# A Simulation-based comparison of the predictive accuracy of the random survival forest and the lasso-regularized Cox Model in Survival Analysis

---

Willem Van Der Merwe

*Supervisor(s):*
Dr. Alphonce Bere


A research proposal submitted in partial fulfillment of the requirements for the degree of Master of Science, Artificial Intelligence

in the

Faculty of Science
University of the Witwatersrand, Johannesburg

24 April 2024

**Using this template**

The template below has been constructed in line with the Faculty of Science requirements as well as the conventions of the School. Guidance on each section can be found in the blue boxes at the start of the section. Note that conventions and preferences for structuring a research proposal can vary from discipline to discipline and supervisor to supervisor. Both the template and the guidance on the different sections are suggestions for structuring the proposal. If your supervisor has a different preferred convention or template, it is recommended that you consult with them regarding the differences.

Further details on using the template can be found in Appendix B.

You should ensure that you either delete or comment out the blue boxes before submission of the proposal.

# Declaration

I, Willem Van Der Merwe, declare that this proposal is my own, unaided work. It is being submitted for the degree of Master of Science, Artificial Intelligence at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

Signature

Willem Van Der Merwe

24 April 2024

The declaration is an important formal requirement. Ensure that you upload an image of your signature and that you change the file name in the main.tex file to include it.

# *Abstract*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

The abstract is a brief informative summary of the proposed research. It can be read independently and should

- locate the proposed research within the background relevant to it,
- state the research question or aim,
- briefly describe the proposed methods for answering the question or achieving the aim, and,
- emphasise the contribution that the proposed research will make to current research in the field.

It is recommended that your abstract is no more than 300 words.

# Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque.

The acknowledgements section allows you to thank those who contributed to the preparation of the proposal. It is usual to acknowledge supervision, financial assistance or funders, and any special facilities provided for the research.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Literature review

In medical studies, the paradigm of survival analysis is used to determine outcome events based on patient survival data. Due to the censoring complexities and high dimensionality these datasets often entail, formal statistical approaches have been developed chronologically, each iteration improving and building upon fundamental statistical properties of survival data and the underlying result interpretation.

### 1.1.1 Background

Survival analysis is used to examine the time until the occurrence of an event like disease relapse. A major challenge in this area is handling censored data, where the event information is incomplete. Censoring can be of different types: right-censored data is when the event has not occurred by the end of the observation period; left-censored data is when the event occurred before the study began; and interval-censored data is when the event occurred between two observed times. In order to analyse such data, statistical methods have been developed. Non-parametric methods like the Kaplan-Meier estimator and the Logrank test do not assume any specific distribution for the time-to-event data, making them robust against misspecifications of the event-time distribution. Parametric methods like the Exponential and Weibull models assume a known distribution that models the time-to-event data. They are typically more precise, at the risk of introducing bias when the assumed distribution is wrong. The Proportional Hazards Model, which can be used

in both semi-parametric (Cox model) and parametric forms, is employed to estimate the hazard ratio, which is a measure of effect size regarding the time to event. An example of this is found with the study [1] could be that of analysing motion-sickness data, where survival functions are estimated and treatments are compared. For instance, studies may compare the time until onset of motion sickness under different conditions to assess treatment effectiveness.

Traditional statistical methods require explicit programming and often suffer from user bias in variable selection. Machine Learning (ML) operates under a paradigm where algorithms autonomously identify patterns in large data sets, which potentially increases accuracy and efficiency. [2] Show that, the literature on ML in orthopaedics, predominantly composed of preliminary studies, frequently lacks depth in addressing complex ML concepts and falls short in providing comprehensive frameworks for result interpretation. Deep Learning, a prominent subset of ML, utilises neural networks to process both structured and unstructured data, enhancing the capability to handle diverse data types like images and texts. [2] [3] Also show out of their methodical study selection process that only a handful of studies have attempted such comparisons at an acceptable standard, while most studies focus predominantly on machine learning techniques neglecting the broader spectrum of statistical methods. Furthermore [3] point out authors often omit interaction terms and non-linear covariate effects which are essential components for enhancing model robustness and accuracy. The predominance of studies failed to relax the proportional hazards assumption which underscores a critical oversight in adapting models to more complex datasets. [3] show that there is a need for comprehensive methodological improvements and enhanced reporting standards to ensure reproducibility and a fair assessment of method capabilities[3]

**Important issues in comparative simulation studies**

Simulation studies are a crucial statistical tool used for evaluating and comparing different statistical methods, particularly when analytic solutions are hard or impossible to achieve. These studies generate data through pseudo-random sampling from known probability distributions, enabling researchers to empirically test the behaviour of statistical methods under varied scenarios. Common uses include validating new statistical methods, ensuring accuracy in mathematical models and

