\frac{\frac{1}{n(n-1)}\sum_{1\leq i<j\leq n}\phi^g(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\frac{1}{n}\sum_{i=1}^{n}\left[I_{\{T_i<T_{max}\}}I_{\{T_i<D_i\}}g^{-1}(T_i)\right]}\xrightarrow{a.s.}\frac{F\otimes F(\phi^G)}{\mathbb{E}\left[I_{\{T_i<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{G(T_i)}\right]}$$

where the right-hand side equals

$$\frac{\mathbb{E}\left[\left(I_{\{T_i\leq T_j\}}-\widehat{\mathbb{P}}\left[T_i\leq T_j\middle|T_i<T_{max}\right]\right)^2\frac{I_{\{T_i<D_i\}}}{G(T_i)}\frac{I_{\{T_j<D_j\}}}{G(T_j)}I_{\{T_i<T_{max}\}}\right]}{\mathbb{E}\left[I_{\{T_i<T_{max}\}}\frac{I_{\{T_i<D_i\}}}{G(T_i)}\right]}.$$

In conclusion, we have shown that

$$\widehat{\mathrm{PC}}_n^1\xrightarrow{a.s.}\mathrm{PC}$$

as $n\to\infty$. □

In this chapter, we have proposed a new measure called pair calibration, which is a calibration measure of a discrimination quantity. Several estimators of pair calibration and truncated pair calibration were also introduced in this chapter. However, we also have found some pitfalls of the estimators, resulting in difficulties in evaluating the model's predictive performance in some cases. Those pitfalls have motivated us to propose two types of reference values: (1) reference values based on the worst prediction, and (2) reference values based on the outcomes proportion. We found that the first reference value was impractical. On the other hand, the second reference value is useful for the pair calibration. At the end of this chapter, we also proved the convergence of the pair calibration estimators. In the next chapter, we will show how pair calibration behaves along with the other proposed measures when evaluating the models' performance simultaneously.

Chapter 6

Comprehensive Evaluation of Proposed Measures

In Chapter 3 up to Chapter 5, we have proposed several novel measures, such as the modified integrated Brier scores, time-dependent Uno's C-index, and pair calibration. While there is no doubt that time-dependent Uno's C-index is a discrimination measure, we categorise normalised integrated Brier score, centered integrated Brier score, and normalised centered integrated Brier score as calibration measures. Although the integrated Brier score and pair calibration are calibration measures, they calibrate the discrimination quantities. Hence, the integrated Brier score and pair calibration may be affected by model's discrimination ability as we will study in some simulation studies of this chapter.

In the previous chapters, we also have applied the proposed measures through simulation studies and real-world examples depending on the objectives we investigated in each topic. However, the numerical experiments were conducted separately in assessing the performance of the fitted models. This section, therefore, will apply the proposed measures, such as $\widehat{\text{IBS}}_n, \widehat{\text{CIBS}}_n, \widehat{\text{NCIBS}}_n^{\varepsilon}, \widehat{\text{C}}_n^{w}, \widehat{\text{PC}}_n^{1}, \widehat{\text{PC}}_n^{2}$ and $\widehat{\text{PC}}_n^{1,T_{max}}$, in evaluating the performance of a single model altogether, resulting in a more comprehensive understanding of the measures' behaviour. We will only report $\widehat{\text{NCIBS}}_n^{\varepsilon}$ because we can automatically obtain $\widehat{\text{NIBS}}_n^{\varepsilon}$ once we have $\widehat{\text{NCIBS}}_n^{\varepsilon}$. We will report only their differences to the respective reference values for $\widehat{\text{IBS}}_n, \widehat{\text{PC}}_n^{1}, \widehat{\text{PC}}_n^{2}$ and $\widehat{\text{PC}}_n^{1,T_{max}}$. We also will use $\widehat{\text{NCIBS}}_n^{\varepsilon}$ instead of its original measure $\widehat{\text{NCIBS}}_n$ to prevent the measure from incomputable values. In all

implementation settings applied in this chapter, we will choose 0.01 for the value of the ε.

6.1 Simulation 6: PH Data with Varying Censoring Rates

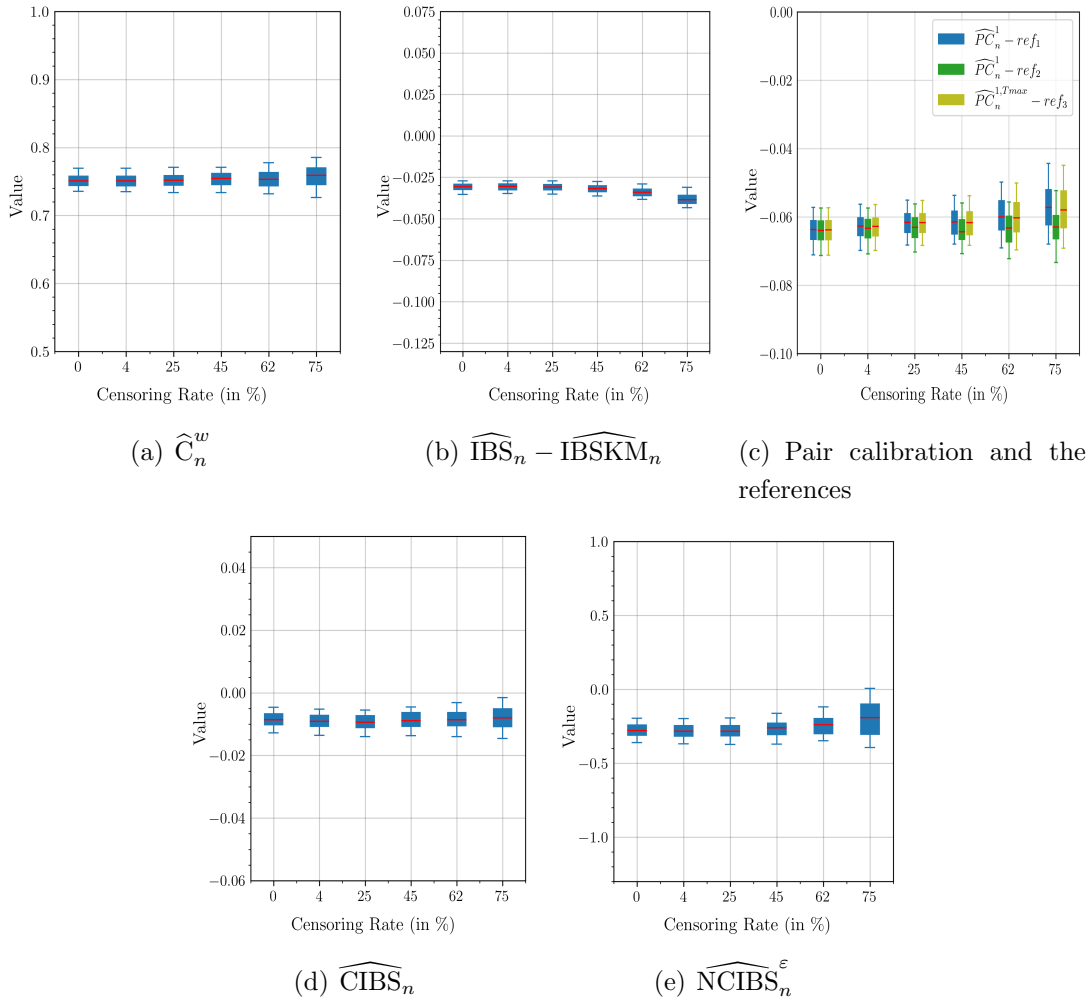
This simulation study aims to investigate thoroughly the behaviour of the proposed measures when evaluating a fixed model's performance from almost fully uncensored train data, but varying censoring rate in test data. This objective is the same as in Section 4.5.1 (Chapter 4) except for the applied measures. We used only one type of data, namely the PH data generated in Simulation 1 (Section 3.1.1 of Chapter 3). We chose the PH data for this simulation mainly because we have shown in Chapter 4 and Chapter 5 that time-dependent Uno's C-index and pair calibration were unbiased as censoring rates increased in the test data regardless of PH assumption in data. A train data ($n_{\text{train}} = 1000$) and 100 independent test data ($n_{\text{test}} = 1000$) were discretised by $\mathscr{D}_{14}$={0, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, ∞} as defined in Section 4.5.1 (Chapter 4). The censoring rate in the train data was very close to 0%, meanwhile for the test data, the used censoring rates were from almost 0% to about 4%, 25%, 45%, 62%, and 75%. We fitted the Sim.3: PH Nnet-survival architecture given in Table B.5 (Appendix B.1). Finally, we evaluated the model performance on the 100 independent test data using $\widehat{\text{IBS}}_n, \widehat{\text{CIBS}}_n, \widehat{\text{NCIBS}}_n^{\varepsilon}, \widehat{\text{C}}_n^{w}, \widehat{\text{PC}}_n^{1}, \widehat{\text{PC}}_n^{2}$ and $\widehat{\text{PC}}_n^{1,T_{max}}$.

The main results of this numerical implementation can be seen in Figure 6.1. For all computed measures, we assume that their "true" values are obtained when the censoring rates in the train and test data are 0%. From Figure 6.1, the proposed measures are generally unbiased since their change as the censoring rates increase is minimal. Panels (b), (d), and (e) show the results for $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n), \widehat{\text{CIBS}}_n$, and $\widehat{\text{NCIBS}}_n^{\varepsilon}$, respectively. Since all individuals in the sample always contribute to the measures, the sum of all squared errors $\left(I_{\{T_i>t\}} - S(t;\mathbb{Z}_i)\right)^2$ in their numerators will be proportional to the number of individuals in the denominator regardless of the number of censored individuals. As a result, $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n), \widehat{\text{CIBS}}_n$, and $\widehat{\text{NCIBS}}_n^{\varepsilon}$ do not significantly change

when censoring rates vary. The large standard deviations in the larger censoring rates are mainly due to the misspecification of $\widehat{G}_n$. The large standard deviation of $\widehat{\text{NCIBS}}_n^{\varepsilon}$ when the censoring rate is 75% is because of the high variabilities of (3.13).

Panels (a) and (c) of Figure 6.1 display the results for $\widehat{\text{C}}_n^w$ and the differences of the pair calibration estimators with their reference values, respectively. Although they are generally unbiased as the censoring rate increases, the standard deviations become larger in large censoring rates, i.e. 62% and 75%. The reason for those deviations is different to the other proposed measures. While large standard deviations in $\widehat{\text{NCIBS}}_n^{\varepsilon}$ are mainly due to the variability of small values of (3.13), in $\widehat{\text{C}}_n^w, \widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ they are due to the higher number of ignored pairs when censoring rates are significant. In the pair calibration estimators, for example, pairs of two different individuals ($i \neq j$) will not have any contributions to their respective measures when T_i and T_j are both censored. We need higher values of weights to penalise that loss, and we have to include many more errors due to misspecification of $\widehat{G}_n$. As a result, we will see significant bias when the censoring rates in the test data are large. When censoring rates are small, there is no significant amongst $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$. However, when the censoring rate is much higher, $\widehat{\text{PC}}_n^1$ and $\widehat{\text{PC}}_n^2$ are preferable because they account for the effect of the censored individuals.

In summary, in this section, we have demonstrated the unbiasedness of the proposed measures. In particular, they are generally unbiased as censoring rates in the test data increase. We have also discussed why the biases in some of the measures are larger when the test data contain severely censored individuals. In the next section, we will study the behaviour of our proposed measures when assessing the performance of different models.



(a) $\widehat{\mathrm{C}}_n^w$ (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$ (c) Pair calibration and the references

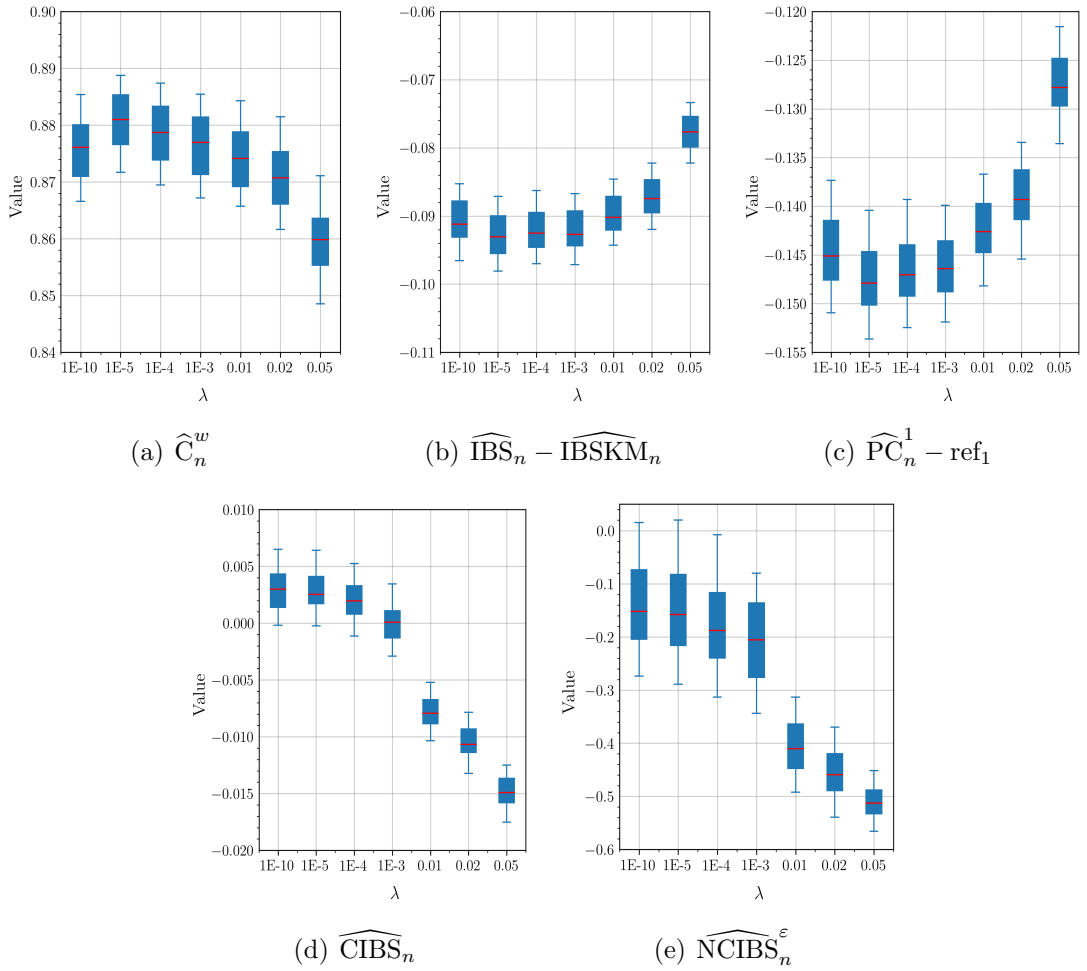
(d) $\widehat{\mathrm{CIBS}}_n$ (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$

Figure 6.1: Boxplots of (a) $\widehat{\mathrm{C}}_n^w$, (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$, (c) pair calibration and the references, (d) $\widehat{\mathrm{CIBS}}_n$, and (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ over various censoring rates. They were computed from 100 test data in Simulation 6.

