Norwegian University
of Life Sciences

Master's Thesis 2023 30 ECTS
Faculty of Science and Technology

A Comparison of Methods for Modelling Survival Time for Cancer Patients

Mikko Johan Vatterdal Rekstad
Data Science

ABSTRACT

In this thesis, we used three types of survival-analysis models to model the overall survival time for patients suffering from rectal cancer and head and neck cancer. These models were Cox proportional hazards, Aalen's additive fitter and accelerated failure time models. The goal was to compare the performance in terms of the measured concordance index and Brier scores.
The performance metrics were estimated using a repeated stratified k-folds cross-validation scheme. With four splits and 25 repeats, we achieved 100 estimates of the performance for each model. This was done for both data sets.

The Cox proportional hazards model achieved the highest concordance index measured on both data sets.
When we visualised the measured Brier scores over the time period of 12 to 60 months in order to interpret the models' overall performance for the five first years. All models showed a rising trend in the measured Brier score. This indicates less accurate predictions over time. The models had similar Brier scores, with the exception of Aalen's additive fitter. This model had a slightly poorer result when time increased.

SAMMENDRAG

I denne oppgaven brukte vi tre typer overlevelsesanalysemodeller for å modellere overlevelsestiden til pasienter som lider av endetarmskreft og pasienter som lider av hode- og halskreft.
Disse modellene var Cox-regresjon, Aalens additive regresjonsmodell og akselererte levetidsmodeller. Målet med denne oppgaven var å sammenligne den målte ytelsen til disse modellene ved hjelp av å bruke concordance index og Brier score som ytelsesberegninger.

Vi estimerte disse ved å bruke en metode som heter "repeated stratified k-folds" for å kryssvalidere de målte resultatene. Vi delte datasettene opp i fire og gjentok dette 25 ganger, for å oppnå totalt 100 "folds". Dette gav oss muligheten til å kalkulere ytelsesberegningene 100 ganger per modell. Vi benyttet denne løsningen på begge datasettene.

Cox-regresjon oppnådde høyest concordance index på begge datasettene.
For å forstå modellenes nøyaktighet de første fem årene visualiserte vi Brier scoren over tidsperioden tolv til 60 måneder. Alle modellene viste en trend. Dette indikerte at modellene blir mindre nøyaktige over tid. De fleste modellene hadde svært liknende resultater målt med Brier score, men Aalens additive regresjonsmodell hadde noe svakere resultater.

PREFACE

I would like to thank everyone who has helped and supported me during the time of writing this thesis. A special thanks to my supervisors Oliver Tomic and Cecilia Marie Futsæther. They have done an excellent job guiding and assisting me with my work throughout this semester.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

List of all abbreviations in alphabetic order:

- **AAF** Aalen's Additive Fitter
- **AIC** Akaike information criteria
- **AJCC** American Joint Committee on Cancer
- **AFT** Accelerated failure time
- **BMI** Body mass index
- **C-index** Concordance index
- **CDF** Cumulative Density Function
- **CPH** Cox Proportional Hazard
- **CRT** Chemoradiotherapy
- **CT** Computed tomography
- **CV** Cross-validation
- **DBSCAN** Density-based spatial clustering of applications with noise
- **FDG** Fluorodeoxyglucose
- **HNC** Head and neck cancer
- **HR** Hazard ratio
- **IBS** Integrated Brier score
- **IQR** Interquartile range
- **KM** Kaplan Meier
- **LNAFT** Lognormal accelerated failure time
- **LOF** Local outlier factor
- **LR** Likelihood ratio

- **MAR** Missing at random
- **MCAR** Missing completely at random
- **MNAR** Missing not at random
- **MRI** Magnetic resonance imaging
- **NMBU** Norwegian University of Life Sciences
- **OS** Overall survival
- **PET** Positron emission tomography
- **Q-Q** Quantile-quantile
- **RT** Radiotherapy
- **RENT** Repeated elastic net technique
- **TNM** Tumour node metastasis
- **UICC** Union for International Cancer Control
- **WAFT** Weibull accelerated failure time

CHAPTER

ONE

INTRODUCTION

1.1 Background and Motivation

According to the Cancer Registry of Norway, 4 out of 10 Norwegians are going to develop a form of cancer by the time they reach the age of 80 years old [1].

Oncology is a branch of medicine that aims to study and treat cancers and tumours. Within this field, events of interest can be death, recurrence of tumours or any other disease-related event.

Other master students have previously considered these data sets for binary classification problems in terms of overall survival and disease-free survival as binary categories. They were interested in predicting whether the patients experienced an event of interest. In this thesis, we are interested in using survival analysis methods for modelling the time until the event of interest occurs.
Compared to other regression methods, survival analysis can use information from the observations that do not experience the event during the time of follow-up. We say that these observations are censored because we do not have the complete information. Three common reasons for censored data to occur are due to the end of the trial or patients either withdrawing before the end or being lost to follow-up [2]. With survival analysis, we can estimate survival functions and hazard functions and even make predictions about the estimated time until the event of interest. Some methods for survival analysis can also assess the effect of predictor variables [3].
In this thesis we are interested in comparing different methods for survival analysis on two data sets: A rectal cancer data set, and a head and neck cancer data set. Performance metrics and visual tools are utilised to compare the performance between the different models.

1.2 Project Description and Research Questions

This thesis takes into consideration some traditional methods for survival analysis methods and applies them to two data sets. These are OxyTarget, a rectal cancer data set, and HNC, a head and neck cancer data set.
We are interested in using Aalen's additive hazards (AAF), Cox proportional hazards (CPH) and accelerated failure (AFT) models to model the overall survival

time (OS). The concordance index and Brier score are used in order to measure their performance.
To acquire the performance metrics from each model, we utilise a repeated stratified k-fold cross-validation scheme.
In this thesis the following **research questions** are considered:

1. Which of the three survival models CPH, AAF and AFT achieves the highest performance measured using the concordance index?

2. How do the three survival models CPH, AAF and AFT perform in the time period of twelve to 60 months measured using the Brier score?

1.3 Related Work

As mentioned above, other master students have considered the binary classification problem on the OxyTarget and HNC data sets. Among these are Engesæth and Fjellvang.
Engesæth applied the feature selection method RENT [4] [5] on the OxyTarget data set in their master thesis [6]. Fjellvang applied radiomics and RENT on the HNC data set in their master thesis [7].
Madadizadeh et al. analysed survival for patients suffering from colorectal cancer (CRC) and compared CPH with additive hazards models, including AAF [8].
Orbe et al. compared AFT models with CPH on breast and gastric cancer [9].
These papers showed that both AFT and AAF can be an alternative to CPH when the proportional hazards assumption fails.

CHAPTER

TWO

THEORY

2.1 Oncology

Oncology is a branch of medicine which specialises in the diagnosis and treatment of cancer and tumours [10]. In this section, a brief introduction to rectal cancer and head and neck cancer is given.

2.1.1 Rectal Cancer

Cancer that forms in the rectum is called rectal cancer [10]. The rectum is located in the pelvic and connects the colon to the anus. We can divide it into the proximal rectum (upper) and the distal (lower) rectum.
In the time period 2017 to 2021, the rectal cancer survival rate was 71.9% and 73.5% for Norwegian men and women respectively. The median age at diagnosis was 70 years old for patients suffering from cancers in the rectum and rectosigmoid areas [1].

2.1.1.1 Diagnosis and Staging

When suspicion about rectal cancer is raised, a colonoscopy or a medical imaging study is required [11]. Magnetic resonance imaging (MRI), endorectal ultrasound and computed tomography (CT) scans are some tools for digital rectal examination that can be used for diagnosing and staging rectal cancer [12].

The tumour node metastasis (TNM) staging system was developed by the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC). This system for cancer staging assumes that the tumour cells spread from the primary site to either the adjacent organs or via blood vessels or lymphatic vessels. Variations of the TNM staging have also been proposed for each tumour site. In this brief introduction, we do not go into more detail. Based on the classifications in the categories T, N and M, the overall cancer stage can be determined. This also depends based on the cancer type and site. In general, stages I-II are categorised by no metastasis, while stage III involves metastatic lymph nodes. Stage IV involves distant metastasis. [13] [11].

2.1.1.2 Treatment

The primary goal for the treatment of rectal cancer is the oncological cure and overall survival (OS). Other priorities are sparing the anal sphincters with the functioning digestive system, and sparing of genitourinary and reproductive organs [12].

Several factors play in when determining the treatment for rectal cancer. Amongst these are staging and distance to the colon. Cancers close to the colon are often treated with surgical resection followed by adjuvant treatment. More distal cancers are often treated with RT before surgery. This also depends on staging [12].

2.1.2 Head and Neck Cancer

HNC is a group of cancers that occurs in the head and neck region. This includes the nasal cavity, sinuses, lips, mouth, salivary glands, throat and larynx (voice box) [10].

2.1.2.1 Diagnosis and Staging

When diagnosing a patient who is potentially suffering from HNC, the first steps are examining the patient's history, physical examination and radiologic imaging. Taking large biopsies can result in anatomical distortions and false positive test results, and it is therefore preferably used after completion of the initial steps [14].

For HNC, the staging process varies between different sites. In this thesis, we do not go into detail on each location. The patients with smaller tumours without prominent lymph involvement are classified as stages I-II. Stages III-IV involves locally advanced cancer tumours that are invading surrounding tissue or a higher amount of lymph nodes. Patients with evidence of distant metastases are classified as IV [14].

2.1.2.2 Treatment

Treatment of HNC depends on cancer staging, site and surgical accessibility. The treatment can encompass surgery, RT and medical oncology and can involve a multispeciality team for evaluation [14]. In this brief description, we do not go into further detail.

2.2 Survival Analysis

In survival analysis, we are interested in modelling the time until an event occurs. This can be any event of interest, for instance, the time of tumour recurrence or death.
We refer to the duration from some initial event, for instance, treatment or diagnosis, until the event occurred as the survival time.
In general terms we refer to this event as a failure. Survival analysis is a collection of statistical procedures and methods that considers this survival time as the variable of interest [2].

The use of life tables, a statistical tool for modelling survival probabilities, can be traced back as early as 1662 when it was used by the merchant John Graunt in the book "Natural and Political Observations upon the Bills of Mortality" [15]. In the 19th century, mortality tables acted as a tool for actuaries in their work for estimating the price for life annuities [16]. The mortality tables estimate the probability of death for a person at a certain age given that they had survived until reaching their current age [17].
The Kaplan-Meier estimator was introduced in the mid-20th century as a method for estimating the survival curves for a population that includes censored data points [18]. The proportional hazard model was introduced in 1972 by D. R. Cox[19]. These tools are still frequently used in survival analysis.