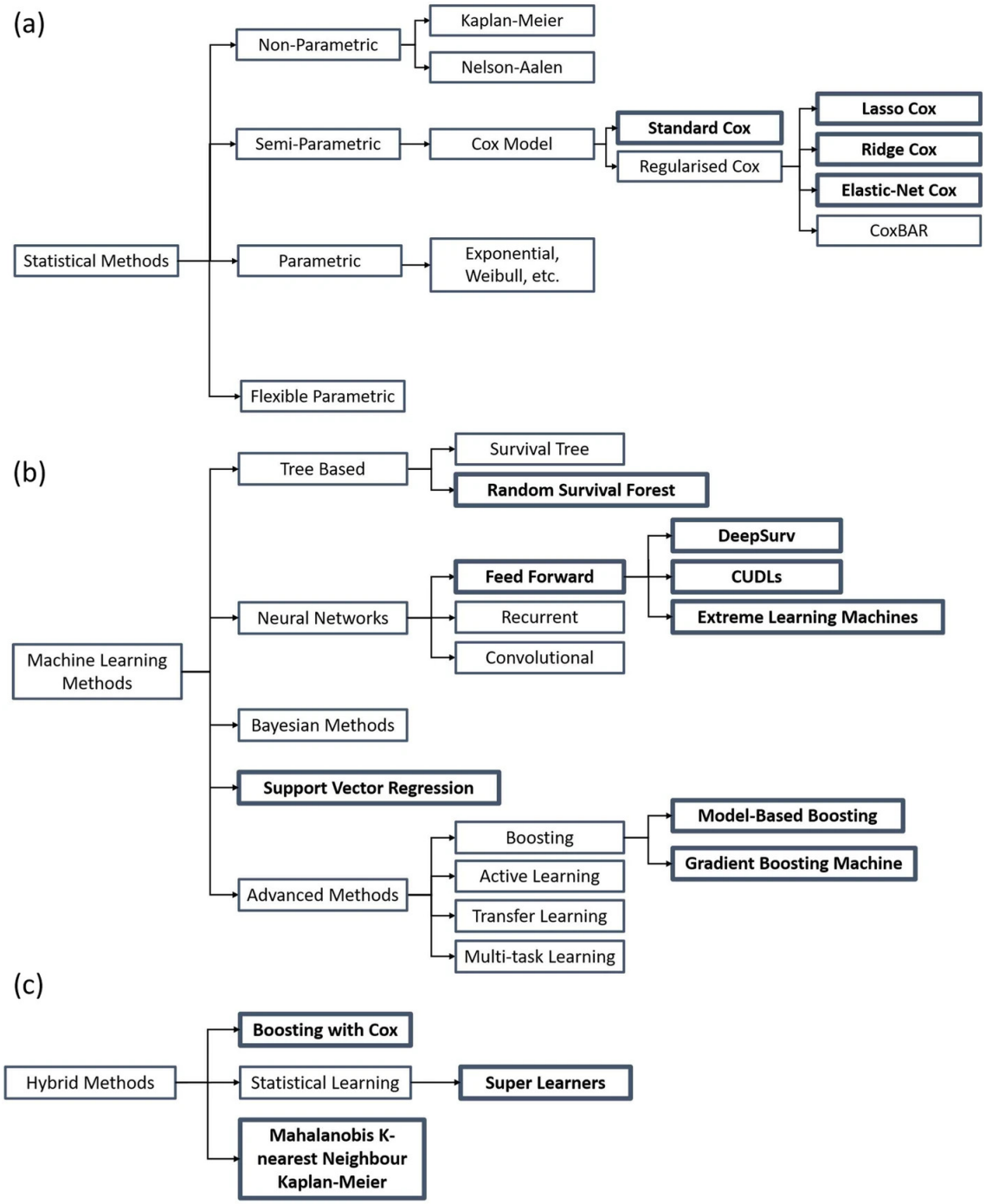



FIGURE 1.1: Shows the breakdown of methods analysis preformed by **[<empty citation>]** during a method review. The Study ran a literature selection process based on qualitative and quantitative metrics of methodology used in studies.

code, and comparing the effectiveness of various approaches. Particularly in medical statistics, simulation studies help in designing experiments, determining sample sizes, and estimating power under specific assumptions about data generation [4]. Despite their widespread use, many statisticians face challenges in properly conducting simulation studies due to a lack of understanding and experience [4]. Common issues include inadequate design and reporting that lead to uncritical acceptance of results. This lack of rigour can result in misleading conclusions, for example the variability introduced by different sets of random numbers in Monte Carlo simulations that is sometimes ignored.

[3] Found a notable scarcity of quality comparative research between statistical and machine learning methods. Predominantly, these studies focus on machine learning techniques while traditional statistical methods are often neglected.

For instance, it was common for some authors to overlook the inclusion of interaction terms and non-linear covariate effects in the Cox model as well as time-dependent effects which are key elements for effectively handling complex datasets.

The reporting standards of the reviewed studies were also generally poor. Important details such as data-generating mechanisms (DGMs), estimands, and method implementations are frequently underreported, which impedes the reproducibility of the research and the ability to conduct fair comparisons between methods.

[3] Also pointed out that a significant bias could be observed in the selection of DGMs, which tend to favour machine learning approaches, especially in scenarios where the number of variables exceeds the sample size. This predisposition can lead to biassed results unless the study incorporates specific statistical variable selection techniques that are suited for high-dimensional data.

Additionally, the prevalent use of the C-index as the sole performance metric, without accounting for calibration is noted. By relying solely on this metric results analysis may not provide a complete picture of the model's predictive accuracy over time, particularly when the proportional hazards assumption is not valid.

Finally, [3] exclaim that there is a concerning lack of expertise in implementing complex statistical methods thoroughly. This deficiency often results in potentially misleading outcomes that do not genuinely reflect the true performance capabilities of the methods being compared. The findings underscore the need for improved methodological rigour and enhanced collaboration among researchers to ensure that both statistical and machine learning methods are implemented to their

full potential and evaluated fairly.