6.2 Simulation 7: Non-PH Data with Varying L2-Regularisation

This simulation study investigates how the proposed measures behave in evaluating the performance of different models. To obtain the models, we varied the value of L^2-regularisation ($\lambda \geq 0$) in the fitted Nnet-survival architecture while the other hyper-parameters were fixed. The setting of this study was the same as the real-world example in breast cancer data discussed in Section 3.1.2 (Chapter 3) except for the applied measures. We chose the non-PH generated in Simulation 2 (see Section 4.1 in Chapter 4) as the used data for this simulation study. We employed almost fully uncensored train and test data to minimise the misspecification effect of $\widehat{G}_n$. We fitted Nnet-survival architecture given in the left column of Table B.4 (Appendix B.2) to a fixed train data (n_{train}=1000). Then, the model's performance was evaluated on 100 independent test data (n_{test}=1000) using $\widehat{\text{IBS}}_n, \widehat{\text{CIBS}}_n, \widehat{\text{NCIBS}}_n^{\varepsilon}, \widehat{\text{C}}_n^{w}$, and $\widehat{\text{PC}}_n^{1}$. Note that we do not implement $\widehat{\text{PC}}_n^{2}$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ since they are the same as $\widehat{\text{PC}}_n^{1}$ when the censoring rate in the test data is 0%.

Figure 6.2 displays the simulation results in five panels. From panels (a)-(c), $\lambda = 1E-5$ is the optimum L^2-regularisation for $\widehat{\text{C}}_n^{w}, \widehat{\text{IBS}}_n$, and $\widehat{\text{PC}}_n^{1}$. Meanwhile, $\widehat{\text{CIBS}}_n$ and $\widehat{\text{NCIBS}}_n^{\varepsilon}$ choose $\lambda = 1E-3$ and $\lambda = 1E-10$ as the best value for L^2-regularisation, respectively. These results show that $\widehat{\text{C}}_n^{w}$ differs from $\widehat{\text{CIBS}}_n$ and $\widehat{\text{NCIBS}}_n^{\varepsilon}$ in tuning the optimum L^2-regularisation. In this case, we can also see that $\widehat{\text{IBS}}_n$ and $\widehat{\text{PC}}_n^{1}$ more or less align with $\widehat{\text{C}}_n^{w}$ in choosing the best L^2-regularisation for our model. We can also see that the differences amongst λ around the optimum λ for $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ and $\widehat{\text{CIBS}}_n$ are very close, resulting in difficulties in choosing the optimum one.



(a) $\widehat{\mathrm{C}}_n^w$ (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$ (c) $\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$

(d) $\widehat{\mathrm{CIBS}}_n$ (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$

Figure 6.2: Boxplots of (a) $\widehat{\mathrm{C}}_n^w$, (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$, (c) $\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$, (d) $\widehat{\mathrm{CIBS}}_n$, and (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ over various L^2-regularisation. They were computed from 100 test data in Simulation 7.

To sum up, in tuning neural networks hyper-parameters, we need to carefully choose the type of measures to use depending on our study objective. Our numerical experiments show that calibration and discrimination measures differ in choosing the optimum L^2-regularisation for our Nnet-survival architecture. If our study aims to obtain a model with good discrimination, we should use $\widehat{\mathrm{C}}_n^w$. Otherwise, $\widehat{\mathrm{CIBS}}_n$ and $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ should be used if we need a model with good calibration ability. In the subsequent section, we will investigate calibration and discrimination in more comprehensive experiments.

6.3 Further Insights into Calibration and Discrimination

In this section, we will provide two examples of models with different discrimination and calibration capabilities. We will also investigate how the proposed measures behave in such situations.

Since we will only use train and test data with almost 0% censoring rates, we do not compute $\widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ and will report $\widehat{\text{PC}}_n^1$ only. The two cases will be discussed: (1) good discrimination but poor calibration, and (2) good calibration but poor discrimination. Our decision to choose one of the calibration and discrimination measures depends on the research objectives. For instance, Burke *et al.* (1997) in their paper only reported a discrimination measure, namely the ROC curve, since their study objective was to differentiate the risks amongst patients with different characteristics. Most studies on survival predictions using deep learning approaches have focused on the model's discrimination capability (Kamran & Wiens, 2021). However, having survival models with good calibration will improve our confidence in our prediction. In addition, it can make our survival models more beneficial in real-world implementations (Gneiting & Katzfuss, 2014).

6.3.1 Example Case 1: Good Discrimination but Poor Calibration

A well-defined threshold is not available in current literature for distinguishing the quality of discrimination measures, e.g. C-index. However, most researchers agree that when C-index is equal to 0.5, then the model's discrimination capability is no more than random guessing. Since the C-index estimates the population concordance probability, then 0.5 is only the same as the chance of being head or tail in flipping a coin for prediction task. Hence, the higher C-index-based measures, such as $\widehat{\text{C}}_n^w$, than 0.5, the better the model performance. Of course, the worst model performance happens when $\widehat{\text{C}}_n^w$ is less than 0.5. Meanwhile, calibration measures are more straightforward than discrimination measures. A well-calibrated model is achieved when the model's output is very close to the

ground truth as the model's outcome. Our proposed calibration measures, namely $\widehat{\text{CIBS}}_n$ and $\widehat{\text{NCIBS}}_n^{\varepsilon}$, are computed based on the squared difference for each individual between the model's outcomes $I_{\{T_i>t\}}$ as the ground truth and the model's outputs $S(t;\mathbb{Z}_i)$.

This section will present a case where our model performance has good discrimination but poor calibration. To accomplish this objective, we provide two simulation scenarios: (1) multiplying $S(t;\mathbb{Z}_i)$ for all $t \in \{1,\cdots,T_{max}\}$ by constant c, and (2) shifting $S(t;\mathbb{Z}_i)$ for some $t \in \{1,\cdots,T_{max}\}$. In real-world applications, the applied setting in the first scenario may relate to the case when some individuals in the train data experience the event of interest more often than usual. As a consequence, the fitted models will be much easier in distinguishing individuals in the data, but the models will not provide the accurate predicted survival curves. It can be said also that our train data is not a sample from the original population, but it is from a part of the population that is more risky to experience the event of interest.

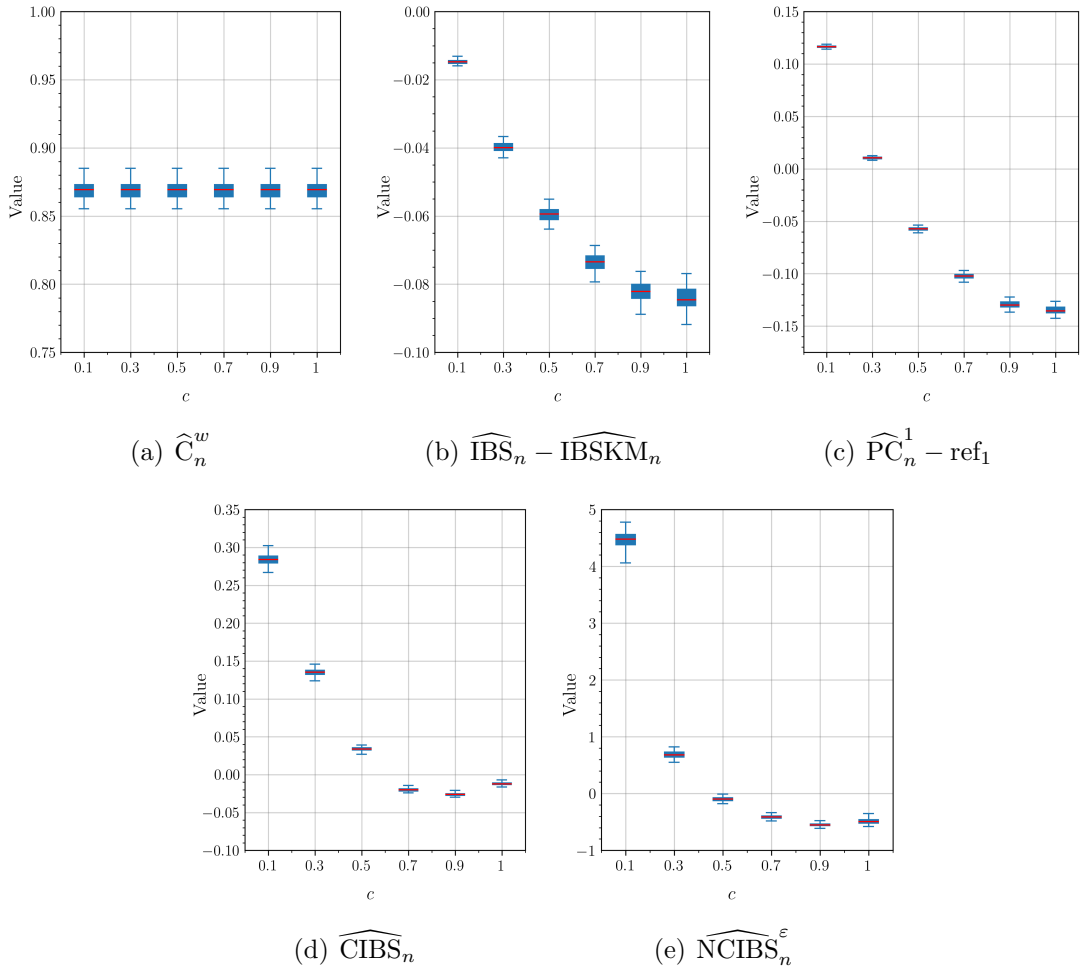
The second scenario will correspond to the case when we mismatch the assessment time for some individuals in our sample. For instance, suppose we are interested in returning time to the hospital after surgery for cancer patients. We have recorded some patients' event times since the surgery time. However, we can only record the other patients a month after their surgery. Unfortunately, when we evaluate the model performance on the data, we unwittingly treat those two different groups of individuals the same and mix them. In particular, if the train data does not contain the mixed individuals while the test data suffers such a mismatch, we will see poor accuracy in our prediction. This human error can happen in practice, resulting in the risk order shifting of some individuals by a month earlier. This risk shifting may not highly affect the risk order of the individuals, but it will affect the accuracy of our prediction.

Scenario 1: Multiplying The Model's Outputs by Constants

The data used in this scenario are the non-PH data obtained from Simulation 2 in Section 4.1 (Chapter 4). In particular, we only used the train data (n_{train}=1000)

and 100 independent test data (n_{test}=1000) with almost 0% censoring rates generated in that simulation. The train and test data were discretised by the same discretisation setup $\mathscr{D}_{13}$={0, 5, 10, 15, 20, 25, 30, 40, 80, 90, 150, ∞} as introduced in Chapter 4. After we fitted the non-PH data architecture in the third column of Table B.5 (Appendix B.2), the models' performance was evaluated using $\widehat{\text{IBS}}_n$, $\widehat{\text{CIBS}}_n$, $\widehat{\text{NCIBS}}_n^{\varepsilon}$, $\widehat{\text{C}}_n^{w}$, and $\widehat{\text{PC}}_n^{1}$ on 100 independent test data. Next, we spoiled the prediction accuracy by multiplying the predicted survival curve values $S(t;\mathbb{Z}_i)$ of each $i(i = 1, \ldots, n)$ with some constants $c \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$ for $t \in \{1, \cdots, T_{max} - 1\}$. By applying this scenario, we forced all individuals to experience the event of interest earlier than their actual event times, depending on the used c. Note that $c = 1$ provides the original values of the predicted survival curves. Indeed, in the real-world applications, this setting is not a proper approach because we will underestimate the survival curves of each individual. However, it may happen sometimes as we discussed previously, resulting in poor accuracy of the predicted survival curves.

The main results of this study can be seen in Figure 6.3. We plot the boxplots of the measures into five panels. From panel (a), we can see that $\widehat{\text{C}}_n^{w}$ is almost unaffected by the values of c. It shows an excellent calibration capability (i.e. around 0.87) for each c. On the other hand, the calibration measures, such as $\widehat{\text{CIBS}}_n$ and $\widehat{\text{NCIBS}}_n^{\varepsilon}$, are sensitive to the change in prediction accuracy, as shown in panel (d) and panel (e). The smaller the value of c, the worse the models' calibration capability. The same behaviour also happens to $\widehat{\text{IBS}}_n$ and $\widehat{\text{PC}}_n^{1}$ (see panel (b) and (c)), where the models' performance worsens as c decreases. These results show that $\widehat{\text{IBS}}_n$ and $\widehat{\text{PC}}_n^{1}$ may also affected by the prediction accuracy the same as in the calibration measures, such as $\widehat{\text{CIBS}}_n$ and $\widehat{\text{NCIBS}}_n^{\varepsilon}$. The more negative the values of $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$ and $(\widehat{\text{PC}}_n^{1} - \text{ref}_1)$, the better the model performance. In this case, $\widehat{\text{IBS}}_n$ and $\widehat{\text{PC}}_n^{1}$ align with $\widehat{\text{CIBS}}_n$ and $\widehat{\text{NCIBS}}_n^{\varepsilon}$.