2.2.1 Censored Data

Survival analysis takes censored data into consideration. When we do not know the exact time of the failure, we have a censored observation. Figure 2.2.1 shows 5 observations in a hypothetical clinical study. We observe these observations for a time period, but only observations B and D experience the event of interest. These observations are not censored.
Observations A, C and E are right-censored because we do not know the time of experiencing the event. What we do know is that A survived from the time of inclusion until being lost to follow-up, C until dropping out and E until the final follow-up.

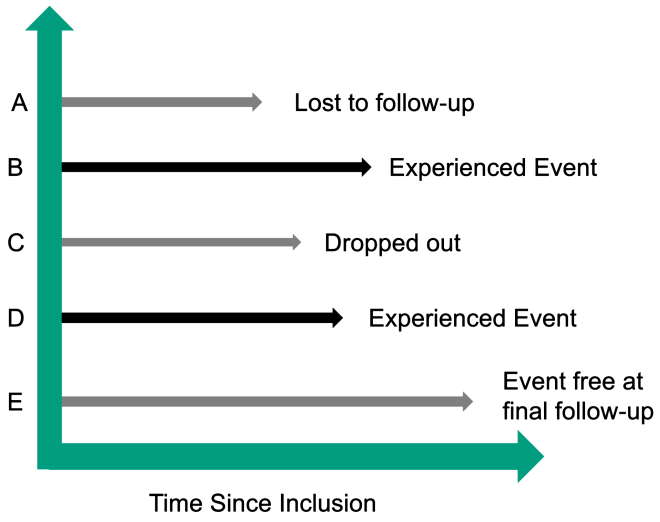

Figure 2.2.1: Demonstration of censored data. Inspired by [20].

If the event occurs between two registered time-points, the observation is interval-censored because we only know that the true survival time is in between the two check-ups [2].
When we only know that the true survival time is less than the registered time, it is left censored [2].
Censored data is also referred to as incomplete because it is missing information.

2.2.2 Truncated Data

In reality, there is a subset of subjects that fails before the study even starts. This is called left truncation, and it is a real problem because it artificially skews the survival distribution [3]. For instance, if we are testing a treatment for a new disease with a high mortality rate, some of the observations may have succumbed before enrollment.

2.2.3 Hazard and Survival

The survival function can be described as the probability that the true survival time T is higher than some specified time t. This is described in equation 2.1 [2].

$$S(t) = P(T > t) \tag{2.1}$$

The hazard is the current risk of event/failure at the time point t. It is denoted as $h(t)$ and described in equation 2.2 [2]. The hazard function is frequently denoted as $\lambda(t)$.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{2.2}$$

Figure 2.2.2 shows how the survival function, hazard function and cumulative hazard function, $H(t)$, are related to each other and the probability distribution functions (PDF), and the cumulative density function (CDF) [2] [21]. In survival analysis, we say that $f(t)$ is the probability of death at time t, and $F(t)$ is the probability of having experienced the event by time t.

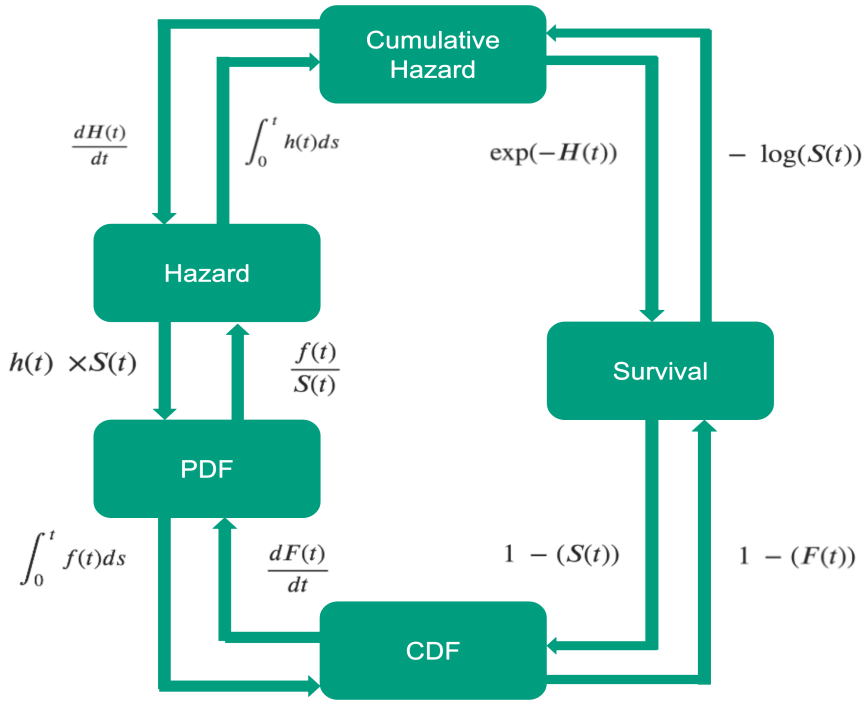

Figure 2.2.2: Relationships between $S(t), h(t), H(t), f(t)$ and $F(t)$. Inspired by [21].

2.2.4 Kaplan Meier - The Product Limit Estimator

At the beginning of this chapter, we mentioned the importance of the Kaplan Meier estimator. In this thesis, it is referred to as KM.
In the real world, the true survival distribution is rarely known. The KM estimator is a non-parametric estimator of the survival function $S(t)$ [3].

In equation 2.3, the KM estimator it is described as the product of 1 minus the fraction of failures, d_i and subjects at risk, n_i for all time points i until t [3].

$$\hat{S}_{KM}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{2.3}$$

The KM estimated survival function is undefined for values of t higher than the highest follow-up time duration, even if we have surviving observations at that time [3]. From equation 2.3, n_i is 0 when there are no remaining subjects at risk due to censoring and no more follow-up.

It is often useful to visualise the KM-estimated survival curves. It can give a general impression of a population's estimated survival time, and if we separate by a categorical variable, we can also visualise the difference in survival time between multiple groups. An example of this is provided in figure 2.2.3. The figure shows the KM estimated survival functions for the OxyTarget data set. The left axis shows estimates separated by the binary feature "Suspected metastatic lesions at diagnosis". The right axis shows the estimated groups separated by age at the median of 65.0. From visual inspection, it appears to be a significant difference in the estimated survival curves to the left and little to no difference to the right.
The more observations included in the KM estimated survival function, the smoother it will be. This is because it includes more data points. As the number of observations increases, it will also approach the population's theoretical survival function [2].
With fewer observations, the KM estimated curves will be more grainy. For each time step, the KM estimated survival probability is updated by multiplying the previous with 1− the proportion of remaining observations experiencing the event. Because of this, when there are a low amount of observations left, just a couple of events will make a great difference in the estimated survival probability between the time steps .

2.2.5 The Logrank Test

In 1966, Mantel proposed a chi-square procedure for comparing life tables in their entirety [22]. A similar procedure was performed by D. R. Cox on a different problem with multiple events in 1959 [23]. In 1972, Cox presented the logrank test. It is also frequently referred to as the Mantel-Cox test [19].
The hypothesis tested by the logrank test is:

h_0 : no significant difference in hazard distributions

h_1 : significant difference in hazard distributions

The logrank test is useful for giving a statistical comparison between two or more groups in terms of their hazard distributions.

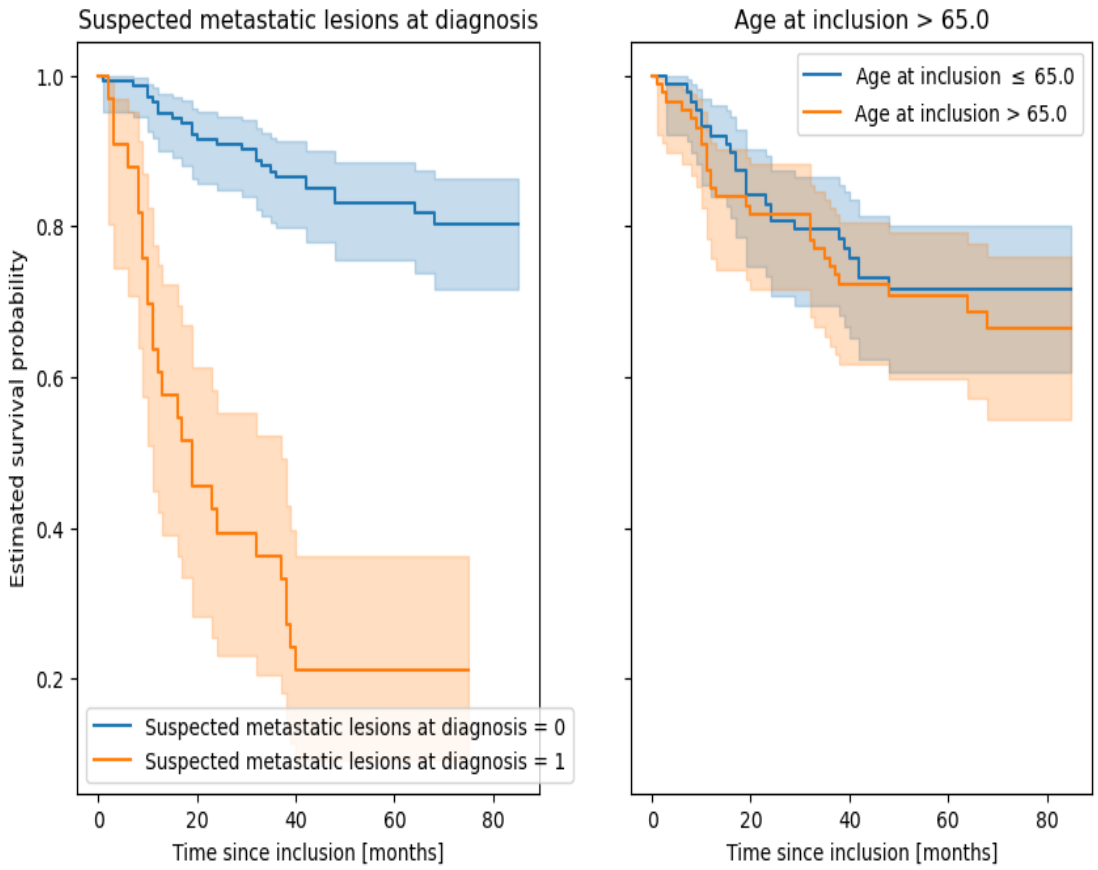

Figure 2.2.3: Example of KM estimated survival curves with 95% confidence intervals for groups in the OxyTarget data set.