As a framework [5] formalises the use of indicators, defined as "questionable research practices" (QRP) which indicate faulty research methods used widely throughout simulation studies, which should get necessary attention. The QRP's are categorised during phases of comparative simulation, namely design phase, execution phase and reporting phase. [5] Labels specific components of simulation studies, an example being D1 which references "the data-generating process", and cross correlates with other components to define QRP occurrence and relationships.

| Tag | Related | Type of QRP |
|---|---|---|
| *Design* | | |
| D1 | E1, R1 | Not/vaguely defining objectives of simulation study |
| D2 | E2, R1 | Not/vaguely defining data-generating process |
| D3 | E3, E4, R1 | Not/vaguely defining which methods will be compared and how their parameters are specified |
| D4 | E1, E5, R1 | Not/vaguely defining estimands of interest |
| D5 | E1, E5, R1 | Not/vaguely defining evaluation criteria |
| D6 | E6, R1 | Not/vaguely defining how to handle missing values (for example, due to non-convergence of methods) |
| D7 | E7, E8, R3 | Not justifying number of simulations |
| *Execution* | | |
| E1 | D1, R2 | Changing objective of the study to achieve desired outcomes |
| E2 | D2, R2 | Adapting data-generating process to achieve desired outcomes |
| E3 | D3, R2 | Adding/removing comparison methods to achieve desired outcomes |
| E4 | D3, R2 | Selective tuning of method hyperparameters to achieve desired outcomes |
| E5 | D4, D5, R2 | Choosing evaluation criteria to achieve desired outcomes |
| E6 | D6, R2 | Adapting inclusion/exclusion/imputation rules to achieve desired outcomes |
| E7 | D7, R3 | Choosing number of simulations to achieve desired outcomes |
| E8 | D7, R3 | Choosing random number generator seed to achieve desired outcomes |
| *Reporting* | | |
| R1 | D1–D6 | Justifying design decisions which lead to desired outcomes *post hoc* |
| R2 | E1–E6 | Selective reporting of results from simulations that lead to desired outcomes |
| R3 | D7, E7, E8 | Failing to report Monte Carlo uncertainty |
| R4 | | Failing to assure computational reproducibility (for example, not sharing code and sufficient details about computing environment) |
| R5 | | Failing to assure replicability (for example, not sufficiently reporting design and execution methodology) |

FIGURE 1.2: shows the outline of QRP classes and instances

**Applied examples of simulation studies**

[6] evaluated and compared the effectiveness of Cox regression analysis (CRA) and random survival forests (RSF) through both simulated and actual breast cancer data scenarios. Initially, the study utilised Monte Carlo simulations to assess how both methods performed across various sample sizes, specifically observing their performance metrics based on Harrell's concordance index. The results indicated that CRA consistently outperformed RSF under simulation conditions, particularly when using the concordance index for evaluation. Following the simulations, the methods were applied to a real dataset comprising 279 breast cancer patients to identify major risk factors influencing disease-free survival (DFS). In this practical application, RSF slightly edged out other methods, offering marginally better performance according to the concordance index when using the approximate log-rank splitting rule.

Approximate log-rank splitting rule

CRA was noted for its predictive accuracy across different sample sizes, making it suitable for a broad range of survival data applications. Conversely, RSF was recommended for its interpretative power, especially beneficial in handling complex datasets where multiple survival trees are analysed. [6]

Furthermore, [7] shows a comparison between a machine learning method, termed survival neural networks (SNNs) and compared it with the Cox proportional hazards model, using clinical trial data for survival outcomes. The models are formulated subject to the European Osteosarcoma intergroup trial data, which is used as the foundation for the synthetic data generation that would ultimately be used for simulation training. The original dataset contains various instances of censoring, and the authors approach this issue, by segmenting the datasets into samples with degrees of censoring present (20

where j = 1, 2, . . . , J are the nodes in the input layer, h = 1, 2, . . . , H are the nodes in the hidden layer, and w are the weights of the network. The training is performed with training sets and validation sets, using cross-validation and hyperparameter tuning. Furthermore, evaluation techniques are explored, specifically looking at factors like discrimination (C-index) by using the average time dependant nonlinear prognostic index,

Accuracy (Brier) interpreted in continuous form as the integrated Brier score for prediction error over the total period,

And lastly Miscalibration (mean squared error) for censored groups. The results indicated comparable predictive performance but highlighted a lack of accuracy for calibration measures with SNNs. The authors point out that although machine learning techniques are attractive for survival analysis scenarios because of the ability to model interactions and nonlinearities with a no assumption approach, the robustness of the Cox model, regarding ease of implementation as well as interpretability of covariates makes it formidable in situations where limited sample sizes and variables are available. The paper ties in nicely with the other literature in support of the need for clear and better implementation of calibration metrics specifically with machine learning models, and caution against indiscriminate application of these models.

### 1.1.2 Cox's Proportional Hazards Model

In the seminal work by [8], the cox model is introduced as an extension to prior work formalised as the Kaplan-Meier estimator, by exploring time-to-event data (life tables). The major benefit is that it addresses censored data, which is a known concept in survival analysis, that there is missing information within the data, specifically, event occurrence without observation on a continuous time scale. The proposition consists of covariates, known as attributes regarding a unit in a distribution of data, which is associated with a coefficient $\beta$ scaling the impact of said covariates; this product is then bound by the baseline hazard h0(t).