(a) $\widehat{\mathrm{C}}_n^w$ (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$ (c) $\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$

(d) $\widehat{\mathrm{CIBS}}_n$ (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$

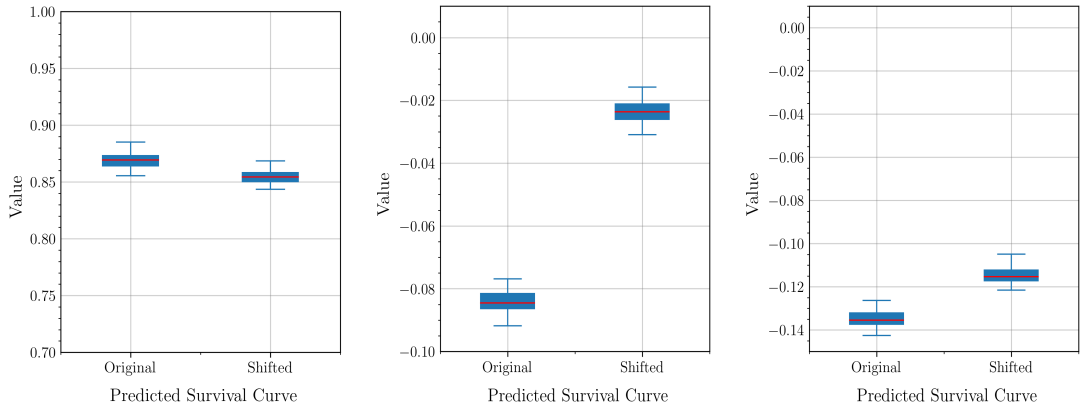
Figure 6.3: Boxplots of (a) $\widehat{\mathrm{C}}_n^w$, (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$, (c) $\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$, (d) $\widehat{\mathrm{CIBS}}_n$, and (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ computed from 100 test data in Scenario 1 of Example Case 1.

In this section, we have demonstrated that the proposed discrimination measure, namely time-dependent Uno's C-index, is not highly affected by prediction accuracy as long as the risk order in the usable pairs does not significantly change. On the other hand, calibration measures are highly dependent on prediction accuracy. Since the integrated Brier score and pair calibration are not pure discrimination measures, they may also be sensitive to prediction error, especially when the prediction accuracy is terrible, as shown in this simulation scenario. The following section will present our second approach to obtaining good discrimination but poor calibration.

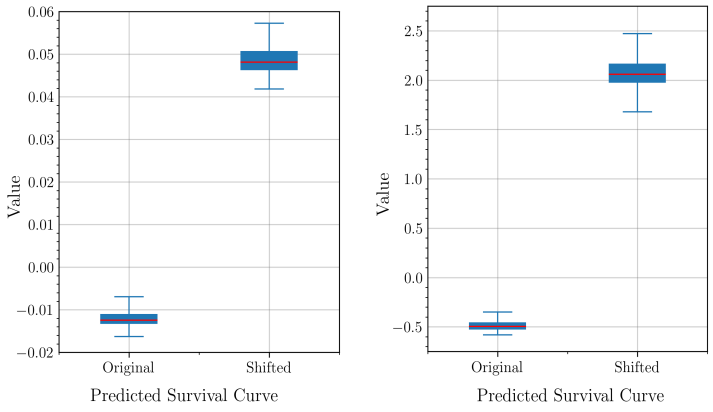
Scenario 2: Shifting The Model's Outputs

In this section, we will demonstrate the second scenario for the case of good discrimination but poor calibration. We used the same non-PH data with 0% censoring rates in the first scenario (see Section 6.3.1), where the train and test data were discretised by $\mathscr{D}_{13}$. Moreover, we fitted the non-PH data architecture in the third column of Table B.5 (Appendix B.2) to the train data (n_{train}=1000), and then evaluated the model's performance on the 100 independent test data (n_{test}=1000) using $\widehat{\text{C}}_n^w$, $\widehat{\text{IBS}}_n$, $\widehat{\text{CIBS}}_n$, $\widehat{\text{NCIBS}}_n^{\varepsilon}$ and $\widehat{\text{PC}}_n^1$. Then, we replaced $S(t;\mathbb{Z}_i)$ by $S(t+2;\mathbb{Z}_i)$, for $t=1$ up to $t=T_{max}-6$. For $t=T_{max}-1$ and $t=T_{max}-2$, $S(t;\mathbb{Z}_i)$ were replaced by $S(t+1;\mathbb{Z}_i)$. We kept the original outputs $S(t;\mathbb{Z}_i)$ for the rest of the periods. By applying this setting, we spoiled the predicted survival curves for several periods, but we kept the order of the outputs for most usable pairs ($i \neq j$), $S(T_i;\mathbb{Z}_i) < S(T_i;\mathbb{Z}_i)$, fixed.

The main results of this simulation can be seen in Figure 6.4. As we can see from panel (a) of the figure, $\widehat{\text{C}}_n^w$ almost does not change after some $S(t;\mathbb{Z}_i)$ are shifted. This is because our setting minimises the concordance change of each usable pair in $\widehat{\text{C}}_n^w$. Since we did not completely change $S(t;\mathbb{Z}_i)$ for all periods $t \in \{1,\cdots,T_{max}\}$, there are still lots of $S(t;\mathbb{Z}_i)$ that are not changed. Thus, the concordance of usable pairs in $\widehat{\text{C}}_n^w$ does not significantly change after the shifting. On the other hand, the calibration measures, namely $\widehat{\text{CIBS}}_n$ and $\widehat{\text{NCIBS}}_n^{\varepsilon}$, dramatically change due to the shifting of some $S(t;\mathbb{Z}_i)$ (see panels (d) and (e)). The change of $\widehat{\text{CIBS}}_n$ increases to around 0.05 in the shifted outputs from 0.01 in the original outputs, meaning that the $\widehat{\text{CIBS}}_n$ in the shifted version is four times larger than in the original version. Although the increase is five times larger than the original, 0.05 is still categorised as close to zero and shows good model performance. However, $\widehat{\text{NCIBS}}_n^{\varepsilon}$ significantly increases from around -0.5 to around 2, so the model performance from the shifting is really bad. This result also shows us an advantage of using $\widehat{\text{NCIBS}}_n^{\varepsilon}$.



(a) $\widehat{\mathrm{C}}_n^w$ (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$ (c) $\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$



(d) $\widehat{\mathrm{CIBS}}_n$ (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$

Figure 6.4: Boxplots of (a) $\widehat{\mathrm{C}}_n^w$, (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$, (c) $\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$, (d) $\widehat{\mathrm{CIBS}}_n$, and (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ in the original and the shifted $S(t; \mathbb{Z}_i)$ for some $t \in \{1, \cdots, T_{max}\}$. They were computed from 100 test data in Scenario 2 of Example Case 1.

Panel (b) of Figure 6.4 shows that based on $(\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n)$, the model performance changes a lot after the shifting. The model performance is initially good in the original outputs, but the model performance is poor after the shifting since $(\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n)$ are close to zero. Based on $(\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1)$, the model performance has changed after the shifting, as shown in panel (c). In this case, $\widehat{\mathrm{IBS}}_n$ and $\widehat{\mathrm{PC}}_n^1$ align with calibration measures. In other words, the calibration parts in $\widehat{\mathrm{IBS}}_n$ and $\widehat{\mathrm{PC}}_n^1$ are much stronger than their discrimination parts.

In this section, we have shown one of the unique cases in performance evaluation tasks: good discrimination but poor calibration. We proposed two simulation scenarios to achieve such situations. In the next section, we will move on to another case where we have good calibration but poor discrimination.

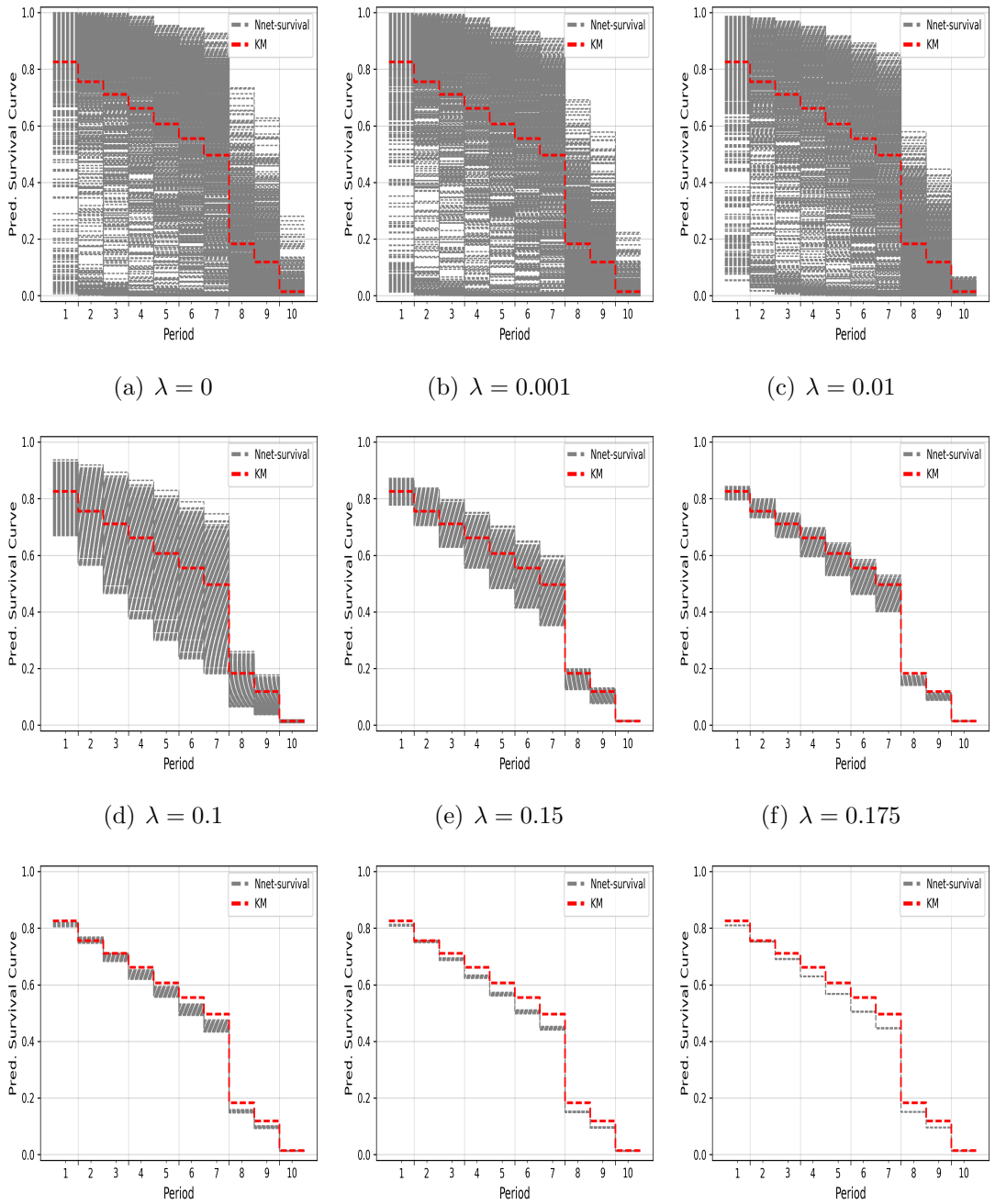
6.3.2 Example Case 2: Good Calibration but Poor Discrimination

This section will present our second example case where we have good calibration but poor discrimination. We used the PH data obtained from Simulation 1 in Section 3.1.1 of Chapter 3. We moreover only employed the train data (n_{train}=1000) and 100 independent test data (n_{test}=1000) with almost zero censoring rates, where we discretised them by $\mathscr{D}_{14}$. We fitted the Nnet-survival architecture in the third column of Table B.6 (Appendix B.2). Then, we computed $\widehat{\text{C}}_n^w$, $(\widehat{\text{IBS}}_n - \widehat{\text{IBSKM}}_n)$, $\widehat{\text{CIBS}}_n$, $\widehat{\text{NCIBS}}_n^{\varepsilon}$, and $(\widehat{\text{PC}}_n^1 - \text{ref}_1)$ from the 100 test data.

The hyper-parameters in Table B.6 (Appendix B.2) were fine-tuned such that we have good calibration but poor discrimination. Since $\widehat{\text{C}}_n^w$ depends on the variability of the predicted survival curves at each t, we can control the values of $\widehat{\text{C}}_n^w$ by varying L^2-regularisation (λ). When we have small values of λ while the number of nodes in each hidden layer is large, the complexity of the fitted models will be high. As a result, we will have higher variability of the predicted survival curves at each period t and hence may have higher values of $\widehat{\text{C}}_n^w$. On the other hand, the more complex architecture usually increases the accuracy of the prediction. Hence, simultaneously, we expect the calibration to be good using the hyper-parameters in Table B.6. For example, the results from a test data when we tuned the value of λ so that we obtained the worst value $\widehat{\text{C}}_n^w$ can be seen in Figure 6.5 and Table 6.1. In Figure 6.5, we have nine panels (a)-(i), where each panel displays the predicted survival curves for all individuals $i (i = 1, \cdots, 1000)$ obtained from Nnet-survival (grey colour) and KM estimator (red colour) with a specific value of λ. As we can see from Figure 6.5, the variability of the predicted survival curves is high when we have minimal values of λ, i.e. $\lambda = 0, 0.001, 0.01$, as shown in panels (a)-(c). Then, the variability decreases as λ increases, resulting

in the predicted survival curves for each individual being almost uniform and very close to the KM estimator survival curve (see panels (d)-(i)). The larger L^2-regularisation applied to the Nnet-survival, the lesser the effects of individuals' covariates on the model. Therefore, the number of tied pairs, $(i \neq j)$ with $S(T_i; \mathbb{Z}_i) = S(T_i; \mathbb{Z}_j)$, increases when λ is large. Since our proposed $\widehat{\mathrm{C}}_n^w$ ignores such tied pairs completely, $\widehat{\mathrm{C}}_n^w$ will decrease closer to 0.5. These results can be seen clearly from Table 6.1, where $\widehat{\mathrm{C}}_n^w$ peaks at $\lambda = 0.01$, and $\lambda = 0.22$ gives the lowest value of $\widehat{\mathrm{C}}_n^w$. Note that when λ slightly rises to higher than 0.22, $\widehat{\mathrm{C}}_n^w$ almost does not change. Furthermore, when $\lambda = 0.24$, we cannot compute $\widehat{\mathrm{C}}_n^w$ any more since all pairs are tied so that no more usable pairs are left. In this case, $\lambda = 0.22$ is suitable for our simulation objective. From Table 6.1, we have an example case for which we have good calibration but poor discrimination at $\lambda = 0.22$. In addition, $\widehat{\mathrm{IBS}}_n$ and $\widehat{\mathrm{PC}}_n^1$, have poor prediction performance since their differences with the respective reference values are very close to zero. In other words, they align with the discrimination measure, $\widehat{\mathrm{C}}_n^w$.