Table 2.2.1 shows the logrank test statistics and p-values for the survival populations shown in figure 2.2.3. The logrank test rejected the h_0 when comparing male and female patients because there is some significant difference in the hazard distributions.

The logrank test failed to reject the h_0 for the groups separated by the median age.

2.3 Cox Proportional Hazards

The Cox Proportional Hazards was introduced by D. R. Cox in the paper "Regression Models and Life-Tables" in 1972 [19]. In this thesis, this model is referred to as CPH. It is considered to be one of the most used models within survival analysis [3].

When comparing CPH to a fully parametric model, it can be just as efficient. When the assumptions for the parametric model are not true, CPH will likely be more efficient than the parametric model [3].

Table 2.2.1: Logrank for groups separated by the binary category suspected metastatic lesions at diagnosis and by median age at inclusion.

Feature	P-value	h_0 given 0.05 significance
Suspected metastatic lesions at diagnosis	<0.001	Rejected
Age at inclusion > 65.0 (median)	0.295	Not rejected

2.3.1 Definition

CPH is a semi-parametric model. This means that the model does not make any assumptions on the specific distribution of the response variable, but it makes an assumption about the effect the covariates have on the hazard function [3].
Equation 2.4 shows how the CPH is stated in terms of the hazard function [3]. The baseline hazard, h_0, "varies" with time, and the relative hazard is represented by $e^{X\beta}$, which is constant and multiplied by the baseline hazard. With this formulation, the linear combination of the predictors is denoted $X\beta$. The hazard is the product of the baseline hazard and the relative hazard [3].

$$h(t \mid X) = h_0(t)e^{X\beta} \tag{2.4}$$

The hazard ratio between two groups is the fraction of the relative hazards for each group. In the subsection 2.3.2 we describe the proportional hazard assumption, which considers the fact that the relative hazard has the same proportion between groups for all values of t.

2.3.2 The Proportional Hazard Assumption

If the hazard ratio is the same for all values of t, one can say that it is proportional [24].
Some researchers are questioning the need for checking the proportional hazard assumption, but in general it is considered to be good conduct to include it [24].
We can check for the proportional hazards assumption by visual inspection of the Schoenfeld residuals, or conducting a statistical test based on the correlation between these residuals and time [3]. The Schoenfeld residuals are calculated per feature, and one residual is achieved per observation. If these residuals are related to time, it is an indication that the proportional hazards assumption does not hold for this feature. We get the following hypothesis test [2]:

h_0 : No correlation between Schoenfeld residuals and ranked failure time

h_1 : Correlation between Schoenfeld residuals and ranked failure time

If the h_0 is rejected, the conclusion is that the proportional hazards assumption is violated [2].

2.3.3 Limitations

When the criteria for the proportional hazard assumptions fail to be fulfilled, and features other than treatment indicator is included, the model can yield incorrect

standard deviation for the estimations. Bootstrapping methods can allow for valid confidence intervals to be created in this scenario [24]. We can also stratify the categorical columns that do not follow the proportional hazard assumption [3].

2.3.4 Coefficient Estimation

In order to estimate β we first let $t_1 < ... < t_k$ be the order of unique survival times. For now, assume no ties in observed survival times. We denote R_i as the set of individuals at risk at timepoint t_i. The observations in R_i are denoted with j. Observation j is at risk at time t_i if $Y_i \geq t_i$, where Y_i is the survival time or time until censoring for observation j. If there is a failure at time point t_i, the conditional probability for observation i being the one experiencing the failure, given at risk is:[3]:

$$P(i \text{ fails at time } t_i \mid R_i \text{ and 1 failure at } t_i) = \frac{P(i \text{ fails at time } t_i \mid R_i)}{P(1 \text{ failure at } t_1 \mid R_1)} \tag{2.5}$$

The probability that i fails at time point t_i given R_i is the hazard for that time point for that observation. This is described in equation 2.4. The probability for a failure at time t_i, is then the sum of all the hazards in the risk set R_i. Because the baseline hazard at time t_i is a constant, we can remove it from both sides of the fraction [3].

$$\frac{h_0(t)e^{X\beta}}{\Sigma_{j\in R_i} h_0(t)e^{X\beta}} = \frac{e^{X\beta}}{\Sigma_{j\in R_i} e^{X\beta}} = \frac{e^{X\beta}}{\Sigma_{Y_j \geq t_i} e^{X\beta}} \tag{2.6}$$

The baseline hazard is now out of the picture. The equation 2.7 shows the partial likelihood for β, and equation 2.8 shows the log partial likelihood. In order to obtain the estimates of β, these partial likelihoods are treated as normal likelihoods [3].

$$L(\beta) = \prod_{Y_i \text{uncensored}} \frac{e^{X_i\beta}}{\Sigma_{Y_j \geq Y_i} e^{X\beta}} \tag{2.7}$$

$$\log(L(\beta)) = \sum_{Y_i \text{uncensored}} (X_i\beta - \log(\sum_{Y_j \geq Y_i} e^{X_j\beta})) \tag{2.8}$$

The lifelines python library [21] use Breslow's approximation [25], described in equation 2.9, for computational approximation of the log-likelihood, and handles ties with Efron's approximation [26], described in equation 2.10 [3].

$$\log(L(\beta)) = \sum_{i=1}^{k} \big\{ S_i\beta - \log\big[\sum_{Y_j \geq t_i} e^{X_j\beta} \big] \big\} \tag{2.9}$$

For Effron's approximation, we let S_i be $\Sigma_{j\in D_i} X_j$, and D_i be the set of indices j at risk at t_i and d_i is the failures at t_i.

$$\log(L(\beta)) = \sum_{i=1}^{k} \left\{ S_i\beta - \log\left[\sum_{Y_j \geq t_i} e^{X_j\beta} - \frac{j-1}{d_i} \sum_{l\in D_i} e^{X_l\beta} \right] \right\} \tag{2.10}$$

2.3.5 Interpretation

The **Wald Z-score** for a coefficient is found by dividing the coefficient itself by its standard error. The p-value is the probability that $p >\mid Z \mid$ [2]. A coefficient with a high p-value suggests that the feature does not discriminate towards the response variable [2].
The likelihood ratio (LR) is found by multiplying the Log likelihood by -2. When looking for confounders, one can compare the likelihood ratios by subtracting the LR from the model trained on the reduced subset by the LR from the LR of the model trained on the full data set. With the h_0 of no interaction effect, we can use the fact that this statistic has a χ^2 distribution with the number of features being assessed as the degrees of freedom. With this distribution, we can conclude whether the features assessed are in fact confounders or not [2].

2.4 Aalen's Additive Fitter

Aalen's Additive Fitter was introduced by Odd O. Aalen in 1989 in the paper "A Linear Regression Model for the Analysis of Life Times" [27]. It is, as the title of the article suggests, a linear model for survival analysis. This model is referred to as AAF in this thesis. In comparison to the Cox PH model, AAF does not rely on the assumption of proportional hazards. However, it is more limited in the amount of features it can consider [28].

2.4.1 Definition

Equation 2.11 describes the AAF hazard function. The coefficients b_1 to b_n each represent the weight for the covariates, and b_0 is the bias / intercept value. The resulting hazard is the dot product between the b vector and $1, x_1, ..., x_n$.

$$h(t \mid x) = b_0(t) + b_1(t)x_1 + ... + b_n(t)x_n \tag{2.11}$$

AAF was presented by Aalen as a non-parametric model, even though one could argue that its linearity could represent an assumption about the distribution [27]. The time-varying nature of the coefficients makes AAF an interesting tool for interpreting the temporal changes in the correlation between the covariates and the target values. In comparison with CPH, which assumes that the effect over time is multiplicative, AAF can yield interesting information about the changes over time. Figure 2.4.1 shows an example of how the AAF models regression coefficients can vary over time. In this plot, we can see the regression coefficients for two features, and the intercept. From this plot, one can see that the feature "Suspected metastatic lesions at diagnosis" have a significant effect on the response variable, which is overall survival.

2.4.2 Coefficient Estimation

Let r be the number of features, n be the number of observations, R_t be the risk set at time t like in section 2.3.4.
We define $Y(t)$ as a $n \times (\mathrm{r} + 1)$ matrix. In this matrix, row i is a vector of the value 1 followed by the covariates for observation i: $Z^i(t) = (1, Z^i_1(t), ..., Z^i_r(t))'$.

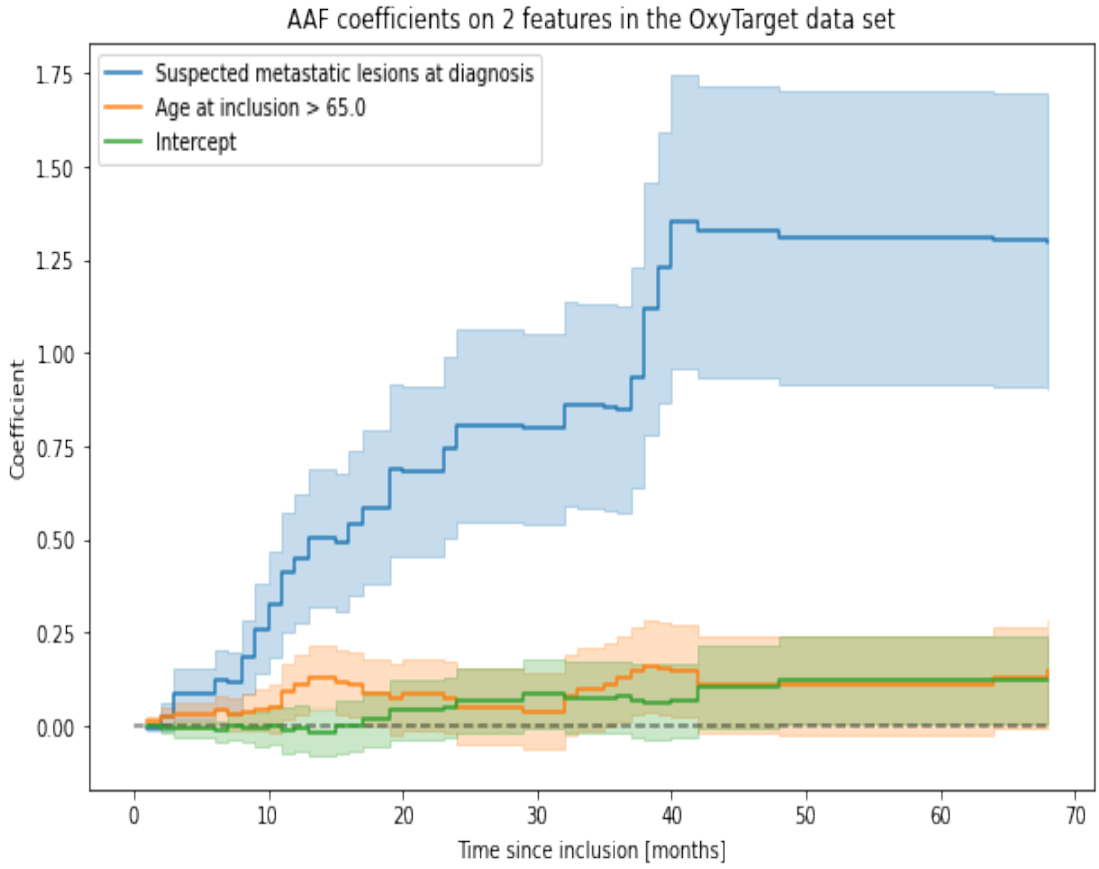

Figure 2.4.1: Example of AAF time-varying coefficients.