Hazard being the estimated conditional probabilities, in line with the observed conditional frequencies of events or simply the risk of event occurrence at a specific time. An assumption of the Cox model is the proportional hazards assumption, suggesting that the hazard ratios for different covariates remain constant over time we see this for two events observations,

This is an operational Assumption and a limitation as this is not always true for survival data. The model can handle censoring, by adjusting the likelihood function for observations where event occurrence did not happen in a particular continuous time frame, and by maximising the likelihood of all observed events, it is possible to estimate the coefficients that could work the best under the Cox formulation.

**Time-Dependant Covariates**

The Cox model incorporates both time-independent and time-dependent covariates. Time-dependent covariates can change over the time, such as Z2(t) = Z1*t. This flexibility allows the model to handle scenarios where hazards are not proportional, which extends its applicability. Relative risk is represented as exp(Z(t)'), showing how risk changes with time and covariate values and depending on the coefficients, the relative risk in a treatment group can increase, decrease, or remain constant over time. External covariates are variables that are independent of the subject's survival process. Whereas internal covariates are variables that might influence and be influenced by survival. Different approaches for modelling survivor functions are required for external and internal covariates due to their nature.

**Discrete vs Continuous Model**

Discrete Models address survival data that is categorical or not continuously distributed while the continuous model proposed by cox, utilises continuous data to model hazard functions. The survivor function can also be represented as a product integral which accommodates both discrete and continuous survival data. This approach allows for the unified handling of mixed data types in survival analysis. For the continuous model, estimation is based on maximising the conditional likelihood across observed failure times, while the discrete model uses a logistic framework for estimation, treating survival as a sequence of binary outcomes.

**Likelihood Function**

The various approaches to the Cox model handle data ties and time-dependent covariates differently, ....recommended specific methods based on the data structure (e.g., number of ties). The concept of partial likelihood is particularly important as it provides a way to focus on relevant factors in the presence of complex data types, enhancing both the theoretical understanding and practical application of the Cox model.

### 1.1.3 Lasso Regularisation And Variable Selection

### 1.1.4 Random Surivival Forest

### 1.1.5 Methods For Data Generating Mechanisms For Simulation

### 1.1.6 Methods For Model Execution And Computation

### 1.1.7 Methods For Model Evaluation And Result Interpretation

In evaluating the performance of survival prediction models, it is crucial to employ metrics that not only assess accuracy but also ensure fairness by minimising bias. This involves selecting evaluation measures that effectively balance calibration (the agreement between predicted and observed outcomes) and discrimination (the model's ability to distinguish between different outcomes). (1…) (2….) provides a guidelines and comprehensive framework for assessing model performance against true event times, ensuring that the models are both fair and accurate across different scenarios. These are essential for advancing survival analysis in ethically sensitive domains, thereby supporting more reliable and equitable outcomes.

#### C-Index

The C-index, or concordance index, is a vital statistical tool used to evaluate the predictive accuracy of survival models, quantifying their ability to correctly rank the order of patient outcomes based on their predicted risks. It is formally defined as the proportion of all "comparable" pairs of subjects where the predictions and actual outcomes are in agreement. A pair is considered comparable if it is possible to determine the order of their events, that is, who experienced the event first.

$$C\text{-index} = \frac{\sum_{i,j} \mathbf{1}(t_i < t_j) \cdot \mathbf{1}(\eta_i > \eta_j) \cdot \delta_i}{\sum_{i,j} \mathbf{1}(t_i < t_j) \cdot \delta_i}$$

Where, $\mu_i$ is the risk score of subject $i$, $\delta_i$ is the event indicator being 1 in case of occurance Variants of the C-index include; time-independent C-index (Cti) which Uses the negative of predicted time or survival probability at a specified time. It assesses if the sequence of actual event times matches the predicted times, time-dependent C-index (Ctd) introduced by Antolini et al. (2005), this version accounts

for varying amounts of censoring over time, attempting to provide a more accurate assessment by being a weighted average of time-dependent area under the curve (AUC) scores. The C-index can be biassed upwards with a high level of censoring in the data. This issue is discussed by Antolini et al. (2005) and addressed through the Ctd rule, although it is not a proper scoring rule (Rindt et al., 2022). The C-index, while useful, does not always align with other metrics such as the Mean Absolute Error (MAE). It is possible for a model to have a high C-index (accurately ranking the order of events) while still having large discrepancies in the actual predicted times of those events.