In this section, we have shown a simulation study showing good calibration but poor discrimination. In the subsequent section, we will apply the proposed measures to the performance evaluation of one of the most common non-linear survival models, i.e. random survival forests (Ishwaran *et al.*, 2008).



(a) $\lambda = 0$ (b) $\lambda = 0.001$ (c) $\lambda = 0.01$

(d) $\lambda = 0.1$ (e) $\lambda = 0.15$ (f) $\lambda = 0.175$

(g) $\lambda = 0.2$ (h) $\lambda = 0.21$ (i) $\lambda = 0.22$

Figure 6.5: Boxplots of $S(t; \mathbb{Z}_i)(i = 1, \cdots, 1000)$ for each $t \in \{1, \cdots, T_{max} - 1\}$ estimated from a test data using Nnet-survival and KM estimator over various λ in Example Case 2.

λ	$\widehat{\mathrm{C}}_n^w$	$\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$	$\widehat{\mathrm{CIBS}}_n$	$\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$	$\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$
0	0.746	-0.0511	0.0066	0.0185	-0.06636
0.001	0.7477	-0.0512	0.0047	-0.0614	-0.06649
0.01	0.7571	-0.0530	-0.0038	-0.1797	-0.06915
0.1	0.7429	-0.0298	-0.0193	-0.1638	-0.04168
0.15	0.7052	-0.0114	-0.0087	-0.0578	-0.01677
0.175	0.6709	-0.0048	-0.0036	-0.0137	-0.00777
0.2	0.6255	-0.0008	-0.0002	0.0154	-0.00231
0.21	0.6185	0.0006	0.0009	0.025	-0.00065
0.22	**0.6178**	**0.0009**	**0.0013**	**0.0283**	**-0.00004**
0.24	-	0.0011	0.0014	0.0289	0.00001

*) If there is no number in the shell, the estimator cannot be computed.

Table 6.1: $(\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n)$, $\widehat{\mathrm{CIBS}}_n$, $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$, $\widehat{\mathrm{C}}_n^w$ and $\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$ from a test data over various λ in Example Case 2.

6.4 Application of Proposed Measures to Random Survival Forests

Almost all the numerical experiments discussed in this thesis evaluate the model performance of Nnet-survival. Although our proposed measures can be used for any non-linear survival models with a single event of interest and right-censored data, for completeness, this section will show the application of the proposed measures in evaluating the model performance of one of the most applied machine learning approaches, namely random survival forests (Ishwaran *et al.*, 2008).

The random forests method was introduced by Breiman (2001) and has been applied in vast research fields for prediction and classification tasks. Ishwaran *et al.* (2008) proposed random survival forests as an extension of the random forests to handle right-censored survival data. Random survival forests was then adapted by Schmid *et al.* (2020) into the context of discrete-time units, where they called their approach as discrete-time survival forests (DTSF). In this work, we aim to investigate how our proposed measures behave in evaluating the model predictive performance of DTSF. In particular, we will assess the performance

of DTSF when one of its hyper-parameters, namely minimum node size, varies while the other hyper-parameters are fixed.

In this implementation, we used the 100 pairs of train and test data obtained from the TCGA data in Section 3.1.2 (Chapter 3). The train and test data were discretised by $\mathscr{D}_8 = \{0,\ 74,\ 152.003,\ 234.465,\ 321.928,\ 415.039,\ 514.573,\ 621.488,\ 736.966,\ 862.497,\ 1000,\ \infty\}$ as defined in Section 3.1.2. As a result, the data contain about 77.01% censored individuals. We first fitted the DTSF architecture in Table B.7 (Appendix B.3) to each train data using R package **ranger** (Wright & Ziegler, 2017). We mostly used the default hyper-parameters in the package. However, the number of trees was 100, and the splitting rule was Hellinger following Schmid *et al.* (2020). Moreover, the minimum node size was varied, i.e. 3, 10, 25, 75, 150, 250, 500, 750, and 1000. Then, for each minimum node size, we evaluated the model performance from their respective test data using $\widehat{\mathrm{C}}_n^w, \widehat{\mathrm{IBS}}_n, \widehat{\mathrm{CIBS}}_n, \widehat{\mathrm{NCIBS}}_n^{\varepsilon}, \widehat{\mathrm{PC}}_n^1, \widehat{\mathrm{PC}}_n^2$ and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$.

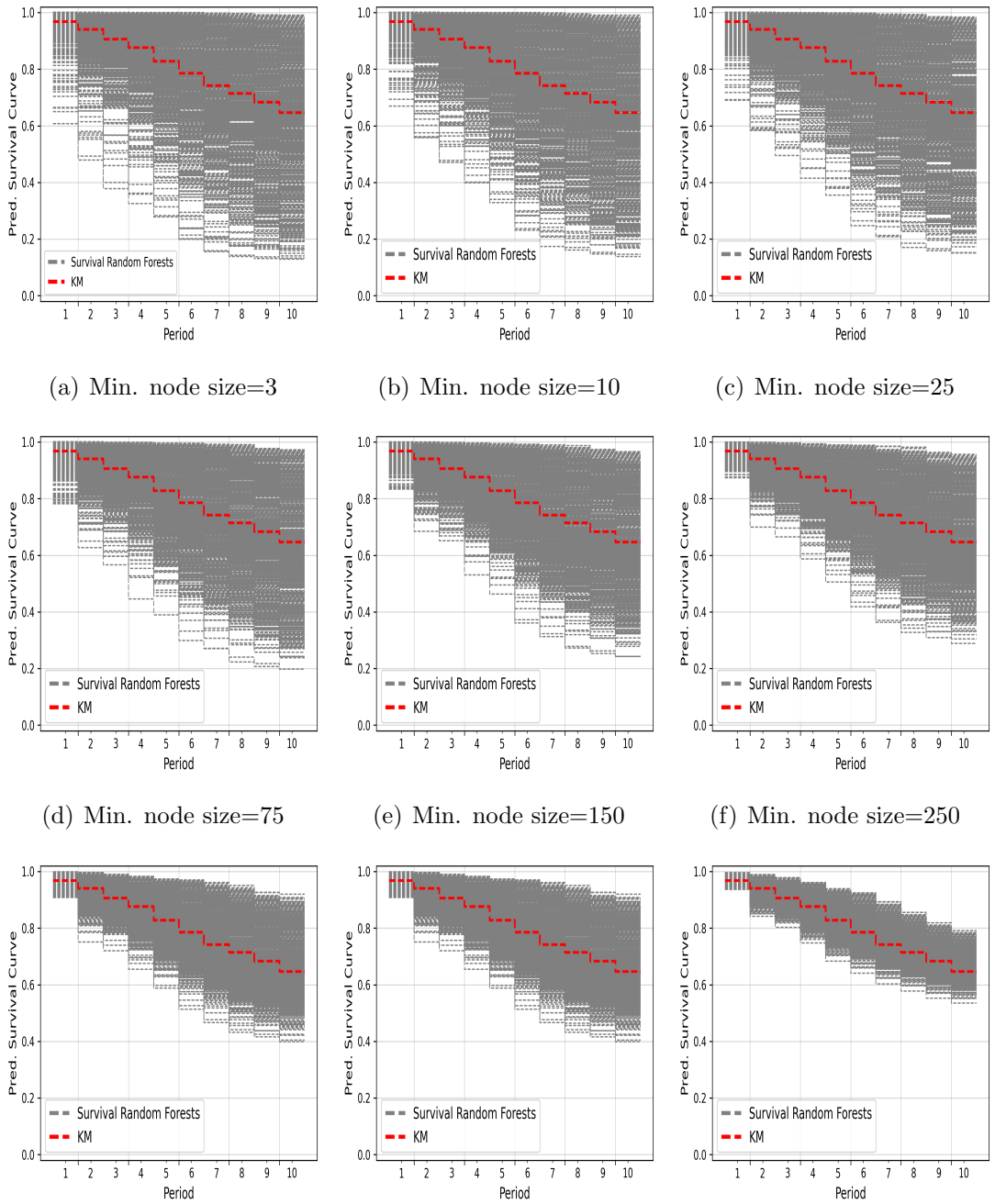
Figure 6.6 shows the predicted survival curves for all periods up to $T_{max}-1 = 10$ from the first test data, and the model was fitted to the first train data. They are obtained from DTSF (grey colour lines) and KM estimator (red colour line). As we can see from Figure 6.6, from smaller minimum node sizes, such as 3-75, the variabilities of the predicted survival curves for each period are high (see panels (a)-(d)). However, starting from minimum node size equals 150, the variabilities get smaller, resulting in the predicted survival curves being closer to the KM estimators (see panels (e)-(i)). This behaviour is the same as in Simulation 7 (Section 6.2) when we varied the L^2-regularisation in Nnet-survival. As a result, for the same reasons as in Simulation 7, $\widehat{\mathrm{C}}_n^w$ firstly increases to reach its peak at around 0.75 when the minimum node size equals 75 but then drops to around 0.71 when the minimum node size equals 1000 (panel (a) of Figure 6.7). From panel (b), $(\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n)$ select 25 as the best value for minimum node size. $\widehat{\mathrm{PC}}_n^1$ and $\widehat{\mathrm{PC}}_n^2$ consistently choose 10 as their optimum values for the minimum node size; meanwhile, $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ choose 75 (panel (c)). $\widehat{\mathrm{CIBS}}_n$ and $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ pick 10 and 25 for their minimum node size, respectively. To be more precise, the results for the first test data (see Table 6.2) show that the optimum values of minimum node size are different.

In this last section of this chapter, we applied the proposed measures to DTSF when we varied its minimum node size while other hyper-parameters were fixed. According to our numerical experiments, the best model's performance was achieved when the minimum node sizes were 10, 25, or 75. It is clear from our results in Section 6.3 that we may have different capabilities between discrimination and calibration when evaluating the performance of one model depending on the fitted models. Thus, the inconsistency amongst the proposed measures in tuning the optimum hyper-parameters may happen in practice. Another possible reason for the inconsistency is that we applied a relatively large distance amongst the minimum node size.

By the end of this chapter, we have revealed that the integrated Brier score and pair calibration can be either more aligned to calibration or discrimination measures depending on the fitted models. When the prediction accuracy of the models is much more dominant than the models' discrimination ability, then the integrated Brier score and pair calibration align with the calibration measure. Otherwise, they are in line with the discrimination measure. The integrated Brier score and pair calibration are sensitive to the quality change in calibration and discrimination measures. Reporting one of discrimination and calibration only may be insufficient for one who wants to comprehensively evaluate the model's performance. Finally, we have two suggestions as follows:

i. The integrated Brier score and pair calibration show excellent model performance. In this case, we are sure that calibration and discrimination are good. Therefore, the pure calibration measures (e.g. the modified integrated Brier scores) and the pure discrimination measures (e.g. time-dependent Uno's C-index) may need not necessarily be reported.

ii. The integrated Brier score and pair calibration show poor model performance. This case suggests further investigation into the model's marginal predictive capabilities: calibration and discrimination. The poor performance of the integrated Brier score and pair calibration can be caused by poor discrimination, poor calibration, or poor discrimination and calibration.

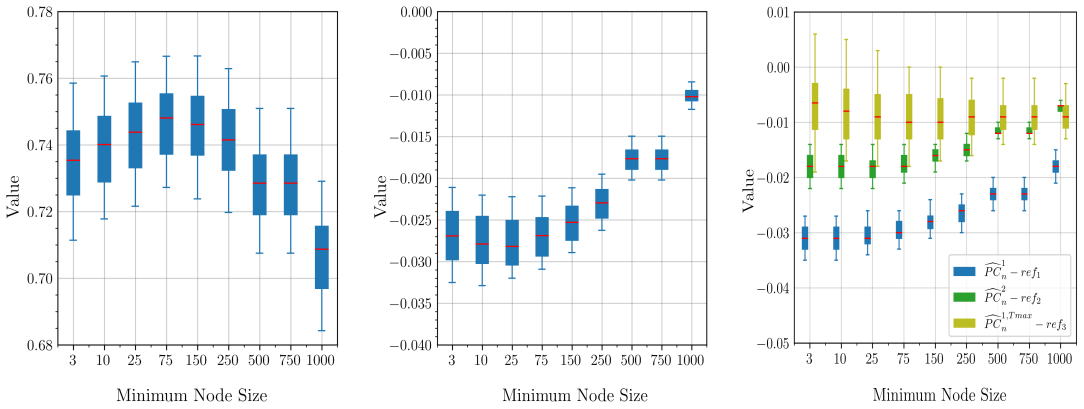


(a) Min. node size=3 (b) Min. node size=10 (c) Min. node size=25

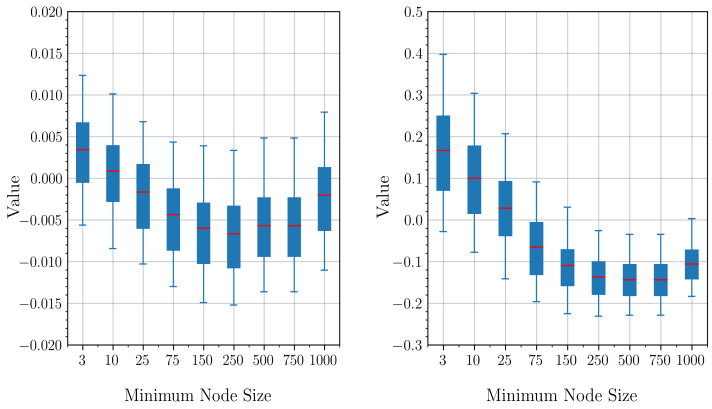
(d) Min. node size=75 (e) Min. node size=150 (f) Min. node size=250

(g) Min. node size=500 (h) Min. node size=750 (i) Min. node size=1000

Figure 6.6: Boxplots of $S(t; \mathbb{Z}_i)(i = 1, \cdots, 1000)$ for each $t \in \{1, \cdots, T_{max} - 1\}$ estimated from the first test data using DTSF and KM estimator over various minimum node size.



(a) $\widehat{\mathrm{C}}_n^w$ (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$ (c) Pair calibration and the references



(d) $\widehat{\mathrm{CIBS}}_n$ (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$

Figure 6.7: Boxplots of (a) $\widehat{\mathrm{C}}_n^w$, (b) $\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$, (c) pair calibration and the references, (d) $\widehat{\mathrm{CIBS}}_n$, and (e) $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$ computed from 100 test data in DTSF.