However, if observation i is not in R_i, i. e is no longer at risk, the vector for the observation consists only of zeros [27].
With this matrix, we can now consider the matrix formulation of the hazard function: $h(t) = Y(t)\beta(t)$. The matrix denoted as β matrix describes the regression coefficients at time point t. We are now interested in estimating the cumulative regression function described in equation 2.12 [27].

$$\beta_j(t) = \int_0^t b_j(s)ds \tag{2.12}$$

In order estimate 2.12, we let $T_1 < T_2 < ...$ be the survival durations in ascending order, and I_k be a vector of zeros for all values except for the subject experiencing the event at time T_k, which has a value of 1. We also define $X(t)$ as the generalized inverse of $Y(t)$: $X(t) = [Y(t)'Y(t)]^{-1}Y(t)'$.
Then we can estimate β:

$$\hat{\beta}(t) = \sum_{T_k \leq t} X(T_k)I_k \tag{2.13}$$

The β estimation is well defined when $Y(t)$ is of full rank and $Y(t)'Y(t)$ is invertible. Because of this, we can only estimate the coefficients when Y is non-singular. Aalen suggests that a solution for the estimation $\hat{B}(t)$ can be found using prior knowledge about the regression coefficients or by using least squares approximation

[27].

2.5 Accelerated Failure Time Models

In this thesis, we consider two accelerated failure time (AFT) models. One based on the Weibull distribution, and one based on the log-normal distribution. We refer to these models as WAFT and LNAFT in this thesis.

These accelerated failure time models assume the distribution of the survival times and is therefore considered to be parametric models [2].
It is worth mentioning that there have also been proposed AFT models which are not based on an assumption about the underlying distribution of survival times [29].
The covariates for AFT models have a direct effect on the estimated survival times. In comparison, the covariate effect for proportional hazards models is on the hazard [2].

2.5.1 Definition

We define the accelerated failure time model in equation 2.14, where ψ denotes the underlying survival distribution function, and σ is the scale parameter [3].

$$S_{t|X} = \psi(\frac{log(t) - X\beta}{\sigma}) \tag{2.14}$$

In equation 2.15 we have defined the AFT model in terms of the log of T (the survival time). In this formulation, ϵ is a random variable from the distribution ψ [3].

$$\log(T) = X\beta + \sigma\epsilon \tag{2.15}$$

2.5.2 Model Estimation

We can use maximum likelihood estimation to find the model parameters for the AFT models. For the observed survival times $t_1, ..., t_n$, we define the maximum likelihood estimation in equation 2.16[30].

$$L(\beta, \sigma) = \prod_{i=1}^{n} (f_i(t_i))^{\delta_i} (S_i(t_i)^{1-\delta_i} \tag{2.16}$$

Where:

$$f_i(t_i) = f_{\epsilon i}(z_i)$$
$$S_i(t_i) = S_{\epsilon i}(z_i)$$
$$z_i(t_i) = \frac{\log(t) - \mu - \beta_1 x_1 - ... - \beta_r x_r}{\sigma}$$

And δ_i is 1 if observation i has experienced the event of interest, and 0 if the observation is censored. $f_{\epsilon i}$ is the density function of the assumed distribution and $S_{\epsilon i}$ is the survival function for the assumed distribution.
Using the Newton-Raphson procedure, we can then find the unknown parameters μ, σ and β [30].

2.5.3 The Accelerated Failure Time Assumption

In section 2.3.2, it was described how the Cox Proportional Hazard assumes that the features have a multiplicative effect in terms of the hazard. On the other hand, the AFT underlying assumption is that the features have a multiplicative effect in terms of survival [2].

$$S_{\text{group}_2}(t) = S_{\text{group}_1}(\gamma t), \text{ for } t \geq 0 \tag{2.17}$$

$$T_{\text{group}_2} = \gamma T_{\text{group}_1} \tag{2.18}$$

A common example of the AFT assumption is to compare the survival times between dogs and humans. Consider equation 2.17, the groups can represent humans in group 2 and dogs in group 1, with the acceleration factor γ describing that humans live γ times longer [2]. Equation 2.18 demonstrates this in terms of random variable T for survival time.

2.5.4 Weibull Accelerated Failure Time

Weibull is the most common distribution to base a parametric survival model on. If T follows the Weibull distribution, then the log of T follows the extreme minimum value distribution [2]. Equation 2.19 describes the survival function for WAFT [3]. This survival function can be derived from the CPH models survival function, and it is therefore also the case that if the proportional hazards assumption holds, the accelerated failure time assumption also holds. The opposite is also true [3] [2].

$$S(t \mid x, y) = \exp\Big[-\exp\Big(\frac{\log(t) - X\beta}{\sigma}\Big)\Big] \tag{2.19}$$

$$f_{\epsilon i}(z_i) = -\frac{1}{\sigma}\exp[z_i - \exp(z_i)] \tag{2.20}$$

For the WAFT model, we can graphically inspect whether the AFT assumption holds by plotting $\log[-\log\hat{S}(t)]$. If the resulting line is linear, the AFT assumption holds [31][2].

2.5.5 Log-Normal Accelerated Failure Time

The LNAFT model assumes that the log of T follows a normal distribution [2]. We define the survival function in equation 2.21 [3].

$$S(t \mid x) = 1 - \Phi(\frac{\log(t) - X\beta}{\sigma}) \tag{2.21}$$

We define the density function [30]:

$$f_\epsilon(z_i) = \frac{1}{2\pi}e^{-\frac{1}{2}z_i^2} \tag{2.22}$$

For the LNAFT we can check if the AFT assumption holds by plotting $\Phi^{-1}[1 - \hat{S}(t)]$. If the resulting line is linear, the AFT assumption holds [31].

2.6 Evaluation of Models

2.6.1 Assessing Appropriateness of Parametric Distributions

In this thesis, we use two methods for determining the appropriateness of an assumption about the underlying distribution of survival times.
These two methods are the Akaike information criteria (AIC) [32] and quantile-quantile (Q-Q) [33] plots.
The AIC is $-2\log \text{likelihood} + 2k$, where k is the degrees of freedom. It is an estimate of the relative amount of information lost by the model. A lower estimated loss of information will give a lower AIC score. A lower AIC can indicate a higher quality model [34].
In a Q-Q plot, the quantiles of one data set is plotted against the quantiles for another. The lifelines library has an implementation of the Q-Q plots which also considers censored observations [21].

2.6.2 Cross Validation

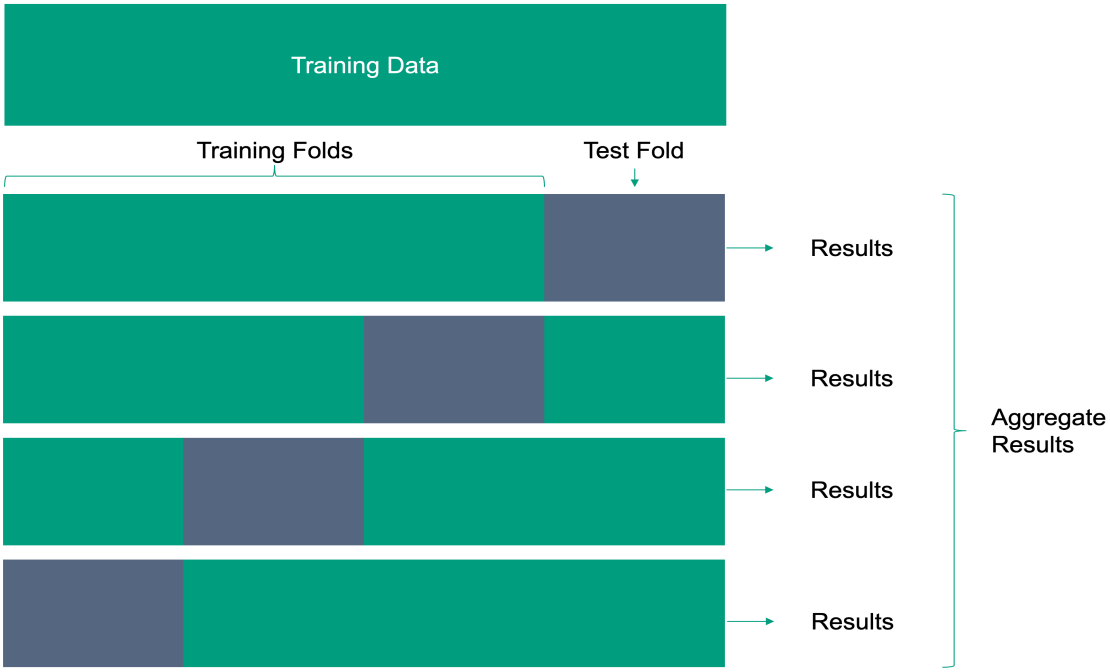

Figure 2.6.1: K-folds cross validation model reproduced as presented by Rashcka [35].