**Brier Score and Integrated Brier Score**

The **Brier Score** and the Integrated Brier score (IBS) are essential metrics used to evaluate the accuracy and reliability of survival models. Both scores measure the calibration and discrimination capabilities of a model, which are crucial for producing unbiased and precise predictions in survival analysis. Below is a detailed explanation of both metrics. The Brier score is a measure used to evaluate the accuracy of probabilistic predictions. It is essentially the mean squared error between the observed outcomes and the predicted probabilities at a specific time $t^*$.

$$BS_{t^*}(VU, \hat{S}(t^*|\cdot)) = \frac{1}{|VU|} \sum_{[\tilde{x}_i, d_i] \in VU} (I[d_i \leq t^*] - \hat{S}(t^*|\tilde{x}_i))^2$$

Where, $VU$ is the validation set, $\hat{S}(t^*|\cdot))$ is the predicted survival probability at $t^*$ and $I[d_i \leq t^*]$ is the event indicator. A perfect model, which perfectly predicts whether events happen by time t* (predicting 1s and 0s accurately), would have a Brier score of 0. A model that always predicts a 50

$$IBS(\tau, VU, \hat{S}(\cdot|\cdot)) = \frac{1}{\tau} \int_0^\tau BS_t(VU, \hat{S}(t|\cdot))\, dt$$

Where, $\tau$ is the maximum event time, $BS_t$ is the brier score at time t. IBS is particularly useful for survival prediction models where it is important to assess model performance comprehensively across time rather than at a single time point. It gives an average score that reflects the overall performance of the model across the specified time interval. For censored data, the Inverse Probability Censoring Weight

(IPCW) technique is often used in conjunction with IBS to adjust the contributions of censored subjects. This method helps to ensure that the model's performance is not unduly biassed by the censoring. IBS is considered a proper scoring rule if the censoring distribution is estimated correctly, meaning it incentivizes truthful forecasting and accurately reflects the model's predictive capabilities. IBS can be particularly impactful in clinical settings where decisions might depend on accurate, time-specific survival probabilities, such as deciding on conservative treatments based on predicted long-term survival chances.

**Hosmer-Lemeshow Calibration (1-Calibration) and D-Calibration**

1-Calibration, also known as Hosmer-Lemeshow calibration, is a statistical test used to evaluate the calibration of a model at a specific time point t*. It measures how well the predicted probabilities of an event (e.g., failure, death) occurring by t* match the actual proportion of those events in the dataset. This test is particularly useful in contexts where predictions need to be reliable at specific critical thresholds. It works by sorting all subjects for the predicted probabilities at time t*. These probabilities are then grouped into a predefined number of groups or bins (typically 10). For each bin, the expected number of events is calculated based on the predicted probabilities, and this is compared to the actual number of events that occurred.

$$\mathrm{HL}(VU, \hat{S}(t^*|\cdot)) = \sum_{j=1}^{B} \frac{(O_j - n_j \bar{p}_j)^2}{n_j \bar{p}_j (1 - \bar{p}_j)}$$

Where, $B$ is the number of bins, $O_j$ is the observed number of events in bin j, $\hat{p}_j$ is the average predicted probability in bin j. A low value of the Hosmer-Lemeshow statistic suggests that the model's predictions are well-calibrated — i.e., the predicted probabilities of survival match the actual rates observed. The statistic follows a chi-squared distribution, allowing for the derivation of a p-value to assess the significance of the results. A model is considered well-calibrated at the chosen significance level if the p-value is greater than 0.05.

**D-Calibration**

D-Calibration extends the concept of 1-calibration over a range of time points or across different distributions of time points, providing a more comprehensive measure of a model's accuracy. It measures the consistency of predicted probabilities across a range of outcomes within a dataset. It assesses whether the distribution of predicted probabilities (over time or across conditions) matches the observed distribution of outcomes. Predicted probabilities are checked across a range of values. For each interval [a,b] within the probability range [0,1], the proportion of subjects with predicted probabilities within this range is compared to the actual proportion of events occurring in this interval.

$$D_{\Theta}([a,b]) = \{[x_i, d_i, \delta_i = 1] \in D | \hat{S}_{\Theta}(d_i) \in [a,b]\}$$

The proportion of subjects in each interval is expected to match the width of the interval $b - a$. For instance, for the interval [0.1, 0.2], approximately 10

**Mean Absolute Error**

## 1.2 Problem Statement

Many studies have compared machine learning with traditional statistics, yet comprehensive simulation-based comparisons are scarce. This gap may lead to biases and sometimes questionable practices, affecting the validity of findings.

## 1.3 Research Aims and Objectives

### 1.3.1 Research Aims

Perform a comparative analysis of survival models using both simulated and real datasets to identify model robustness and effectiveness, adhering to formal frameworks and avoiding common pitfalls outlined in the literature.

### 1.3.2 Objectives

1. Source a Practical Dataset: Acquire a dataset with clear metrics and pre-analyzed statistics for straightforward applicability in survival analysis. This dataset should comply with standards.

2. Dataset Analysis: Run analysis on dataset metrics and formulate appropriate data-generating methods to match distribution.

3. Apply Data Generating Methods: Utilise standard libraries to generate simulated data that closely replicates the statistical properties of the real dataset.

4. Construct Survival Models:

   (a) Random Survival Forest Model: Develop and apply this model using both the real and simulated datasets.

   (b) Cox Proportional Hazards Model: Similarly, develop and apply this model with both datasets.

5. Evaluate and Visualise Predictions: Use common survival analysis metrics for evaluation and employ visualisation tools from survival libraries to illustrate the results effectively.

## 1.4 Limitations

1. **Scope and Focus:** This study primarily focuses on the application and evaluation of established survival analysis methods and their existing extensions as documented in the literature. The comparative aspect of our study does not aim to modify the foundational algorithms of these methods; rather, it seeks to implement and test these pre-existing models in a new dataset context with established frameworks.