Min. Node Size	$\widehat{\mathrm{C}}_n^w$	$\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n$	$\widehat{\mathrm{CIBS}}_n$	$\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$	$\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1$
3	0.7195	-0.0285	0.00461	0.1900	-0.0300
10	0.7270	**-0.0293**	0.00121	0.1040	**-0.0310**
25	**0.7282**	-0.0289	0.00050	0.0872	**-0.0310**
75	0.7255	-0.0269	**-0.00045**	**-0.0190**	-0.0290
150	0.7235	-0.0239	-0.00143	-0.0778	-0.0270
250	0.7152	-0.0213	-0.00112	-0.1103	-0.0240
500	0.7013	-0.0162	0.00051	-0.1071	-0.0200
750	0.7013	-0.0162	0.00051	-0.1071	-0.0200
1000	0.6809	-0.0088	0.00392	-0.0684	-0.0150

Table 6.2: The values of $\widehat{\mathrm{C}}_n^w$, $(\widehat{\mathrm{IBS}}_n - \widehat{\mathrm{IBSKM}}_n)$, $\widehat{\mathrm{CIBS}}_n$, $\widehat{\mathrm{NCIBS}}_n^{\varepsilon}$, and $(\widehat{\mathrm{PC}}_n^1 - \mathrm{ref}_1)$ of DTSF from the first test data of TCGA data over various minimum node size.

Chapter 7

Conclusions

7.1 Research Summary

Throughout this thesis, we contribute to statistical methodology and applications. For the statistical methodology, our main contributions are: (1) proposing three novel measures based on Brier score, such as normalised integrated Brier score, centered integrated Brier score, and normalised centered Brier score, (2) proposing time-dependent Uno's C-index and showing its convergence in detail, and (3) proposing a novel measure, namely pair calibration. For the applications, we mainly contributed: (1) to demonstrate some crucial pitfalls of integrated Brier score through comprehensive numerical experiments, (2) to show the case where time-dependent concordance can be downward biased, and (3) to provide some unique cases where calibration and discrimination may not be in line with each other.

In Chapter 2 of this thesis, we introduced the basic notations, the evaluated non-linear model, the standard measures of predictive performance, and the assumptions used throughout the thesis. It is crucial to understand these assumptions as they form the basis of our work. Our objective is not to discuss the more comprehensive literature review for each topic due to the limited space of this thesis. The discussions can be the most crucial foundations to understand the subsequent chapters, such as distinguishing the two main measures in survival analysis, namely calibration and discrimination, and why we chose Nnet-survival as the main example of non-linear models.

Chapter 3 discusses some calibration measures, namely the integrated Brier score and several of its modified versions. We began with some case studies, including simulation and real-world examples, to demonstrate some crucial pitfalls of integrated Brier score. We found that the integrated Brier score can be predominantly affected by the data structure. When the data structure significantly changes due to administrative censoring, the Brier score at the right tail of $\{1, \cdots, T_{max} - 1\}$ might be ignored from the computation of the integrated Brier score. As a result, the integrated Brier score also changed depending on the contributions of the removed Brier score. Another case was when the data contained a very high proportion of administratively censored individuals. This case results in $S(t; \mathbb{Z}_i) \approx 1$ for each individual $i (i = 1, \cdots, n)$ and $t \in \{1, \cdots, T_{max} - 1\}$. Thus, $\widehat{\text{IBS}}_n$ is always close to zero regardless of the quality of the fitted models. To cope with such issues, we used the integrated Brier score for KM estimator as the reference value of the integrated Brier score. Our results showed that the integrated Brier score for KM estimator can help us convincingly state whether our models' predictive performance is good or bad.

In terms of interpretation, we can only interpret the integrated Brier score as follows: the closer the score to zero, the better the model performance. Therefore, to improve the interpretability of the integrated Brier score, we proposed three modified integrated Brier scores, namely normalised integrated Brier score, centered integrated Brier score, and normalised centered integrated Brier score. Through some numerical experiments, using centered integrated Brier score, we struggled to distinguish the performance of different models because the values lied on the small range of interval. On the other hand, the normalised integrated Brier score and the normalised centered integrated Brier score do not have such drawbacks. Moreover, they could be obtained once we have computed one of them. Thus, they share the same properties and behaviour except their interpretations.

Our study also discovered a severe pitfall of the normalised integrated Brier score (and hence the normalised centered integrated Brier score). Their values can blow up massively or even cannot be computed when some of $S(T_{max} - 1; \mathbb{Z}_i)$ were close to one or even equal one. We therefore proposed their truncated versions at $\varepsilon \in [0, 1]$, where $S(t; \mathbb{Z}_i) = \varepsilon$ if $S(t; \mathbb{Z}_i) \leq \varepsilon$ or $S(t; \mathbb{Z}_i) = (1 - \varepsilon)$

if $S(t;\mathbb{Z}_i) \geq (1-\varepsilon)$. Those truncated versions can effectively cope with such drawback as long as we were able to choose the best value for ε. In the final part of this chapter, we also concluded that the three types of modified integrated Brier scores are pure calibration measures. Meanwhile, the integrated Brier score is combination between calibration and discrimination measures.

In Chapter 4, we moved on to another type measure of performance, namely discrimination. This chapter started with a simulation study that demonstrated the disadvantage of Uno's C-index when we evaluated a non-PH model fitted to non-PH data. This result had motivated us to develop an overall measure that can cope with such drawback of Uno's C-index. In particular, we developed time-dependent Uno's C-index which can be seen as the weighted version of Antolini's time-dependent concordance. Our main contributions in developing time-dependent Uno's C-index can be viewed from theoretical and practical perspective. From the theoretical point of view, it is common that there are a number of authors, such as Gerds *et al.* (2013) and Cheung *et al.* (2019), who directly applied the results of the paper by Uno *et al.* (2011) when they showed the convergence of their proposed discrimination measures. In the convergence proof by Uno *et al.* (2011), there is a crucial step which states that Uno's C-index is a U-statistic such that it converges almost surely to its expectation following Theorem 7 in the paper by Nolan & Pollard (1987). In this thesis, we have shown how our proposed discrimination measure satisfies the assumptions of the theorem in more detail, which is not trivial. From the practical point of view, the major contribution of this chapter is to show that time-dependent concordance can be either upward or downward bias. Since time-dependent concordance is an extension of Harrell's C-index, our results more or less significantly affect the common insight on how Harrell's C-index behave.

Chapter 5 proposes a novel measure, pair calibration, a mean squared error that compares the observed order of events for two individuals with their predicted probability. It serves as a calibration measure for the discrimination of risks (ranking) predicted by the model. It furthermore helps us understand how well our model's predictions align with the outcomes. Thus, pair calibration bridges calibration and discrimination, offering a comprehensive view of prediction assessment. Our research in this chapter was more than just theoretical.

We proposed several estimators of pair calibration and also demonstrated their convergence. Through simulation studies and real-world examples, we showed that pair calibration can be influenced by the data structure. Pair calibration may not be able show a bad model performance when most event times are in the left tail of the follow-up time while most of the data were administratively censored. This practical insight led us to propose two measures as the reference values for pair calibration: (1) the reference values based on the worst case of model performance, and (2) the reference values based on the proportion of the outcomes. However, we concluded that the first reference values were impractical. Meanwhile the second reference values had promising use as the complementary measures for pair calibration. We also introduced the truncated pair calibration, a practical solution for those interested in prediction at sub-intervals of the follow-up.

Throughout Chapter 6, we conducted several comprehensive numerical experiments for all proposed measures. We started with how the measures behaved in a fixed model, but the censoring rate in the test data varied. The results showed that the proposed measures were very close to their "true" values for lower censoring rates, that is, up to 45% censoring rate. However, the differences between the proposed measures and their true values were more significant when we had higher censoring rates, e.g. 62% and 75%. Although the proposed measures are theoretically unbiased, their estimators are usually biased due to the higher number of ignored pairs and the misspecification of $\widehat{G}_n$ when censoring presences in the data. The higher the censoring rates in the data, the more errors we obtained from $\widehat{G}_n$.

Next, we applied the proposed measures to evaluate different models, where we varied the values of L^2-regularisation (λ) when we trained the models to obtain different models. From the experiments, we found that all proposed measures can distinguish different models well. Different measures can choose different optimum λ since each measure evaluates different models' outputs. Thus, we must carefully use the measures to tune the models' hyper-parameters. If our study aims to obtain a model with good discrimination, we should use a discrimination measure. Otherwise, calibration measures should be used if we need a model with good calibration ability.

Chapter 6 also provides several examples of cases where calibration and discrimination measures were different from one another. The cases were: (1) good discrimination but poor calibration, and (2) good calibration but poor discrimination. In the first case, we found that the time-dependent Uno's C-index was not crucially affected by the prediction error as long as the order of the model's outputs did not change significantly. However, the proposed calibration measures are highly dependent on the output accuracies. Meanwhile, the prediction error also affected the integrated Brier score and pair calibration. In the second case, we showed that the integrated Brier score and pair calibration aligned more with the calibration measures. From these results, our suggestions are: (1) when the integrated Brier score and pair calibration are excellent, we may not necessarily report the pure calibration measures (e.g. the modified integrated Brier scores) and the pure discrimination measures (e.g. time-dependent Uno's C-index), and (2) when the integrated Brier score and pair calibration are poor, we highly suggest further investigation into the model's marginal predictive capabilities: calibration and discrimination.

In the final part of Chapter 6, we presented an example where we applied our proposed measures to evaluate the model performance of discrete-time survival forests. In particular, we implemented the proposed measures when we varied the minimum node size while we fixed the other hyper-parameters of the discrete-time survival forests. It is clear from our numerical results in Section 6.3 that we may have different capabilities between discrimination and calibration when evaluating the performance of one model depending on the fitted model. We found the inconsistency among the proposed measures in tuning the best minimum node size. We suggested two reasons for the results: (1) the characteristics of the fitted models and (2) the large distance amongst the used minimum node sizes.

Throughout this thesis, all numerical results of the simulation studies and real-world examples, especially for evaluating the model performance of Nnet-survival, were computed using high performance computation (HPC) facilities (ARC3 and ARC4) provided by the Research Computing Team at the University of Leeds. For example, in the simulation studies with a train data (n_{train}=1000) and 100 independent test data (n_{test}=1000), the evaluation of model performance

using all proposed measures simultaneously was conducted about one hour in the average with maximum memory about 500 MB.

7.2 Further Research and Possible Extensions

To compute $\widehat{\text{IBS}}_n$ given in (2.11), we applied $(1/T_{max} - 1)$ as the weight for each period, meaning that we treated equally each $\widehat{\text{BS}}_n(t)$ for all $t \in \{1, \cdots, T_{max} - 1\}$. As we have shown in our simulation studies and real-world examples, such an approach resulted in the dependence of the integrated Brier score on the data structure. Using such fixed constant weight to all time points within the follow-up is a common practice in the literature (Fotso *et al.*, 2019–; Haider *et al.*, 2020; Moradian *et al.*, 2017). Although we can follow the weights mentioned in Graf *et al.* (1999), namely t/t^* and $(1 - S(t))/(1 - S(t^*))$, where in this thesis $t^* = T_{max} - 1$, these weights may only suitable for specific distribution of $\widehat{\text{BS}}_n(t)$. For instance, if we use t/t^*, we assume that $\widehat{\text{BS}}_n(t)$ linearly increases as t increases so that larger weights need to be given for $\widehat{\text{BS}}_n(t)$ in the right-tail of the follow-up. This situation is unsuitable for most of our numerical results, where most $\widehat{\text{BS}}_n(t)$ oscillated over the follow-up. Therefore, more research about the weights is still required to match the obtained $\widehat{\text{BS}}_n(t)$ characteristics. Furthermore, the research could be conducted also for the normalised integrated Brier score, the centered integrated Brier score, and the normalised centered integrated Brier score.

The discussions in Chapter 4 offer more questions for further research. The obtained simulation results showed that Uno's C-index is not a proper measure for assessing the fitted non-linear model since $\widehat{\text{C}}_n^{\text{uno}}$ varied over the follow-up depending on the non-proportionality behaviour of the data. Along with Harrell's C-index and Gönen's C-index, Uno's C-index can only evaluate models' performance with the PH assumption. In this thesis, we restricted the discussions to a single event of interest and right-censored survival data. The next question is how we extend the time-dependent Uno's C-index to evaluate the performance of the other non-linear survival models developed for survival data with different characteristics. In current literature, several machine learning approaches have been proposed for more complex cases in survival analysis, such as Neural survival recommender (Jing & Smola, 2017) as the long short-term networks (LSTM)

for recurrent events, partial logistic artificial neural network competing risks automatic relevance determination (PLANNCR-ARD) (Lisboa *et al.*, 2003) and random survival forests competing risks (RSFCR) (Ishwaran *et al.*, 2014) for the competing risks, and neural networks for interval-censored survival data (Meixide *et al.*, 2024). In these models, we cannot directly apply the time-dependent Uno's C-index to evaluate their model's performance, requiring further research and extensions.

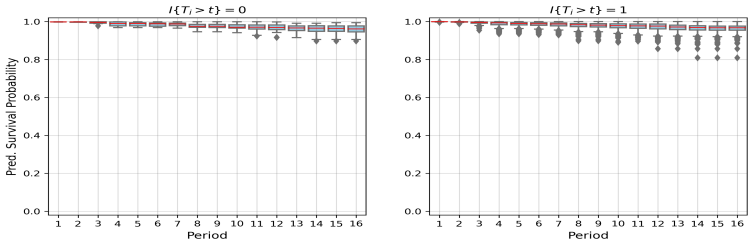
Our proposed pair calibration is based on the mean squared error, which is the same as for the Brier score. Hence, the interpretation of the pair calibration is also limited, as in the Brier score. By using the proposed reference values, we have improved its interpretability. However, we can still improve the interpretability using the same ideas to obtain the modified Brier scores. In particular, we could extend the pair calibration by involving the variabilities of the predicted probabilities of interest in pair calibration. By involving the variance (or the standard deviation of the probabilities), we could modify pair calibration by applying the "central limit theorem" in statistics. As a result, we could conduct a hypothesis testing about the differences between the model's outcome and the model's output. For example, we would extend $\widehat{\text{PC}}_n^1$ to obtain a statistical test regarding the difference between $I_{\{T_i \leq T_j\}}$ and $\pi_{ij}^{T_{max}}$. Of course, this is an interesting extension for further research.