When creating prediction models it is important to estimate the models' performance on unseen data. Testing the model on unseen data allows us to see whether the model is able to generalize or not. In situations with smaller amounts of data, it can be difficult to achieve a traditional train test split and still have enough data to both train models and get sufficient test data size.
K-fold cross-validation can help in this scenario by allowing for multiple test and train splits from the same data set. The data set is separated into k-folds as shown in figure 2.6.1, and in an iterative process, one by one the folds are

held out as a test data set. This will give multiple results and can give a better understanding of the performance of unseen data. Another popular use case for this method is hyper-parameter tuning and model selection [35].

The algorithms separating the data into folds can stratify by one of the categorical features or leave the distributions random [36]. If the response variable is sparse, it can be useful to stratify. This will make sure that we have the same occurrence rate for the category in each fold.

Repeated K-Folds is a method where the K-fold process is repeated n-times [36]. We then get k splits times n repeats folds. Table 2.6.1 shows an example with 2 splits and 2 folds, for a total of 4 folds. Because k is 2, we simply split the data set in half, and train on one half then test on the other. In the next fold, the training and test samples are swapped. In fold 3, we do a new split and swap it again for fold 4.

Table 2.6.1: Example of a repeated k folds split with $n = 2$ and $k = 2$.

	train_samples	test_samples
fold_1	[patient_1, patient_4, patient_5, patient_6]	[patient_2, patient_3, patient_7, patient_8]
fold_2	[patient_2, patient_3, patient_7, patient_8]	[patient_1, patient_4, patient_5, patient_6]
fold_3	[patient_1, patient_4, patient_6, patient_8]	[patient_2, patient_3, patient_5, patient_7]
fold_4	[patient_2, patient_3, patient_5, patient_7]	[patient_1, patient_4, patient_6, patient_8]

2.6.3 Harrell's Concordance Index

Harrell's Concordance Index was introduced as a method for evaluating information provided from medical tests in 1982 [37]. In this thesis, we refer to this metric as the C-index. It is considered to be one of the most used methods for evaluating both traditional survival models as well as modern machine learning models [38].

2.6.3.1 Definitions

The C-index is in general calculated by dividing the number of concordant pairs of observations by the total amount of comparable pairs [38].
A pair of observations is comparable if we are able to determine which of them failed first [39].
This can be determined in two cases [37]:

1. Both observations have experienced the failure
2. One failure and the other observation surpass the survival time of the failed one.

This means that pairs of two surviving observations can not be compared, as well as pairs where the shortest survival time is censored.
Equation 2.23 shows how the C-index is calculated as a fraction between the concordant and the comparable pairs [38].

$$C_H = \frac{\Sigma_{i \neq j} I(\tilde{T}_i < \tilde{T}_j) \cdot I(\eta_i > \eta_j) \cdot \Delta_i}{\Sigma_{i,j} I(\tilde{T}_i < \tilde{T}_j) \cdot \Delta_i} \tag{2.23}$$

Where:

$$I(\tilde{T}_i < \tilde{T}_j) = \begin{cases} 1, & \text{if } \tilde{T}_i < \tilde{T}_j \\ 0, & \text{otherwise} \end{cases}$$

$$I(\eta_i > \eta_j) = \begin{cases} 1, & \text{if } \eta_i > \eta_j \\ 0, & \text{otherwise} \end{cases}$$

$$\Delta_i = \begin{cases} 0, & \text{the pair is comparable} \\ 1, & \text{otherwise} \end{cases}$$

In equation 2.23, i and j are the indices which refers to the pairs of observations. $\tilde{T}_i$ and $\tilde{T}_j$ are the observed survival times, η_i and η_j are the predicted survival times. Δ_i allows the equation to discard inadmissible observations.

The numerator of the fraction is equal to the number of comparable pairs where the correct order is predicted. The denominator of the fraction is equal to the total number of admissible pairs.

Equation 2.24 shows an alternative calculation as demonstrated in the documentation for the Lifelines library [21]. This equation also takes into consideration ties, and gives them a weight coefficient of 0.5. The formulation is also simplified.

$$C_H = \frac{\text{Comparable and correct pairs} + 0.5 \cdot \text{ Comparable and tied pairs}}{\text{Comparable pairs}} \tag{2.24}$$

2.6.3.2 Interpretation

When interpreting the C-index, 1 is the highest achievable score, and it is achieved when both parts of the equations are equal. This is the case when all the comparable pairs are estimated in the correct order. If all predictions are completely random, the probability for each pair to be correct is 50%. When half of the comparable pairs are wrongly estimated, and the other half is correct, the C-index would be 0.5. This score is therefore considered to be random guessing, and a low C-Index [2].

When interpreting the intermediate values of the C-index ($0.5 <$ C-index < 1), there are no answers to whether any of the values are a good or bad representation on the models real-world performance. The conclusion one can draw from the C-index is the amount of concordant pairs relative to the amount of total comparable pairs [39].

The C-index is equivalent to the area under the receiver operator curve [40].

2.6.3.3 Limitations

The C-index can depend on the censoring distribution. If there is a high percentage of censored observations, the C-index tend to have a positive bias. There are multiple alternative approximations to the C-index that takes this into consideration, including Uno's C-index [40].

If comparing the results of the predictions to a specific survival duration, the C-index does not consider the prediction of survival time for a fixed time. It rather focuses on the order of the predictions. The C-index is not an appropriate metric for this problem [40].

2.6.4 Uno's Concordance Index

In section 2.6.3.3 we mentioned that the C-index can be biased if there is a high amount of censored observations, and that Uno's C-index can be an alternative with less bias.

This modification of the C-index takes into consideration the distribution of censored data points in both the training- and test data sets because it is based on the inverse probability of censoring weights.
Equation 2.6.4 shows how the Uno's C-index is estimated, where $\hat{G}(\tilde{T}_i)$ is the KM estimator with censoring as the event of interest, and τ is the truncation time. The truncation time should be chosen so that the probability of being censored after t should be none 0. Uno states that the support for the distribution of censored data usually is shorter than the one of the survival time and that this can lead to an unstable estimation of the tail part of the survival function [40].

$$\hat{C}_U = \frac{\sum_{i=1}^{n}\sum_{i=1}^{n}\Delta_i\{\hat{G}(\tilde{T}_i)\}^{-2}I(\tilde{T}_j < \tilde{T}_i, \tilde{T}_i < \tau)I(\eta_i > \eta_j)}{\sum_{i=1}^{n}\sum_{i=1}^{n}\Delta_i\{\hat{G}(\tilde{T}_i)\}^{-2}I(\tilde{T}_i < \tilde{T}_j, \tilde{T}_i < \tau)} \tag{2.25}$$

2.6.5 Brier Score

The concept behind the Brier score was introduced by the American statistician and meteorologist Glenn W. Brier [41]. In survival analysis and medical statistics, the Brier score is used as an overall performance measure [42].

2.6.5.1 Definition

In contrast to the C-index, the Brier score is a time-dependent performance measure. This means that we can use the Brier score to measure the overall performance of a model at a specific point in time.
The Brier score is the mean squared distance between the predicted survival probability and observed survival status [43].
Equation 2.26 shows how the Brier score is calculated. $\hat{G}(t)$ is a KM estimation of the censoring distribution, $\hat{\pi}(t \mid \mathbf{x}_i)$ is the estimated survival probability at time t given the covariates x_i, y_i is the time until failure or censoring. δ_i is 1 if the time until the event is lower than the hypothetical time under observation, otherwise 0.
We have three categories for the observations:

1. $y_i \leq t$ and $\delta_i = 1$
2. $y_i > t$ independent of δ_i

3. $y_i \leq t$ and $\delta_i = 0$

The category decides the error calculation and weighting for the observation. The weighting compensates for the information lost due to censoring.
$I(y_i \leq t \wedge \delta_i = 1)$ is 1 if the observation belongs to category 1, otherwise 0. An observation belongs to category 1 if the survival time y_i is less than or equal to t, and the time until the event is lower than the hypothetical censoring time. For this category, the event status is 0. We calculate the squared error $0 - \hat{\pi}(\mathbf{x}_i)^2$ and scale by multiplying with $1/\hat{G}(y_i)$ to find the contribution to the Brier score.
$I(y_i > t)$ is 1 if the observation is in category 2, otherwise 0. Observations in category 2 have experienced the event, and we calculate the contribution: $0-\hat{\pi}(\mathbf{x}_i)^2$ and scale by multiplying with $1/\hat{G}(t)$.
Category 3 gets a weight of 0 and does not contribute to the Brier score. This category consists of observations with unknown event status. The Brier score calculation is shown in equation 2.26 [43].

$$\mathrm{BS}(t) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq t \wedge \delta_i = 1)\frac{(0 - \hat{\pi}(t \mid \mathbf{x}_i))^2}{\hat{G}(y_i)} + I(y_i > t)\frac{(1 - \hat{\pi}(t \mid \mathbf{x}_i))^2}{\hat{G}(t)} \quad (2.26)$$

2.6.5.2 Interpretation

When interpreting the measured Brier score, 0 is a perfect score. An uninformative model would achieve a Brier score of 0.25. The Brier score can be interpreted as the mean squared error between the predictions and the observed event status adjusted for the probability of censoring [43].

2.6.5.3 Integrated Brier Score

In this thesis, we refer to the integrated Brier score as IBS. The IBS is the integral of the Brier score between two points in time. This can for instance be $t = 1$ and $t = 5$ years. This allows for a numeric performance metric for evaluating the overall performance of the predictions over a time period. As with the Brier score, 0 is a also perfect IBS score, while the worst case depends on the distributions of censoring.
IBS is described in euqation 2.27, where we have the weighing function: $w(t) = t/t_{stop}$.

$$\mathrm{IBS} = \int_{t_{start}}^{t_{stop}} \mathrm{BS}(t)dw(t) \quad (2.27)$$

2.7 Data Pre-Processing

In order to be able to conduct data science experiments on a data set, it might be necessary to do some pre-processing of the data. There are many reasons for this, it can improve the results, improve the reliability and robustness of the models and the results, and it can make the models even work at all. This section will describe the theory behind the data pre-processing methodology used in this thesis.

2.7.1 Encoding Variables

Many data science libraries are requiring integer-encoded class labels. It is also considered good practice, even if the libraries used can handle other data types [35]. This subsection describes some methods for encoding these categories.

When encoding a categorical variable, it is important to consider whether the variable is **ordinal** or **nominal**. When the order of the categories matters, it is considered an ordinal feature. If the order does not matter, it is a nominal feature [35].