2. **Redundancy in Literature:** Furthermore, comprehensive comparative studies like those conducted. please add source have already evaluated these methods extensively. These studies provide a solid foundation of knowledge regarding the performance and limitations of traditional and modified survival analysis models across various types of data.

3. **Innovation vs. Application:** Consequently, this proposal does not venture to innovate on the algorithmic core of these methods. Instead, it is designed to apply these well-established techniques to derive insights from specific datasets, thereby contributing to empirical evidence and practical applications rather than theoretical advancements.

## 1.5 Overview

In addressing the noted shortcomings in comparative simulation studies, this literature review methodically examines simulation work in segments relevant to each section of the study. We begin with an overview of the Cox method and its various extensions, illustrating how these foundational techniques are implemented. Following that, we explore proofs and extensions of the Lasso method, which build on the base Cox method, enhancing its predictive power and flexibility. The discussion then moves to Random Survival Forests (RSF), detailing recent advancements in RSF algorithms that provide a solid reference for current implementations. Two comparative studies are highlighted; these utilise simulations to evaluate the methods mentioned above, offering insights into their practical applications and effectiveness. Finally the last section categorises the literature into subgroups that align with the specific components of the proposed research framework…..., facilitating easy reference and integration into the research design and methodology in Chapter 2, ensuring a coherent and structured approach to applying these methods in proposed study.

# Chapter 2

# Research Methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat.

The methodology chapter provides an overview and description of all the important elements of *how* the research will be carried out. It differs from the methodology chapter of the final report insofar as it includes discussion

- both of available and chosen methods, and,
- of any foreseeable limitations of the methods.

## 2.1 Research design

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque.Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elitposuere quam, quis adipiscing urna leo nec orci.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu.

Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus.

This section provides a brief, high-level description of the broad research method to be used in carrying out the research. This research method will vary from field to field. The section should

- identify the broad research method, and,
- provide a brief description of the method.

## 2.2 Methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat.

This section describes and *motivates* the instruments and procedure to be used in carrying out the research. These should not be discussed in a chronological way (e.g. first, this step will be taken, and then this step will be taken, etc.), but should instead be grouped together systematically. A more detailed discussion of the relevant algorithms or models should be included here. The discussion should show an understanding of the (available) methods and put forward a *motivation for those chosen*.

## 2.3 Limitations

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros.

This section identifies limitations with the methodology. These limitations might have a number of sources. For example, constraints on time and/or computational power could limit the number of methods or models tested. In this section, you should identify and discuss any such limitations that are foreseeable at the time of proposal.

## 2.4 Ethical Considerations

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

This section assesses the need for ethical clearance for the proposed research.In cases in which no ethical clearance is required, the section will indicate as such and include a brief motivation for this. In cases in which ethical clearance is required, the section will indicate as such, include a brief motivation for this, and describe the procedure that will be followed for obtaining clearance.

# Chapter 3

# Schedule of Work

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat.

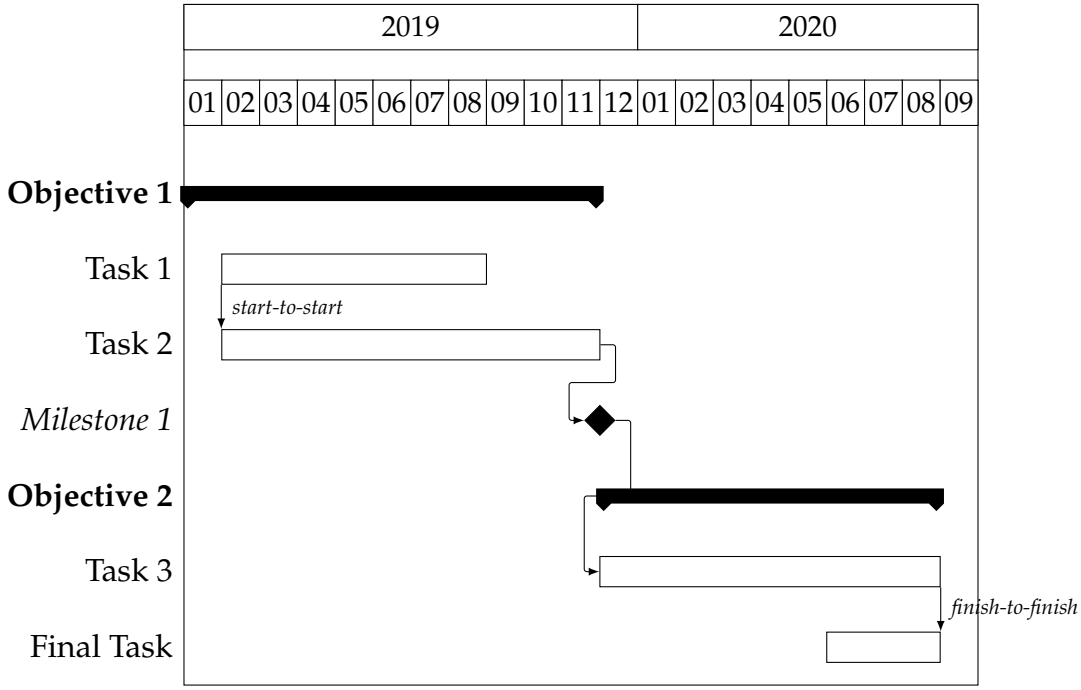
The chapter puts forward the planned schedule of work for the research. In it, you should this chapter, you should provide an overview and *discussion* of the planned schedule for the research. Your discussion should

- present and discuss the schedule of work,
- identify any foreseeable difficulties to carrying out the project according to the planned schedule,
- provide a brief conclusion to the chapter.