When we applied the reference values for pair calibration and integrated Brier score, we categorised the model performance based on the following conditions: (1) if the measures are much less than the reference values, the model performance is excellent, (2) if the measures are close to the reference values, the model performance is poor, and (3) if the measures are much greater than the reference values, the model performance is very bad. Using these conditions, for example, we are sometimes confused when the measures show the values between the excellent and poor model performance. In this thesis, we did not define how much the values to be achieved by the estimators so that we can convincingly categorise the model performance as excellent, poor, very bad, or even in between. Therefore, this problem should be investigated for further research. In the modified integrated Brier scores, we also have such "threshold" issue. Thus, the same further research should also be conducted for these measures.

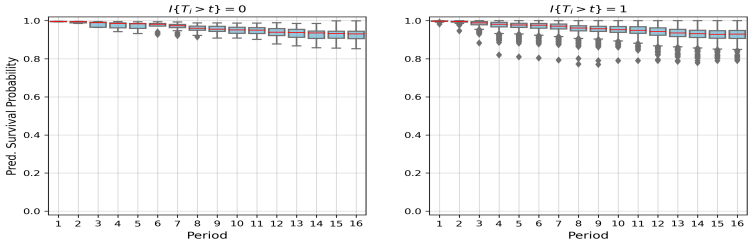
Appendix A

Plots

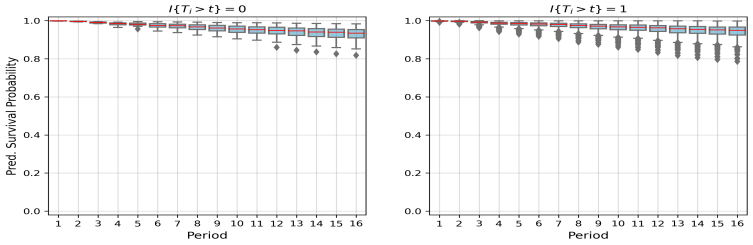
A.1 Plots in Chapter 3



(a) $\mathscr{D}_{11}$ with $\lambda = 0$

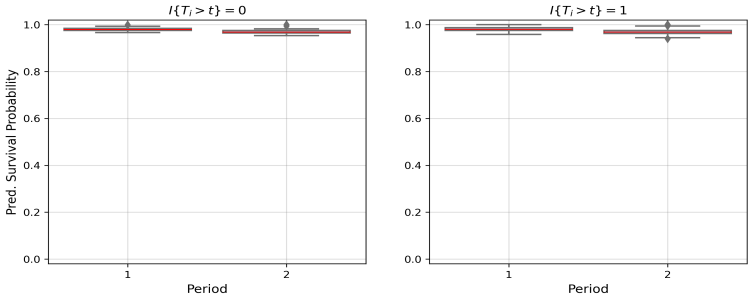


(b) $\mathscr{D}_{11}$ with $\lambda = 1E-4$

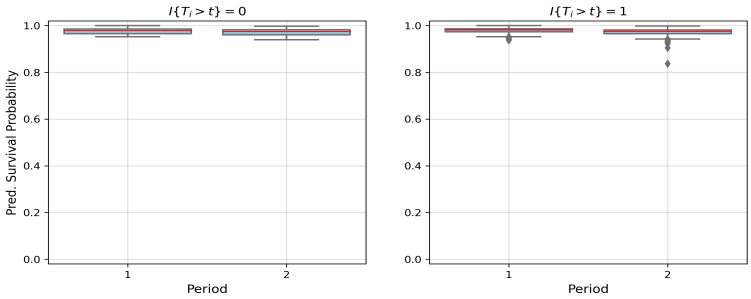


(c) $\mathscr{D}_{11}$ with $\lambda = 0.2$

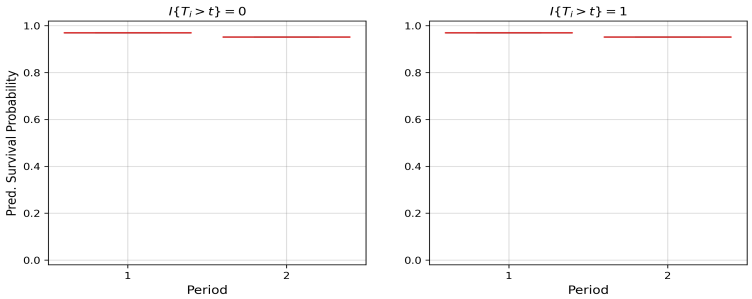
Figure A.1: $S(t; \mathbb{Z}_i)$ for $i = 1, \cdots, 1000$ categorised by $I_{\{T_i > t\}}$ over $\{1, \cdots, T_{max} - 1\}$ in the breast cancer data. The prediction was conducted in a test data based on a model fitted to the respective train data with three values of L^2-regularisation (λ), i.e. $0, 1E-4$, and 0.2. The train and test data were discretised by $\mathscr{D}_{11}$.



(a) $\mathscr{D}_{12}$ with $\lambda = 0$



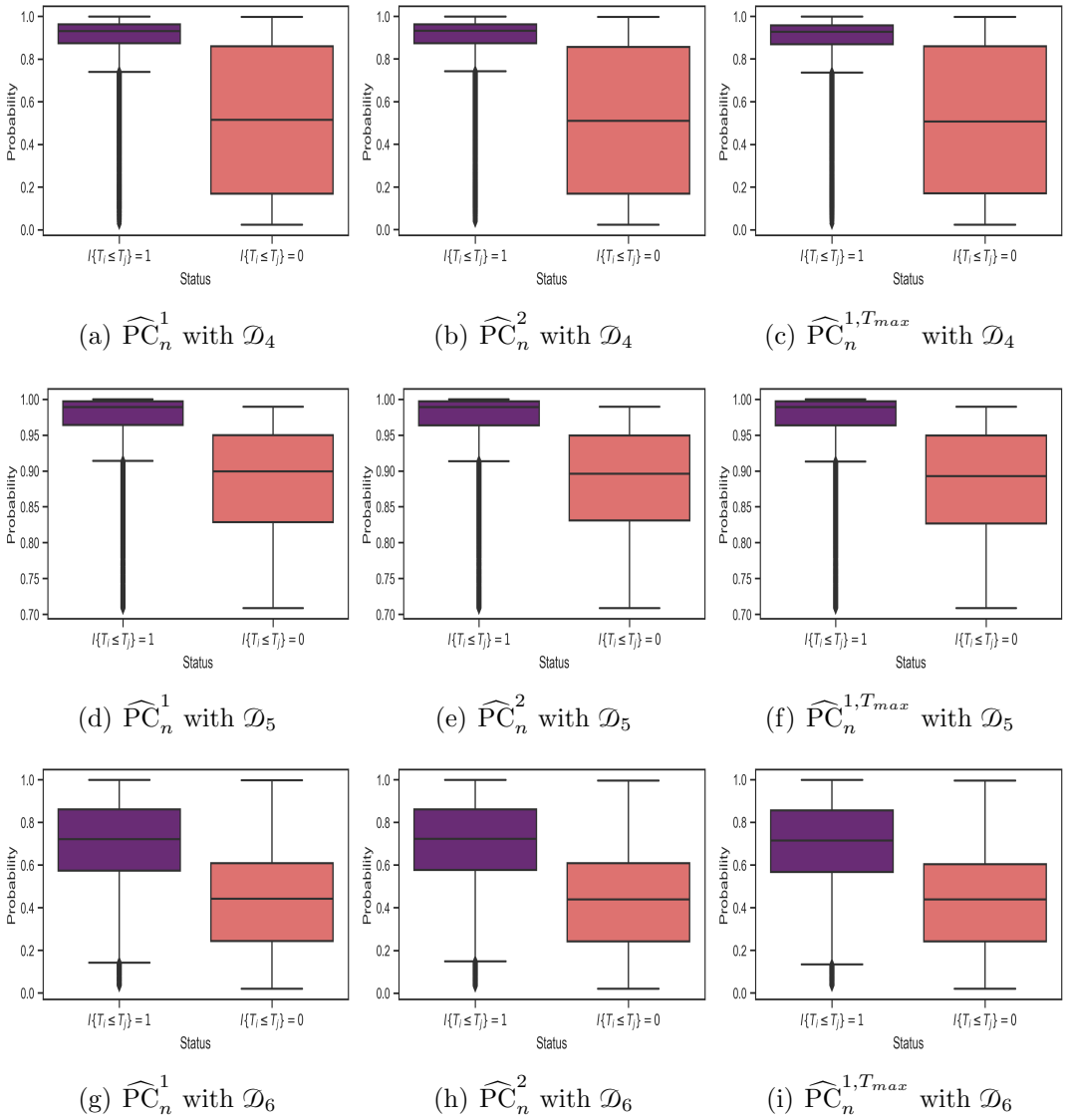
(b) $\mathscr{D}_{12}$ with $\lambda = 1E - 4$



(c) $\mathscr{D}_{12}$ with $\lambda = 0.2$

Figure A.2: $S(t; \mathbb{Z}_i)$ for $i = 1, \cdots, 1000$ categorised by $I_{\{T_i > t\}}$ over $\{1, \cdots, T_{max} - 1\}$ in the breast cancer data. The prediction was conducted in a test data based on a model fitted to the respective train data with three values of L^2-regularisation (λ), i.e. $0, 1E - 4$, and 0.2. The train and test data were discretised by $\mathscr{D}_{12}$.

A.2 Plots in Chapter 5

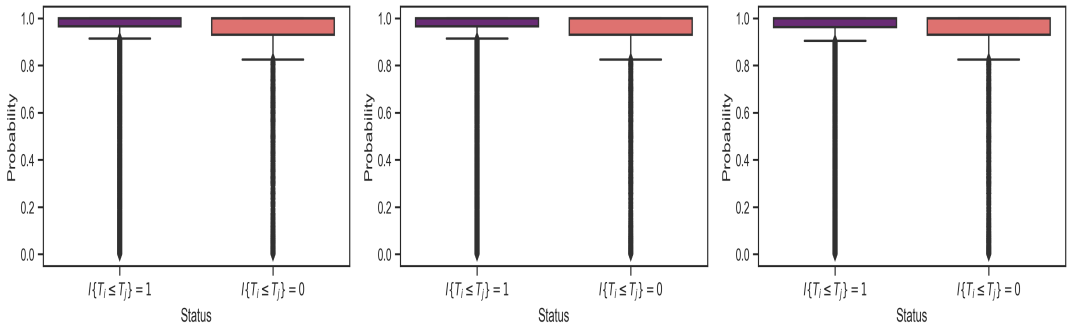


(a) $\widehat{\mathrm{PC}}_n^1$ with $\mathscr{D}_4$ (b) $\widehat{\mathrm{PC}}_n^2$ with $\mathscr{D}_4$ (c) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_4$

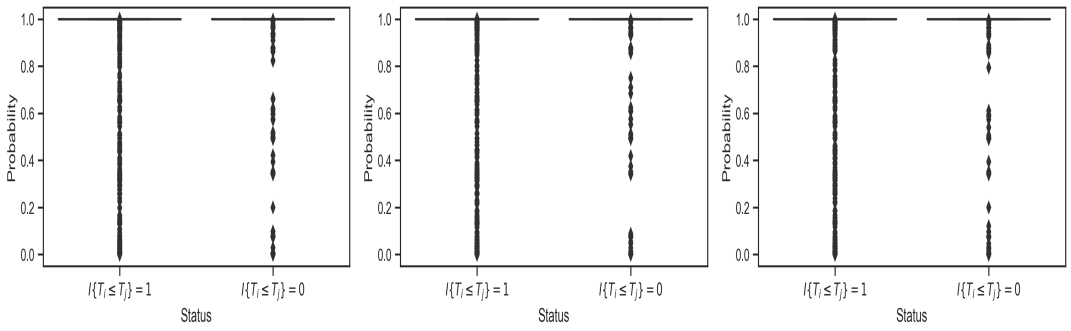
(d) $\widehat{\mathrm{PC}}_n^1$ with $\mathscr{D}_5$ (e) $\widehat{\mathrm{PC}}_n^2$ with $\mathscr{D}_5$ (f) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_5$

(g) $\widehat{\mathrm{PC}}_n^1$ with $\mathscr{D}_6$ (h) $\widehat{\mathrm{PC}}_n^2$ with $\mathscr{D}_6$ (i) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_6$

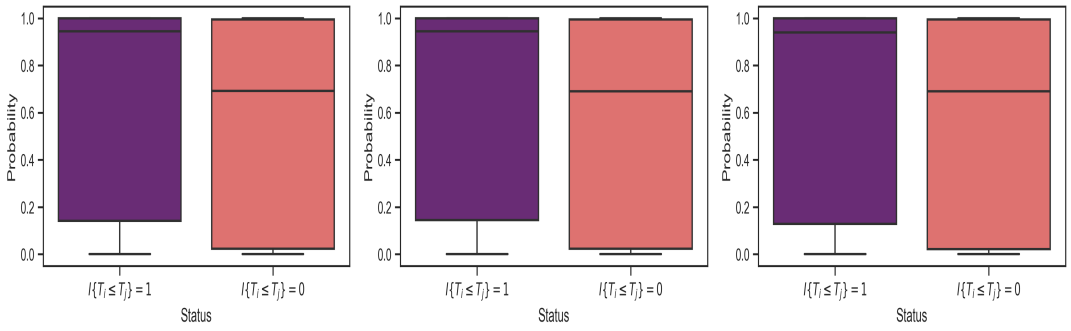
Figure A.3: The predicted probabilities of interest of $\widehat{\mathrm{PC}}_n^1$, $\widehat{\mathrm{PC}}_n^2$, and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for 20% randomly selected pairs $i \neq j$ over each discretisation setup. They were obtained from the good Nnet-survival in Scenario 2 of Simulation 1 fitted to a fixed train data ($n_{\text{train}} = 1000$) and evaluated on the first test data ($n_{\text{test}} = 1000$).