For ordinal encoding, one needs to consider the order of the categories and create a feature mapping such that the non-numerical input values get the correct assigned value [35].

For nominal features, a solution can be one-hot encoding. With this solution, each category in the original feature gets turned into a new binary feature [35]. For instance, if the feature in question is "weather", and the categories are "sunny, "cloudy" and "snowing", there would be one feature for "weather_sunny", "weather_cloudy" and "weather_snowing". If all values are one of the three categories, if it is not "sunny" or "cloudy", it must be "snowing".

When separating these three categories, two of the features can give all the necessary information: if not "sunny" or "cloudy", it has to be "snowing".

Sometimes, there is no value registered. Section 2.7.2.2, it is described how one can create a new feature column for this information as well.

One can also combine multiple categories that are similar, or even encode intervals of continuous variables into categories.

2.7.2 Handling missing values

Most computational tools are not able to interpret missing data, and this can lead to less reliable results. In some cases, the missing values can also result in the tools not even working at all. Removal of missing rows and columns can be one solution. Another solution is to use imputation to estimate a likely value [35].

2.7.2.1 Types of Missing Data

Rubin introduced a framework that considers the mechanism which causes the information to be missing [44]. If there is some reason for the data to be missing, which is related to the analysis and not described by the other variables, it is considered to be missing not at random (MNAR).
If the probability for the data to be missing is related to the other variables, it is considered to be missing at random (MAR).
When the probability of the data to be random is completely random, it is classified as missing completely at random (MCAR) [45].

The mechanisms behind the reason for the missing data can include some information about the data which appears to be missing [45].

2.7.2.2 Removing Missing Data

When removing missing data from the data set, it is important to note that there can be some side effects. If too much information is removed, it can reduce the

models' performance [35]. When removing data, it can be useful to set a threshold for the allowed amount of missing data points per row and column.

2.7.2.3 Interpolating Missing Data

As mentioned in section 2.7.2.2, removal of data can lead to some disadvantages. There are several ways of estimating what the missing value could be. Some of the more "simple" methods are mean, and most frequent imputation [35].

There are also other methods for interpolating missing values, such as machine learning methods, which can consider the context from the other columns for the row containing the missing values. The K-nearest neighbours imputation method is one of these methods. It is an unsupervised clustering algorithm that assigns the value in the likes of the "K" nearest neighbours [46].

2.7.3 Outlier Detection

In statistical analysis one is interested in having as much information as possible. However, some observations can bring information which is so extreme that it can negatively affect the overall performance and robustness.

2.7.3.1 Z-score outliers

One method for finding outliers is by looking at the Z score for each single data input. The Z-score is calculated by subtracting the mean and dividing by the standard deviation of the population, as described in equation 2.28. This is calculated for each feature. The value of the Z-score indicates the amount of standard deviations from the population average.

$$z(x) = \frac{x^{(i)} - \mu_x}{\sigma_x} \tag{2.28}$$

2.7.3.2 Local Outlier Factor

Another method for discovering outliers is by using the local outlier factor, or LOF [47], method. This will allow for assigning to which degree an observation is an outlier. An example is presented in figure 2.7.1. In this example, we can see the data points, and the red rings around indicate the scaled local outlier factor score. We can see that the remote observations have rings with a larger radius around them.

2.7.4 Data Scaling

2.7.4.1 Min Max Scaling

Normalization, or min-max scaling, is one of the two popular approaches for bringing features to the same scale. Normalization is frequently referred to as bringing the values to as scaling all values to between 0 and 1. To normalize the value for x, one can subtract by the minimum value and divide by the difference between the highest and lowest values, as described in equation 2.29. This process is a special case of min-max scaling [35].

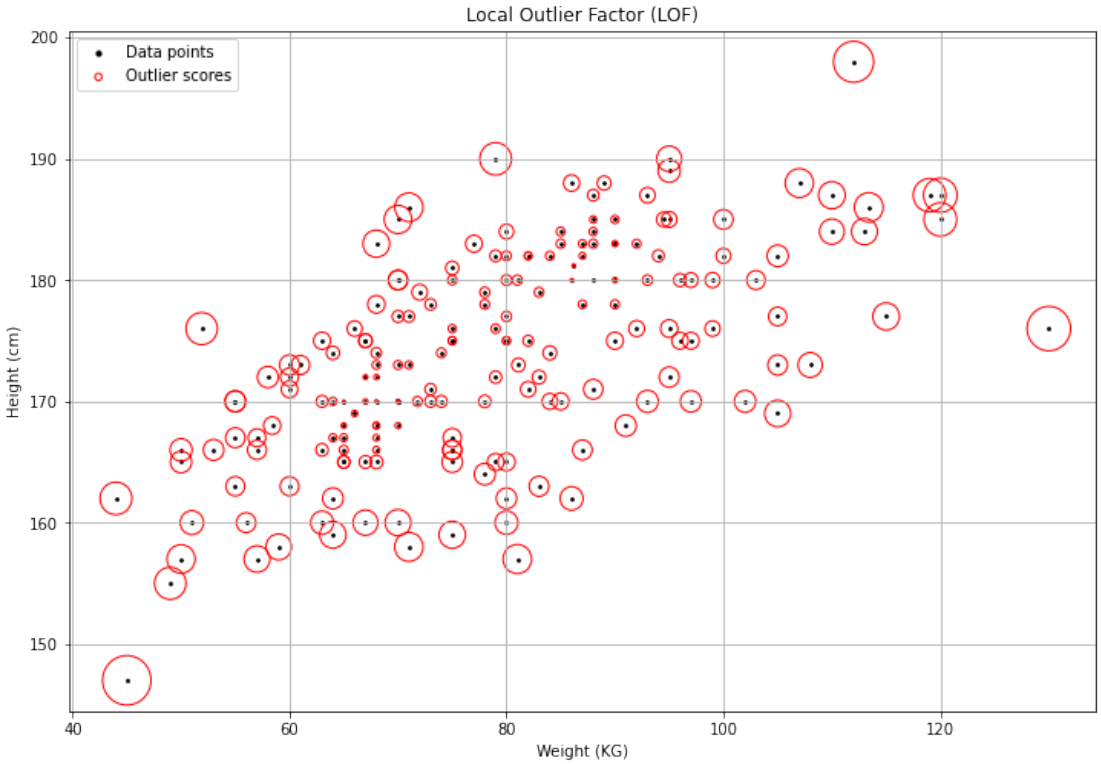

Figure 2.7.1: Example of LOF-based outlier detection with two features from the OxyTarget data set.

$$x_{\text{norm}}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}} \tag{2.29}$$

2.7.4.2 Standard Scaling

When standardizing the feature columns are centered with the mean at 0 and the standard deviation at 1. When performing min-max scaling, we are transforming the values into their corresponding z-score. This is done in the same way as when calculating the z-score outliers in equation 2.28 [35].

2.7.5 Power Transformation

Box and Cox introduced the Box-Cox power transformation in 1964 [48]. The Box-Cox transformation is given in equation 2.30. λ is known as the power parameter and is found using maximum likelihood and goodness of fit tests.

$$x_{\text{Box-Cox}}(\lambda) = \begin{cases} (x^{\lambda} - 1)/\lambda, & \text{if } (\lambda \neq 0) \\ \log(x), & \text{otherwise} \end{cases} \tag{2.30}$$

Yeo and Johnson introduced the Yeo-Johnson power transformation as an extension to the Box-Cox power transformation. This method is capable of considering negative input values as well. The Yeo-Johnson transformation is given in equation 2.31 [49].

$$x_{\text{Yeo-Johnson}}(\lambda) = \begin{cases} \frac{(x+1)^{\lambda}-1}{\lambda}, & \text{if } (\lambda \neq 0, x \geq 0) \\ \log(x+1), & \text{if } (\lambda = 0, x \geq 0) \\ \frac{-((-x+1)^{2-\lambda}-1)}{2-\lambda}, & \text{if } (\lambda \neq 2, x < 0) \\ -\log(-x+1), & \text{if } (\lambda = 2, x < 0) \end{cases} \tag{2.31}$$

CHAPTER

THREE

METHODS

3.1 General Workflow

In this thesis, we modelled the overall survival time (OS) for the OxyTarget and HNC data sets using Aalen's additive fitter (AAF), Cox proportional hazards and accelerated failure time models. The goal was to compare the performance of these models, measured using the concordance index and Brier score.

Figure 3.1.1 shows the experimental setup we used to achieve this.
We started by introducing the data sets, and then conducting pre-processing and data exploration. After these two modules were completed, the data was prepared for modelling.
The experimental setup does not allow us to inspect each model individually. This is because we used a cross-validation solution with 100 folds. This solution gave 100 of each model per data set. The intention of this experimental setup was to evaluate the measured performance. Because of the high amount of features, we included a penalty term in each model.

We did some inspection to see if we were able to fulfil the proportional hazards assumptions on a model trained on the entire data set. We also checked if the assumption of the underlying parametric distributions was reasonable or not.
After this assessment, we applied the cross-validation method to calculate performance metrics on the models for each fold.
The final output was the results from the cross-validation.

3.2 Data Sets

In this thesis, we applied survival analysis methods to a rectal cancer and a head and neck cancer data set. These data sets are described in this section.