## 3.1 Schedule of Work

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus

justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.




Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

## 3.2 Potential Difficulties

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis.

# Chapter 4

# Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

The final chapter provides a retrospective overview of the proposed research and is usually quite brief in relation to the rest of the report. A good way to structure this section is to proceed from the more specific to the more general, by

- reiterating the relevant gaps in the current literature,
- reiterating the aim of the research,
- emphasising the contribution that the proposed research will make to current research in the field.

# Appendix A

# Appendix Title

## A.1 Main Section

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu.

The appendices are sections in which complicated mathematical or other formulae, descriptions of experiments or apparatus, and any other specialised or lengthy material such as computer programme listings, copies of spectra or other instrumental outputs are found.

# Appendix B

# Using this Template

In the following appendix, some of the guidance from the original template is included. This includes the

- files and folders included in the template,
- guidance on filling the main.tex file with your information,
- guidance on including references, tables, figures, and mathematical formulae.

## B.1 What this Template Includes

### B.1.1 Folders

This template comes as a single zip file that expands out to several files and folders. The folder names are mostly self-explanatory:

**Appendices** – this is the folder where you put the appendices. Each appendix should go into its own separate `.tex` file. An example and template are included in the directory.

**Chapters** – this is the folder where you put the thesis chapters. Each chapter should go in its own separate `.tex` file.

**Figures** – this folder contains all figures for the thesis. These are the final images that will go into the thesis document.

### B.1.2 Files

Included are also several files, most of them are plain text and you can see their contents in a text editor. After initial compilation, you will see that more auxiliary files are created by LaTeX or BibTeX and which you don't need to delete or worry about:

**example.bib** – this is an important file that contains all the bibliographic information and references that you will be citing in the thesis for use with BibTeX. You can write it manually, but there are reference manager programs available that will create and manage it for you. Bibliographies in LaTeX are a large subject and you may need to read about BibTeX before starting with this. Many modern reference managers will allow you to export your references in BibTeX format which greatly eases the amount of work you have to do.

**MastersDoctoralThesis.cls** – this is an important file. It is the class file that tells LaTeX how to format the thesis.

**main.pdf** – this is your beautifully typeset thesis (in the PDF file format) created by LaTeX. It is supplied in the PDF with the template and after you compile the template you should get an identical version.

**main.tex** – this is an important file. This is the file that you tell LaTeX to compile to produce your thesis as a PDF file. It contains the framework and constructs that tell LaTeX how to layout the thesis. It is heavily commented so you can read exactly what each line of code does and why it is there. After you put your own information into the *THESIS INFORMATION* block – you have now started your thesis!

Files that are *not* included, but are created by LaTeX as auxiliary files include:

**main.aux** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file.

**main.bbl** – this is an auxiliary file generated by BibTeX, if it is deleted, BibTeX simply regenerates it when you run the `main.aux` file. Whereas the `.bib` file contains all the references you have, this `.bbl` file contains the references you have actually cited in the thesis and is used to build the bibliography section of the thesis.

**main.blg** – this is an auxiliary file generated by BibTeX, if it is deleted BibTeX simply regenerates it when you run the main `.aux` file.

**main.lof** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It tells LaTeX how to build the *List of Figures* section.

**main.log** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It contains messages from LaTeX, if you receive errors and warnings from LaTeX, they will be in this `.log` file.

**main.lot** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It tells LaTeX how to build the *List of Tables* section.

**main.out** – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file.

So from this long list, only the files with the `.bib`, `.cls` and `.tex` extensions are the most important ones. The other auxiliary files can be ignored or deleted as LaTeX and BibTeX will regenerate them.

## B.2 Filling in Your Information in the `main.tex` File

You will need to personalise the thesis template and make it your own by filling in your own information. This is done by editing the `main.tex` file in a text editor or your favourite LaTeX environment.

Open the file and scroll down to the third large block titled *THESIS INFORMATION* where you can see the entries for *University Name*, *Department Name*, etc ...

Fill out the information about yourself and institution. You can also insert web links, if you do, make sure you use the full URL, including the `http://` for this. If you don't want these to be linked, simply remove the `\href{url}{name}` and only leave the name.

When you have done this, save the file and recompile `main.tex`. All the information you filled in should now be in the PDF, complete with web links. You can now begin your thesis proper!

## B.3 Thesis Features and Conventions

To get the best out of this template, there are a few conventions that you may want to follow.

One of the most important (and most difficult) things to keep track of in such a long document as a thesis is consistency. Using certain conventions and ways of doing things (such as using a Todo list) makes the job easier. Of course, all of these are optional and you can adopt your own method.

### B.3.1 References

The `biblatex` package is used to format the bibliography and inserts references such as this one [**Reference1**]. The options used in the `main.tex` file mean that the in-text citations of references are formatted with the author(s) listed with the date of the publication. Multiple references are separated by semicolons (e.g. [**Reference2**, **Reference1**]) and references with more than three authors only show the first author with *et al.* indicating there are more authors (e.g. [**Reference3**]). This is done automatically for you. To see how you use references, have a look at the `Chapter1.tex` source file. Many reference managers allow you to simply drag the reference into the document as you type.