(a) $\widehat{\mathrm{PC}}_n^1$ with $\mathscr{D}_4$ (b) $\widehat{\mathrm{PC}}_n^2$ with $\mathscr{D}_4$ (c) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_4$

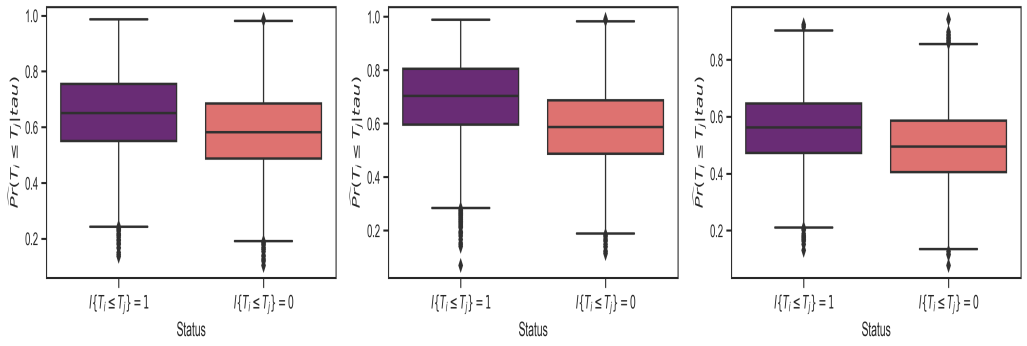


(d) $\widehat{\mathrm{PC}}_n^1$ with $\mathscr{D}_5$ (e) $\widehat{\mathrm{PC}}_n^2$ with $\mathscr{D}_5$ (f) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_5$

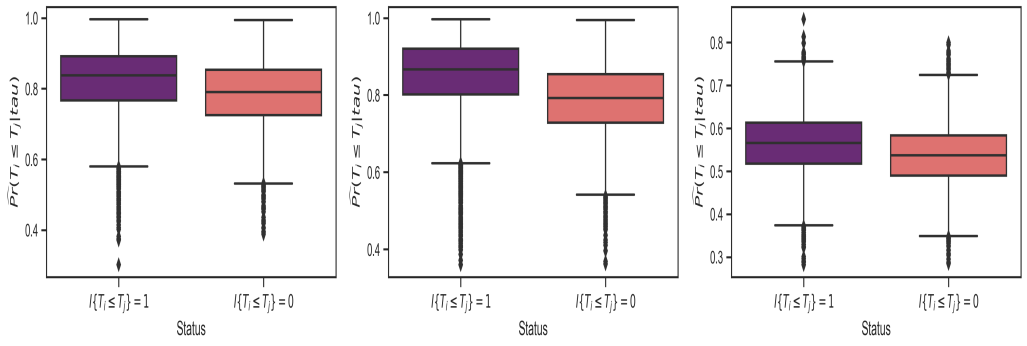


(g) $\widehat{\mathrm{PC}}_n^1$ with $\mathscr{D}_6$ (h) $\widehat{\mathrm{PC}}_n^2$ with $\mathscr{D}_6$ (i) $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_6$

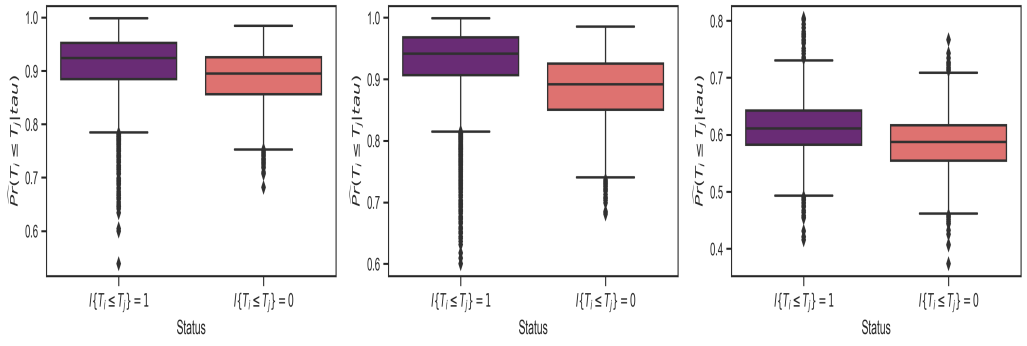
Figure A.4: The predicted probabilities of interest of $\widehat{\mathrm{PC}}_n^1$, $\widehat{\mathrm{PC}}_n^2$, and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for 20% randomly selected pairs $i \neq j$ over each discretisation setup. They were obtained from the overfitted Nnet-survival in Scenario 2 of Simulation 1 fitted to a fixed train data ($n_{\text{train}} = 1000$) and evaluated on the first test data ($n_{\text{test}} = 1000$).



(a) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_7$ (b) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_7$ (c) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_7$

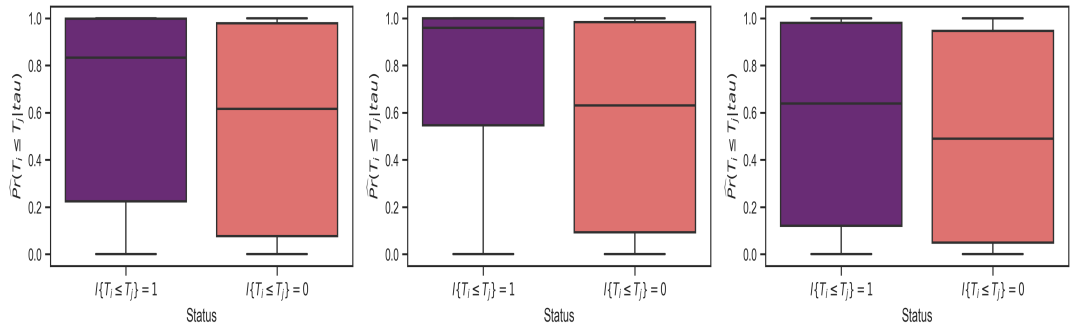


(d) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_8$ (e) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_8$ (f) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_8$

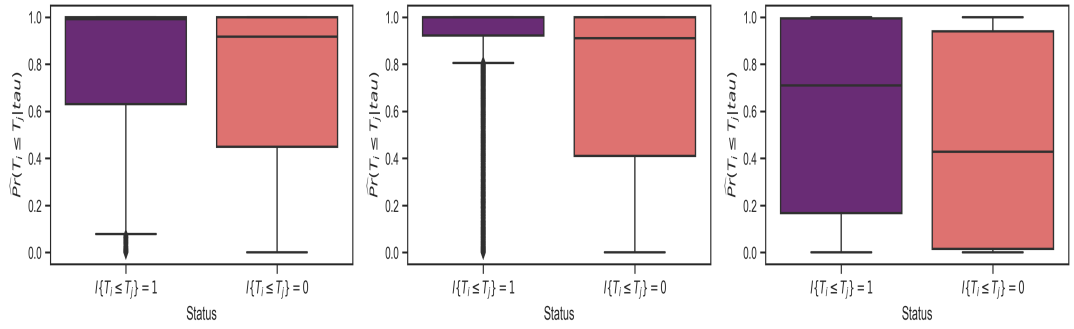


(g) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_9$ (h) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_9$ (i) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_9$

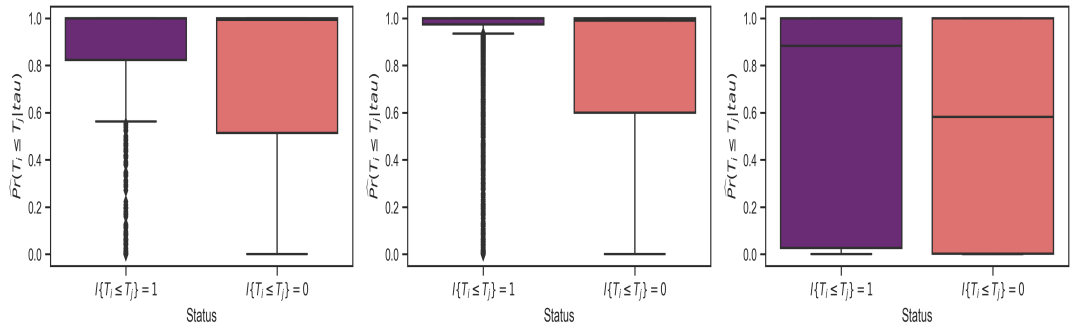
Figure A.5: The predicted probabilities of interest of $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ for 20% randomly selected pairs $i \neq j$ over each discretisation setup in TCGA data. They were obtained from the good Nnet-survival fitted to the first train data ($n_{\text{train}} = 1000$) and evaluated on the first test data ($n_{\text{test}} = 1000$).



(a) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_7$ (b) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_7$ (c) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_7$



(d) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_8$ (e) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_8$ (f) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_8$



(g) $\widehat{\text{PC}}_n^1$ with $\mathscr{D}_9$ (h) $\widehat{\text{PC}}_n^2$ with $\mathscr{D}_9$ (i) $\widehat{\text{PC}}_n^{1,T_{max}}$ with $\mathscr{D}_9$

Figure A.6: The predicted probabilities of interest of $\widehat{\text{PC}}_n^1$, $\widehat{\text{PC}}_n^2$, and $\widehat{\text{PC}}_n^{1,T_{max}}$ for 20% randomly selected pairs $i \neq j$ over each discretisation setup in TCGA data. They were obtained from the overfitted Nnet-survival fitted to the first train data ($n_{\text{train}} = 1000$) and evaluated on the first test data ($n_{\text{test}} = 1000$).

Appendix B

Architectures of The Fitted Machine Learning Approaches

B.1 Architectures in Chapter 3

Hyper-parameters	**Good Model**	**Overfitted Model**
#Hidden layer	2	4
#Nodes in each layer	(8,16)	(256,256,256,256)
Hidden activation function	ReLu	ReLu
Output activation function	Sigmoid	Sigmoid
L^2-regularisation (λ)	1E-2	0
Batch size	250	250
#Epoch	1000	1000
Learning rates	1E-3	1E-3
Early stopping (patience)	Yes (20)	Yes (20)
Lower bound (ϵ) for $\widehat{G}_n$	0.02	0.02

Table B.1: Nnet-survival architectures of Simulation 1.

Hyper-parameters	**Good Model**	**Overfitted Model**
#Hidden layer	2	4
#Nodes in each layer	(64,64)	(256,256,256,256)
Hidden activation function	ReLu	ReLu
Output activation function	Sigmoid	Sigmoid
L^2-regularisation (λ)	0.02	0
Batch size	250	250
#Epoch	1000	1000
Learning rates	1E-3	1E-3
Early stopping (patience)	Yes (20)	Yes (20)
Lower bound (ϵ) for $\widehat{G}_n$	0.02	0.02

Table B.2: Nnet-survival architectures of TCGA data.

Hyper-parameters	**Values**
#Hidden layer	4
#Nodes in each layer	(16,16,16,16)
Hidden activation function	ReLu
Output activation function	Sigmoid
L^2-regularisation (λ)	(0, 1E-4,0.2)
Batchsize	250
#Epoch	1000
Learning rates	1E-3
Early stopping (Patience)	Yes (20)
Lower bound (ϵ) for $\widehat{G}_n$	0.02

Table B.3: Nnet-survival architectures of breast cancer data.

B.2 Architectures in Chapter 4

Hyper-parameters	**Non-PH Data**	**PH Data**
#Hidden layer	2	2
#Nodes in each layer	(16,16)	(16,16)
Hidden activation function	ReLu	ReLu
Output activation function	Sigmoid	Sigmoid
L^2-regularisation (λ)	1E-6	1E-2
Batch size	50	250
#Epoch	150	1000
Learning rates	1E-2	1E-3
Early stopping (patience)	Yes (20)	Yes (20)
Lower bound (ϵ) for $\widehat{G}_n$	0.02	0.02

Table B.4: Nnet-survival architectures of the simulation studies.

Hyper-parameters	**Sim.3: PH**	**Sim.3: Non-PH**	**Sim.4**	**HF Data**
#Hidden layer	2	2	2	2
#Nodes hidden layers	(16,16)	(128,128)	(16,16)	(8,8)
#Nodes output layers	11	11	15	15
Hidden activation function	ReLu	ReLu	ReLu	ReLu
Output activation function	Sigmoid	Sigmoid	Sigmoid	Sigmoid
L^2-regularisation (λ)	0.075	1E-2	1E-6	0.1
Batch Size	250	250	50	52
#Epoch	1000	1000	150	300
Learning rates	1E-3	1E-4	1E-2	1E-3
Early stopping (patience)	Yes (20)	Yes (20)	Yes (20)	Yes (20)
Lower bound (ϵ) for $\widehat{G}_n$	0.02	0.02	0.02	0.02

Table B.5: Nnet-survival architectures of the simulation studies and HF data.

B.3 Architectures in Chapter 6

Hyper-parameters	**Scenario 2 Case 1**	**Case 2**
#Hidden layer	2	2
#Nodes in each layer	(256,256)	(128,128)
Hidden activation function	ReLu	ReLu
Output activation function	Sigmoid	Sigmoid
L^2-regularisation (λ)	1E-3	0.22
Batch size	250	250
#Epoch	1000	1000
Learning rates	1E-3	1E-4
Early stopping (patience)	Yes (20)	Yes (20)
Lower bound (ϵ) for $\widehat{G}_n$	0.02	0.02

Table B.6: Nnet-survival architectures for the example cases in Section 6.3.

Hyper-parameters	**TCGA DTSF**
Minimum node size	Varies (i.e.3,10,25,75, 150,250,500,750, and 1000)
#Trees	100
#Variables to possibly split at in each node	Rounded down square root of the number variables
Splitting rule	Hellinger
Maximal tree depth	Unlimited
Grow a probability forest	Yes
Lower bound (ϵ) for $\widehat{G}_n$	0.02

Table B.7: DTSF architecture of TCGA data.

Appendix C

Numerical Results

C.1 Results of Chapter 5

$\mathscr{D}_4$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.219	0.127	0.220	0.127	0.228	0.132
0.209	0.126	0.209	0.126	0.218	0.131
0.237	0.137	0.237	0.137	0.244	0.141
0.238	0.135	0.237	0.135	0.241	0.140
0.216	0.135	0.215	0.135	0.223	0.138
$\mathscr{D}_5$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.053	0.036	0.053	0.036	0.054	0.036
0.065	0.043	0.065	0.043	0.065	0.043
0.044	0.029	0.044	0.029	0.044	0.029
0.043	0.032	0.043	0.032	0.043	0.032
0.039	0.023	0.039	0.023	0.039	0.023
$\mathscr{D}_6$					
$\widehat{\text{PC}}_n^1$	ref_1	$\widehat{\text{PC}}_n^2$	ref_2	$\widehat{\text{PC}}_n^{1,T_{max}}$	ref_3
0.367	0.242	0.369	0.242	0.373	0.243
0.366	0.242	0.366	0.242	0.368	0.243
0.379	0.242	0.379	0.242	0.381	0.242
0.377	0.241	0.377	0.241	0.380	0.242
0.360	0.241	0.360	0.241	0.363	0.242

Table C.1: The reference values of $\widehat{\text{PC}}_n^1, \widehat{\text{PC}}_n^2$ and $\widehat{\text{PC}}_n^{1,T_{max}}$ for the first five test data discretised by $\mathscr{D}_4, \mathscr{D}_5$, and $\mathscr{D}_6$. These results were obtained from the over-fitted Nnet-survival architecture in Scenario 2 of Simulation 1.