3.2.1 Rectal Cancer Data

In this thesis, we utilised a data set from the "Functional MRI of Hypoxia-mediated Rectal Cancer Aggressiveness" study. It is also referred to as OxyTarget, and it can be identified by the clinical trial number NCT01816607. This is an observational cohort study with the primary objective measure: identify the

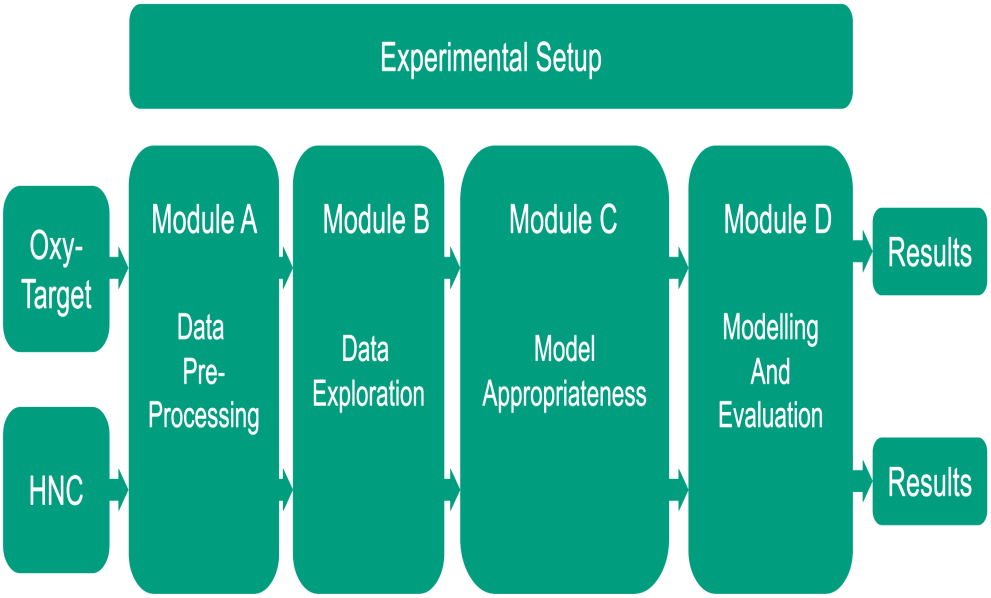

Figure 3.1.1: This figure shows the experimental setup for this thesis.

"presence of metastatic disease 5 years after rectal cancer treatment", with a time frame of 5 years. The study included 192 patients, where 7 of the patients withdrew their consent. There are 185 patients left in this data set for us to study, before the removal of outliers.

3.2.1.1 Inclusion Criteria

The following are the inclusion criteria for the OxyTarget trial [50]:
The patients had to be:

- Willing and able to consent.
- At least 18 years old.
- Scheduled for radical surgery. Either surgery or alone, or with preoperative CRT.

And also have:

- A confirmed rectal cancer diagnosis.
- No prior treatment for rectal cancer.
- At least 60 ml/minute creatinine clearance.
- Signed written informed consent.

The consent protocol was approved by the Regional Ethics Committee. It is worth mentioning that not all observations went through their surgery.

3.2.1.2 Exclusion Criteria

Patients with the following attributes were excluded from the trial [50]:

- Unable to receive MRI due to contradictions to the contrast agent or the MRI itself.
- Wanted to withdraw their consent, for any reason.

3.2.1.3 Response Variables

In this thesis, we are interested in modelling OS for the OxyTarget data set. When considering the OS duration, we measure the time from random assignment to either death or censoring. This includes the death of any source, whether relevant to the study or not [51]. For the OxyTarget data set, we define the response variable OS as the time from inclusion until the time of death for patients experiencing the event, and until the last time point in time registered alive for censored observations.
We define the OS-event response variable as a binary category, 1 for death and 0 when censored.
In this data set, 4 patients died from causes other than rectal cancer. These patients are not excluded from our study, because OS considers any source of death as the OS event.
Some of the features contain information about the patients that can have been acquired after the inclusion time. This is because some of the data was gathered at the time of surgery. This can introduce some bias, and it is important that we consider it when interpreting the results.

3.2.2 Head and Neck Cancer Data

This data set contains both clinical features and features extracted from PET images. The data set consists of 197 HNC patients. These patients were referred to curative CRT at Oslo University Hospital in the time period 2007 to 2013 [52] [53].

3.2.2.1 Inclusion Criteria

The following are the inclusion criteria for the trial, described by Moan et. al. [52]:
The patients had received curative radiotherapy or radio-chemotherapy for squamous cell carcinoma of the oral cavity, larynx, oropharynx or hypopharynx. The patients also needed to have available plans for RT based on FDG PET/CT.

3.2.2.2 Exclusion Criteria

The patients with the following attributes were excluded, as described by Moan et. al [52]:

1. Suffering from nasopharyngeal cancer.
2. Scheduled post-operative RT without residual tumour.

3. Known distant metastases prior to treatment.

Patients without contrast-enhanced planning CT were also excluded [53].

3.2.2.3 Response Variables

For Head and Neck cancer, the overall survival was measured from the first day of radiotherapy until the occurrence of OS-event or censoring [52]. OS-event is a binary response variable, where the value of 1 indicates that the observation experienced death and 0 for censored observations.

3.3 Data Pre-Processing

Figure 3.3.1 shows the general structure for the data pre-processing in this thesis. This is module A in the experimental setup 3.1.1. The input for this module is the raw/unprocessed data sets. We processed the data sets separately, but with similar approaches where applicable.
The output for this module is the pre-processed data sets.

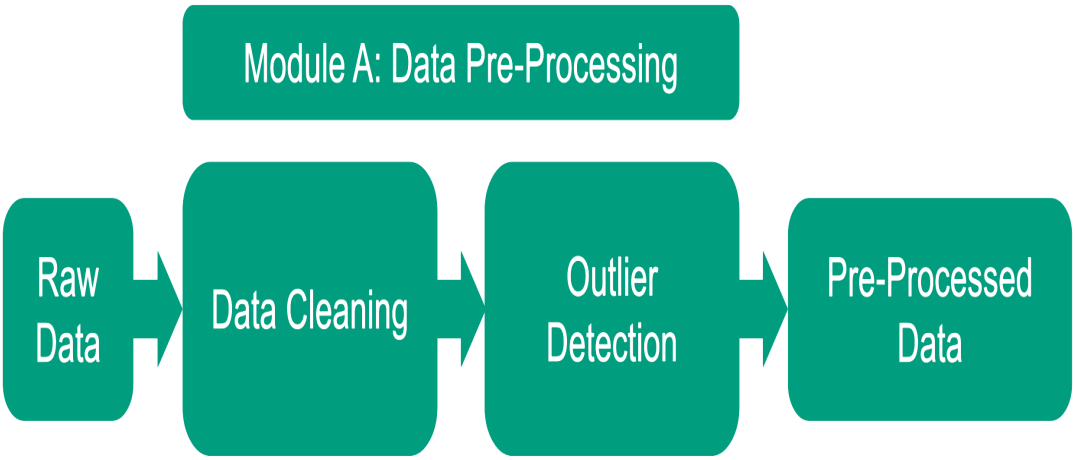

Figure 3.3.1: The figure shows the data pre-processing workflow. This is module A in the general workflow 3.1.1

3.3.1 Data Cleaning

In this section, we discuss the data cleaning process. This process can be divided into three main activities: **removing data**, **formatting columns** and **handling missing values**. Some would also argue that the process of removing outliers would be a part of the data cleaning, but for this thesis, this is listed under the data exploration section.

3.3.1.1 Remove Data

The first thing we did after loading the data set was to remove the patients who had withdrawn their consent.

We separated response variables from the data sets. Any other columns that could indicate the results were then also removed. Columns with uninterpretable values, such as dates and string comments, were also removed from the data set.

We also removed some columns with redundant information and the columns which had more than 25% of the data points missing.

3.3.1.2 Format Columns

Some of the columns in the data set had some single values that were difficult to interpret. These values were encoded to be the right data type.

We inspected the distribution of categories across all categorical features. Categories that occurred only a couple of times were grouped up into a new category "other". This is because a category with few occurrences might not bring enough information to be worth the complexity. It can also confuse the models, especially if the category is so sparse that it does not occur once in the training split. We also considered the missing values in this distribution, and some of the missing values were encoded to a separate binary feature. Other missing values were handled as described in section 3.3.1.3.

Ordinal categories, where the order does matter, were merged into one. Typically, this meant that a group with high values got assigned the value of 1, and the group with low values 0. An inspection of the distribution for all the categories was done, including missing values. For some features with missing values, we encoded them into a separate binary feature.

For **nominal categories**, categories where the order does not matter, we one-hot encoded the features into either single binary columns or multiple.

3.3.1.3 Handling missing values

As discussed in section 3.3.1.2, there was done an inspection of the distributions of the categories, including the missing values. Some of the missing values were interpreted as information and were categorized as a feature on their own. This included "type_of_surgery" in the OxyTarget data set and "hpv_related" and "..." in Head and Neck. Most of the missing values in the "type_of_surgery" column were the patients who had no surgery. For these patients, we created a new feature called "no_surgery". The missing values in the columns "hpv_related" and "..." in the head and neck data set was also related. These missing values were the ones where the HPV status was unknown. For these values, we created one combined column named HPV status unknown. In the Head and Neck data set, these were the only values missing.

The numerical features were imputed using the KNN imputer. The number of neighbours parameter was set to 5, and the weights parameter was set to 'uniform'.

For categorical features, we used the most frequent method before encoding the categories to binary. This was done because it is not practical to use K-nearest imputation to find categories. The KNN imputation algorithm considers the k nearest neighbours and estimates the missing value to be the mean of them, with the weighting as a user-changeable parameter.
This would mean that a value for a binary category could be any number between 0 and 1. This is easy to fix, just change everything over 0.5 to 1 and 0 otherwise. But when considering multiple categories in a column, it would assign the mean of

the k nearest neighbours. If we had encoded them numerically, let's say categories $0, 1, 2$ and 3, the mean of these values would not make any sense.
A solution to this would be to reduce down to one neighbour, because then the mean would make sense, as the same as the closest neighbour. The downside to this approach is that we only get to compare with one neighbour.
Another solution would be to encode the category to multiple binary categories first, and then after imputing the missing values go over each multi-category set of columns to change them in such a manner that there are no observations that are assigned to multiple categories from the same original feature.
This proved to be too cumbersome and difficult to interpret, so we justified with the worse option as we mentioned above: most frequent.

3.3.2 Outlier Detection

In this thesis, we explored multiple methods for discovering outliers, using the theory described in section 2.7.3.

3.3.2.1 OxyTarget

With the extreme value methods, IQR and Z-score, there was a very high amount of observations flagged as outliers. In this thesis, we were interested in keeping as much data as possible without damaging the model's robustness and performance. Because of this, we explored methods such as "Density-based spatial clustering of applications with noise", or DBSCAN for short [54], LOF [47] and isolation forest [55].
These methods can consider multiple dimensions, instead of just looking at one feature at a time.
The results from the DBSCAN analysis were not suitable for our use case. In order to find a "suitable" amount of outliers with this method, we had to change the parameters to such a degree that the method was no longer sensible to use.
Isolation forest, also flagged an unsustainably high amount of outliers.
With the local outlier factor we were able to find 10 outliers, which is 5.4% of the observations. For this to be achieved we used the parameters "n_neighbors = 20" and "contamination = 0.05". The default parameters in sci-kit learn [36] are "n_neighbors = 20" and "contamination = 0.1" for reference.