Scientific references should come *before* the punctuation mark if there is one (such as a comma or period). The same goes for footnotes[1]. You can change this but the most important thing is to keep the convention consistent throughout the thesis. Footnotes themselves should be full, descriptive sentences (beginning with a capital letter and ending with a full stop). The APA6 states: "Footnote numbers should be superscripted, [...], following any punctuation mark except a dash." The Chicago manual of style states: "A note number should be placed at the end of a sentence or clause. The number follows any punctuation mark except the dash, which it precedes. It follows a closing parenthesis."

The bibliography is typeset with references listed in alphabetical order by the first author's last name. This is similar to the APA referencing style. To see how LaTeX typesets the bibliography, have a look at the very end of this document (or just click on the reference number links in in-text citations).

[1] Such as this footnote, here down at the bottom of the page.

**A Note on bibtex**

The bibtex backend used in the template by default does not correctly handle unicode character encoding (i.e. "international" characters). You may see a warning about this in the compilation log and, if your references contain unicode characters, they may not show up correctly or at all. The solution to this is to use the biber backend instead of the outdated bibtex backend. This is done by finding this in `main.tex`: *`backend=bibtex`* and changing it to *`backend=biber`*. You will then need to delete all auxiliary BibTeX files and navigate to the template directory in your terminal (command prompt). Once there, simply type `biber main` and biber will compile your bibliography. You can then compile `main.tex` as normal and your bibliography will be updated. An alternative is to set up your LaTeX editor to compile with biber instead of bibtex, see here for how to do this for various editors.

### B.3.2 Tables

Tables are an important way of displaying your results, below is an example table which was generated with this code:

```
\begin{table}
\caption{The effects of treatments X and Y on the four groups studied.}
\label{tab:treatments}
\centering
\begin{tabular}{l l l}
\toprule
\tabhead{Groups} & \tabhead{Treatment X} & \tabhead{Treatment Y} \\
\midrule
1 & 0.2 & 0.8\\
2 & 0.17 & 0.7\\
3 & 0.24 & 0.75\\
4 & 0.68 & 0.3\\
\bottomrule\\
\end{tabular}
\end{table}
```

TABLE B.1: The effects of treatments X and Y on the four groups studied.

| Groups | Treatment X | Treatment Y |
|---|---|---|
| 1 | 0.2 | 0.8 |
| 2 | 0.17 | 0.7 |
| 3 | 0.24 | 0.75 |
| 4 | 0.68 | 0.3 |

You can reference tables with `\ref{<label>}` where the label is defined within the table environment. See `Chapter1.tex` for an example of the label and citation (e.g. Table B.1).

### B.3.3 Figures

There will hopefully be many figures in your thesis (that should be placed in the *Figures* folder). The way to insert figures into your thesis is to use a code template like this:

```
\begin{figure}
\centering
\includegraphics{Figures/Electron}
\decoRule
\caption[An Electron]{An electron (artist's impression).}
\label{fig:Electron}
\end{figure}
```

Also look in the source file. Putting this code into the source file produces the picture of the electron that you can see in the figure below.

Sometimes figures don't always appear where you write them in the source. The placement depends on how much space there is on the page for the figure. Sometimes there is not enough room to fit a figure directly where it should go (in relation to the text) and so LaTeX puts it at the top of the next page. Positioning figures is the job of LaTeX and so you should only worry about making them look good!

FIGURE B.1: An electron (artist's impression).

Figures usually should have captions just in case you need to refer to them (such as in Figure B.1). The `\caption` command contains two parts, the first part, inside the square brackets is the title that will appear in the *List of Figures*, and so should be short. The second part in the curly brackets should contain the longer and more descriptive caption text.

The `\decoRule` command is optional and simply puts an aesthetic horizontal line below the image. If you do this for one image, do it for all of them.

LaTeX is capable of using images in pdf, jpg and png format.

### B.3.4 Typesetting mathematics

If your thesis is going to contain heavy mathematical content, be sure that LaTeX will make it look beautiful, even though it won't be able to solve the equations for you.

The "Not So Short Introduction to LaTeX" (available on CTAN) should tell you everything you need to know for most cases of typesetting mathematics. If you

need more information, a much more thorough mathematical guide is available from the AMS called, "A Short Math Guide to LaTeX" and can be downloaded from: ftp://ftp.ams.org/pub/tex/doc/amsmath/short-math-guide.pdf

There are many different LaTeX symbols to remember, luckily you can find the most common symbols in The Comprehensive LaTeX Symbol List.

You can write an equation, which is automatically given an equation number by LaTeX like this:

```
\begin{equation}
E = mc^{2}
\label{eqn:Einstein}
\end{equation}
```

This will produce Einstein's famous energy-matter equivalence equation:

$$E = mc^2 \tag{B.1}$$

All equations you write (which are not in the middle of paragraph text) are automatically given equation numbers by LaTeX. If you don't want a particular equation numbered, use the unnumbered form:

```
\[ a^{2}=4 \]
```

Guide written by —
Sunil Patel: www.sunilpatel.co.uk
Vel: LaTeXTemplates.com