$\mathscr{D}_7$					
$\widehat{\mathrm{PC}}_n^1$	ref_1	$\widehat{\mathrm{PC}}_n^2$	ref_2	$\widehat{\mathrm{PC}}_n^{1,T_{max}}$	ref_3
0.263	0.205	0.297	0.205	0.333	0.249
0.216	0.203	0.258	0.203	0.275	0.249
0.249	0.207	0.271	0.207	0.301	0.249
0.267	0.210	0.298	0.210	0.313	0.249
0.273	0.203	0.303	0.203	0.344	0.249
$\mathscr{D}_8$					
$\widehat{\mathrm{PC}}_n^1$	ref_1	$\widehat{\mathrm{PC}}_n^2$	ref_2	$\widehat{\mathrm{PC}}_n^{1,T_{max}}$	ref_3
0.194	0.141	0.188	0.141	0.319	0.248
0.166	0.132	0.165	0.132	0.349	0.248
0.177	0.129	0.161	0.129	0.340	0.247
0.173	0.132	0.175	0.132	0.371	0.248
0.187	0.127	0.184	0.127	0.389	0.248
$\mathscr{D}_9$					
$\widehat{\mathrm{PC}}_n^1$	ref_1	$\widehat{\mathrm{PC}}_n^2$	ref_2	$\widehat{\mathrm{PC}}_n^{1,T_{max}}$	ref_3
0.166	0.069	0.170	0.069	0.393	0.240
0.145	0.065	0.138	0.065	0.405	0.240
0.085	0.063	0.090	0.063	0.428	0.241
0.195	0.066	0.189	0.066	0.410	0.240
0.185	0.062	0.174	0.062	0.392	0.240

Table C.2: The reference values of $\widehat{\mathrm{PC}}_n^1, \widehat{\mathrm{PC}}_n^2$ and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for the first five test data discretised by $\mathscr{D}_7, \mathscr{D}_8$, and $\mathscr{D}_9$. These results were obtained from the over-fitted Nnet-survival architecture in TCGA data.

$\widehat{\mathrm{PC}}_n^1$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_1
0.022	0.022	0.022	0.025
0.022	0.022	0.022	0.024
0.021	0.021	0.021	0.024
0.022	0.021	0.022	0.024
0.022	0.021	0.021	0.024
$\widehat{\mathrm{PC}}_n^2$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_2
0.023	0.023	0.022	0.025
0.022	0.022	0.022	0.024
0.028	0.022	0.022	0.024
0.022	0.022	0.022	0.024
0.029	0.029	0.022	0.024
$\widehat{\mathrm{PC}}_n^{1,T_{max}}$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_3
0.254	0.254	0.249	0.249
0.249	0.248	0.249	0.249
0.253	0.253	0.249	0.249
0.249	0.253	0.249	0.249
0.255	0.258	0.249	0.249

Table C.3: The reference values of $\widehat{\mathrm{PC}}_n^1$, $\widehat{\mathrm{PC}}_n^2$ and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for $\mathscr{D}_{11}$ in breast cancer data.

$\widehat{\mathrm{PC}}_n^1$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_1
0.015	0.015	0.015	0.042
0.014	0.015	0.015	0.04
0.014	0.014	0.014	0.04
0.015	0.015	0.015	0.04
0.014	0.014	0.014	0.04
$\widehat{\mathrm{PC}}_n^2$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_2
0.015	0.015	0.015	0.042
0.015	0.015	0.015	0.04
0.014	0.014	0.014	0.04
0.015	0.015	0.015	0.04
0.014	0.014	0.014	0.04
$\widehat{\mathrm{PC}}_n^{1,T_{max}}$			
$\lambda = 0$	λ=1E-4	λ=0.2	ref_3^b
0.2	0.198	0.188	0.187
0.188	0.191	0.189	0.188
0.189	0.19	0.187	0.187
0.201	0.19	0.189	0.188
0.199	0.187	0.189	0.188

Table C.4: The reference values of $\widehat{\mathrm{PC}}_n^1, \widehat{\mathrm{PC}}_n^2$ and $\widehat{\mathrm{PC}}_n^{1,T_{max}}$ for the first five test breast cancer data discretised by $\mathscr{D}_{12}$.

Appendix D

Summary of Academic Activities

I participated in several academic activities during my PhD study from February 1st 2020 until the thesis submission on July 31st 2024. In the first year of my study, I was scheduled to participate in a series of crucial Academy for PhD Training in Statistics (APTS) modules. However, due to the Covid-19 outbreak, these courses were postponed. I engaged in these APTS modules in 2021, recognizing their importance in the earlier years of the study.

In 2021, I also delivered a talk at an internal PhD student seminar at the Department of Statistics, University of Leeds. I then presented a poster at the IMforFUTURE Workshop 2021 at the School of Mathematics, University of Leeds. In 2022, I gave a talk at the 2022 IMS International Conference on Statistics and Data Science (ICSDS) in Florence, Italy.

In April 2023, I gave a talk in the Research Students' Statistics Seminar, Department of Statistics, University of Leeds. Then, in May 2023, I also gave a talk in another seminar held by research students in the School of Mathematics, University of Leeds.

Finally, in September 2023, I gave a talk at the 46th Research Students' Conference in Probability and Statistics at the University of Sheffield.

Bibliography

Ahmad, T., Munir, A., Bhatti, S.H., Aftab, M. & Raza, M.A. (2017). Survival analysis of heart failure patients: A case study. *PloS one*, **12**, e0181001. 112

Aivaliotis, G., Palczewski, J., Atkinson, R., Cade, J.E. & Morris, M.A. (2021). A comparison of time to event analysis methods, using weight status and breast cancer as a case study. *Scientific reports*, **11**, 14058. 2, 51, 55

Antolini, L., Boracchi, P. & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in medicine*, **24**, 3927–3944. 4, 5, 12, 21, 22, 23, 88

Bender, R., Augustin, T. & Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, **24**, 1713–1723. 31

Breiman, L. (2001). Random forests. *Machine learning*, **45**, 5–32. 175

Brier, G.W. & Allen, R.A. (1951). Verification of weather forecasts. In T.F. Malone, ed., *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium of Meteorology*, 841–848, American Meteorological Society, Boston, MA. 2, 19

Brown, S.F., Branford, A.J. & Moran, W. (1997). On the use of artificial neural networks for the analysis of survival data. *IEEE transactions on neural networks*, **8**, 1071–1077. 9

Burke, H.B., Goodman, P.H., Rosen, D.B., Henson, D.E., Weinstein, J.N., Harrell Jr, F.E., Marks, J.R., Winchester, D.P. & Bostwick, D.G. (1997). Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, **79**, 857–862. 166

Cade, J.E., Burley, V.J., Alwan, N.A., Hutchinson, J., Hancock, N., Morris, M.A., Threapleton, D.E. & Greenwood, D.C. (2017). Cohort profile: the uk women's cohort study (ukwcs). *International journal of epidemiology*, **46**, e11–e11. 51

Cheung, L.C., Pan, Q., Hyun, N. & Katki, H.A. (2019). Prioritized concordance index for hierarchical survival outcomes. *Statistics in medicine*, **38**, 2868–2882. 183

Chicco, D. & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical informatics and decision making*, **20**, 16. 112

Cox, D.R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**. 1

D'Agostino, R.B. & Nam, B.H. (2003). Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, **23**, 1–25. 2, 3

DeGroot, M.H. & Fienberg, S.E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **32**, 12–22. 2, 19

Deng, Y., Liu, L., Jiang, H., Peng, Y., Wei, Y., Zhou, Z., Zhong, Y., Zhao, Y., Yang, X., Yu, J. *et al.* (2023). Comparison of state-of-the-art neural network survival models with the pooled cohort equations for cardiovascular disease risk prediction. *BMC Medical Research Methodology*, **23**, 22. 14

Fang, H. & Gough, J. (2014). Thednet'approach promotes emerging research on cancer patient survival. *Genome medicine*, **6**, 1–16. 46

Fotso, S. *et al.* (2019–). PySurvival: Open source package for survival analysis modeling. 186

Gensheimer, M.F. & Narasimhan, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ*, **7**, e6257. 1, 9, 12, 14, 21, 47, 57, 88

Gerds, T.A., Kattan, M.W., Schumacher, M. & Yu, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, **32**, 2173–2184. 4, 5, 11, 12, 24, 25, 57, 90, 111, 113, 183

Gijbels, I. (2010). Censored data. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 178–188. 11

Gneiting, T. & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, **1**, 125–151. 166

Goldstein, M., Han, X., Puli, A., Perotte, A. & Ranganath, R. (2020). X-cal: Explicit calibration for survival analysis. *Advances in neural information processing systems*, **33**, 18296–18307. 2

Gönen, M. & Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, **92**, 965–970. 4, 113

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016). *Deep learning*, vol. 1. MIT press Cambridge. 14

Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, **18**, 2529–2545. 2, 19, 57, 186

Haider, H., Hoehn, B., Davis, S. & Greiner, R. (2020). Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, **21**, 1–63. 2, 19, 186

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L. & Rosati, R.A. (1982). Evaluating the yield of medical tests. *Jama*, **247**, 2543–2546. 3, 21, 25, 107

Heagerty, P.J. & Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, **61**, 92–105. 4, 22, 88

Hosmer, D.W. & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, **9**, 1043–1069. 2

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S. *et al.* (2008). Random survival forests. *The Annals of Applied Statistics*, **2**, 841–860. 1, 46, 173, 175

Ishwaran, H., Gerds, T.A., Kogalur, U.B., Moore, R.D., Gange, S.J. & Lau, B.M. (2014). Random survival forests for competing risks. *Biostatistics*, **15**, 757–773. 186

Jadhav, A., Pramod, D. & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, **33**, 913–933. 46

Jing, H. & Smola, A.J. (2017). Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 515–524. 186

Kamran, F. & Wiens, J. (2021). Estimating calibrated individualized survival curves with deep learning. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 240–248. 166

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A. *et al.* (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339. 46

Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53**, 457–481. 1

Kleinbaum, D.G. & Klein, M. (2010). *Survival analysis*. Springer. 1, 11

Kvamme, H. & Borgan, Ø. (2021). Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, **27**, 710–736. 12

Kvamme, H. & Borgan, Ø. (2023). The brier score under administrative censoring: Problems and a solution. *Journal of Machine Learning Research*, **24**, 1–26. 4

Kvamme, H., Borgan, Ø. & Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, **20**, 1–30. 21

Lisboa, P., Wong, H., Harris, P. & Swindell, R. (2003). A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine*, **28**, 1–25. 186

Loève, M. & Loève, M. (1977). *Elementary probability theory*. Springer. 158

Meixide, C.G., Matabuena, M., Abraham, L. & Kosorok, M.R. (2024). Neural interval-censored survival regression with feature selection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **17**, e11704. 187

Memon, S.M., Wamala, R. & Kabano, I.H. (2023). A comparison of imputation methods for categorical data. *Informatics in Medicine Unlocked*, **42**, 101382. 46

Moradian, H., Larocque, D. & Bellavance, F. (2017). L_1 l 1 splitting rules in survival forests. *Lifetime data analysis*, **23**, 671–691. 186

Murphy, A.H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, **12**, 595–600. 6

Nolan, D. & Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics*, 780–799. 24, 90, 91, 100, 183

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830. 46

Pepe, M.S., Zheng, Y., Jin, Y., Huang, Y., Parikh, C.R. & Levy, W.C. (2008). Evaluating the roc performance of markers for future events. *Lifetime data analysis*, **14**, 86–113. 4

Rahman, M.S., Ambler, G., Choodari-Oskooei, B. & Omar, R.Z. (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Medical research methodology*, **17**, 60. 3

Royston, P. & Parmar, M.K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*, **21**, 2175–2197. 57

Schmid, M., Welchowski, T., Wright, M.N. & Berger, M. (2020). Discrete-time survival forests with hellinger distance decision trees. *Data Mining and Knowledge Discovery*, **34**, 812–832. 9, 175, 176

Schumacher, M., Graf, E. & Gerds, T. (2003). How to assess prognostic models for survival data: a case study in oncology. *Methods of information in medicine*, **42**, 564–571. 3, 57

Serfling, R.J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons. 25

Shao, J. (2003). *Mathematical statistics*. Springer Science & Business Media. 27, 29, 93, 105, 159

Singer, J.D. & Willett, J.B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of educational statistics*, **18**, 155–195. 9

Sonabend, R., Bender, A. & Vollmer, S. (2022). Avoiding c-hacking when evaluating survival distribution predictions with discrimination measures. *Bioinformatics*, **38**, 4178–4184. 21, 87

Song, X., Zhou, X.H. & Ma, S. (2012). Nonparametric receiver operating characteristic-based evaluation for survival outcomes. *Statistics in medicine*, **31**, 2660–2675. 57, 121

Stephenson, D.B. (2000). Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting*, **15**, 221–232. 3

Suresh, K., Severn, C. & Ghosh, D. (2022). Survival prediction models: an introduction to discrete-time modeling. *BMC medical research methodology*, **22**, 207. 14

Tutz, G. & Schmid, M. (2016). *Modeling discrete time-to-event data*. Springer. 9

Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B. & Wei, L.J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, **30**, 1105–1117. 5, 12, 22, 24, 25, 57, 90, 183

Wang, J.G. (1987). A note on the uniform consistency of the kaplan-meier estimator. *The Annals of Statistics*, **15**, 1313–1316. 11

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J.M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**, 1113–1120. 46

Wright, M.N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, **77**, 1–17. 176

Yang, Y. & Zou, H. (2013). A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions. *Statistics and its Interface*, **6**, 167–173. 2

Yates, J.F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, **30**, 132–156. 6