For this data set, we found that the features extracted from a blood sample test at the time of inclusion were the source of many of the continuous variables. Amongst these features were also several potential outliers that frequently got flagged by the Z-score and IQR methods. A subset without these features had far fewer outliers in terms of these methods, but with the LOF method, the subset had the same amount of outliers as the full data set. Although, there was a slight difference in the outliers selected.

3.3.2.2 HNC

With the HNC data set, we did not have the same problem with outliers using the Z-score and IQR methods. Therefore, we decided to use the outliers extracted from these methods.

We assume the reason for this is that the HNC data set consists of far fewer continuous variables compared to the OxyTarget data set.

We decided to remove the outliers based on the Z-score method after separating the data set by sex. This way allowed female and male observations to have different distributions.
In this thesis we also trained our models on the data sets consisting of only PET features and only clinical features. As mentioned in section REF this is in order to understand the importance of the different subsets, and is only used as additional information in this thesis. The methods for detecting outliers for these subsets are the same as with the full data set, except that the PET-only subset is not separated by sex first. This is because that information is not available in that particular subset without extracting information from the clinical subset.

3.4 Exploratory Data Analysis

In this section, the methods for exploring the data sets are described in detail. Figure 3.4.1 shows the workflow for the exploratory data analysis. This section represents module B in the experimental setup 3.1.1.
In this module we considered the pre-processed data. The data was explored in order to determine feature correlations, before conducting a method for reducing the feature space.
We also performed a univariate analysis. The results from this analysis are presented in the results and discussion chapter 4.

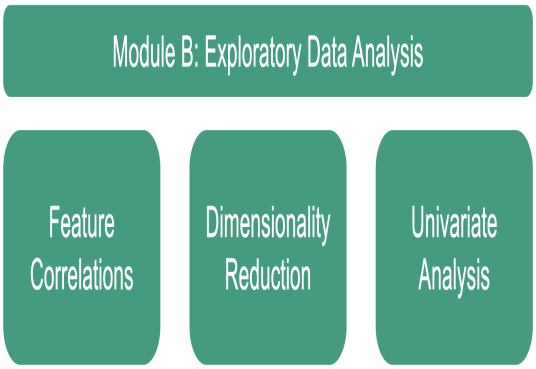

Figure 3.4.1: This figure describes module B in the experimental setup 3.1.1

3.4.1 Feature Correlations

To get an overview of the correlation between the predictor variables, we constructed correlation matrices and visualised them using heatmaps.
The heatmap for Oxytarget is shown in figure 3.4.2. It shows that there are some features with high correlation. Some examples of this are weight and BMI, and the features that both measure the distance between the tumour and anus, either with MR or rigid rectoscopy. For both these cases, some of the underlying information is described by both features.

Figure 3.4.3 shows that the PET-features in the HNC data set are highly correlated.

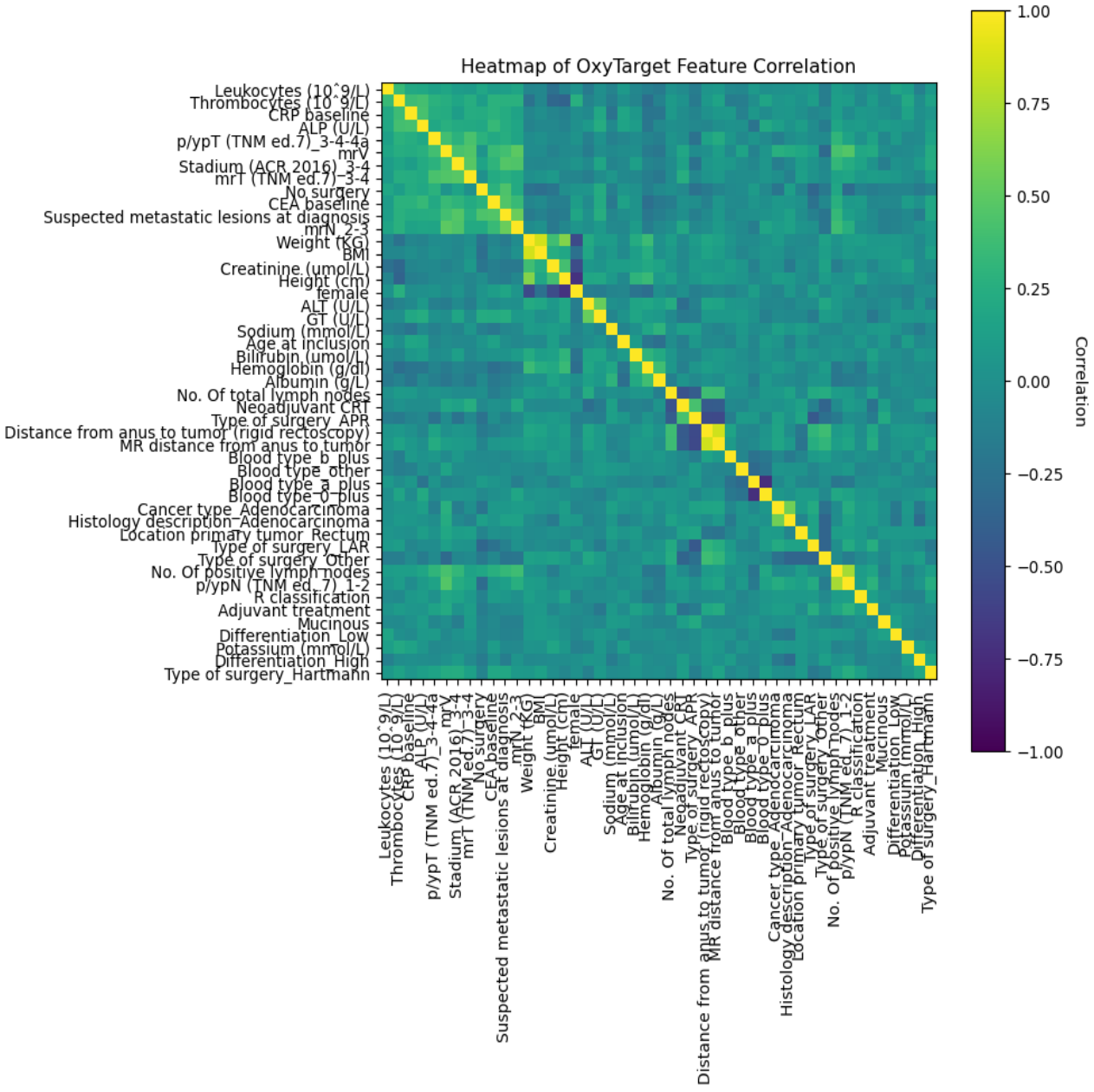

Figure 3.4.2: A heatmap of the correlation matrix for the OxyTarget data set.

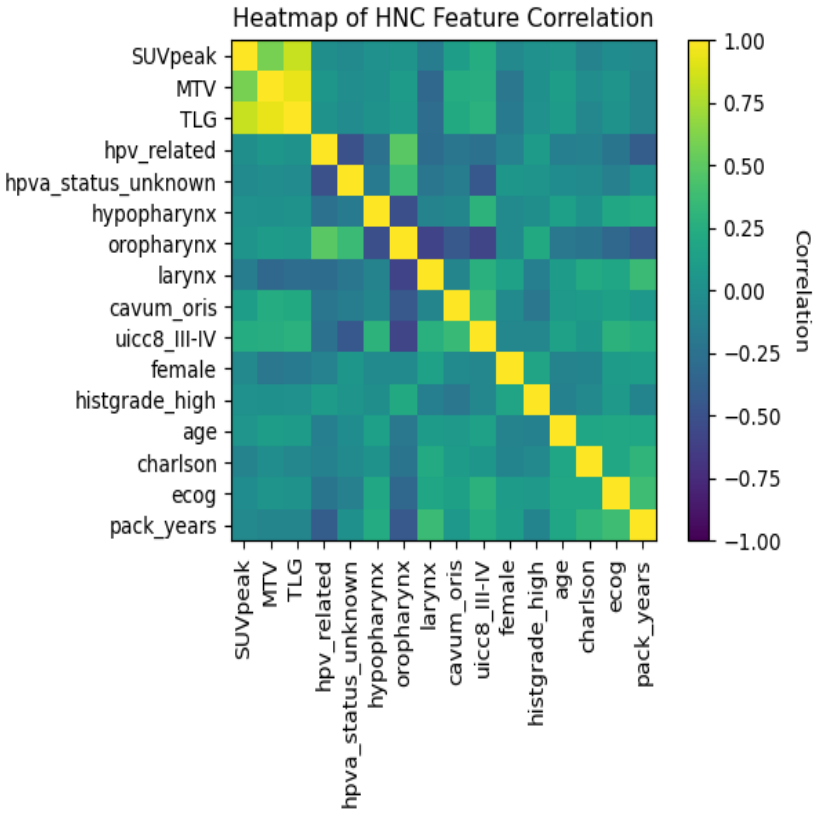

Figure 3.4.3: A heatmap of the correlation matrix for the HNC data set.

3.4.2 Dimensionality Reduction

In section 3.4.1 we described that the OxyTarget data set had features with high correlation. A result of this was that we decided to conduct some dimensionality reduction. The goal of this is to reduce the amount of correlation in the feature set and to reduce the number of features.
This problem was not as present when analysing the HNC data set. This data set also had a lower amount of features, and a higher amount of observations in comparison.
We decided to only reduce the dimensionality of the OxyTarget data set.

3.4.2.1 Cluster Analysis

The selected method for performing feature selection is by using clustering analysis. The goal of the clustering analysis is to select meaningful features that describe data with a lower amount of "overlapping" information. An unsupervised clustering analysis will allow us to group features together in clusters. We can then select one feature from each cluster.
We performed agglomerative hierarchical clustering analysis. This method starts with one cluster per feature and groups the two closest clusters into one. This is repeated until all features are in one cluster. We use Ward linkage to determine the distance between each cluster. This process is utilised on a distance matrix based on the Spearman rank-order correlations. Figure 3.4.4 shows the dendrogram from the hierarchical clustering analysis. We selected the threshold distance 0.9 in order to reduce the number of features to 17. We found this threshold by looking at the clusters in the dendrogram. With 17 features, we would have approximately the same amount of features in both data sets. We chose such a "high" amount of features in order to increase the probability of including important features. The features are shown in table univariate results table 4.1.1.

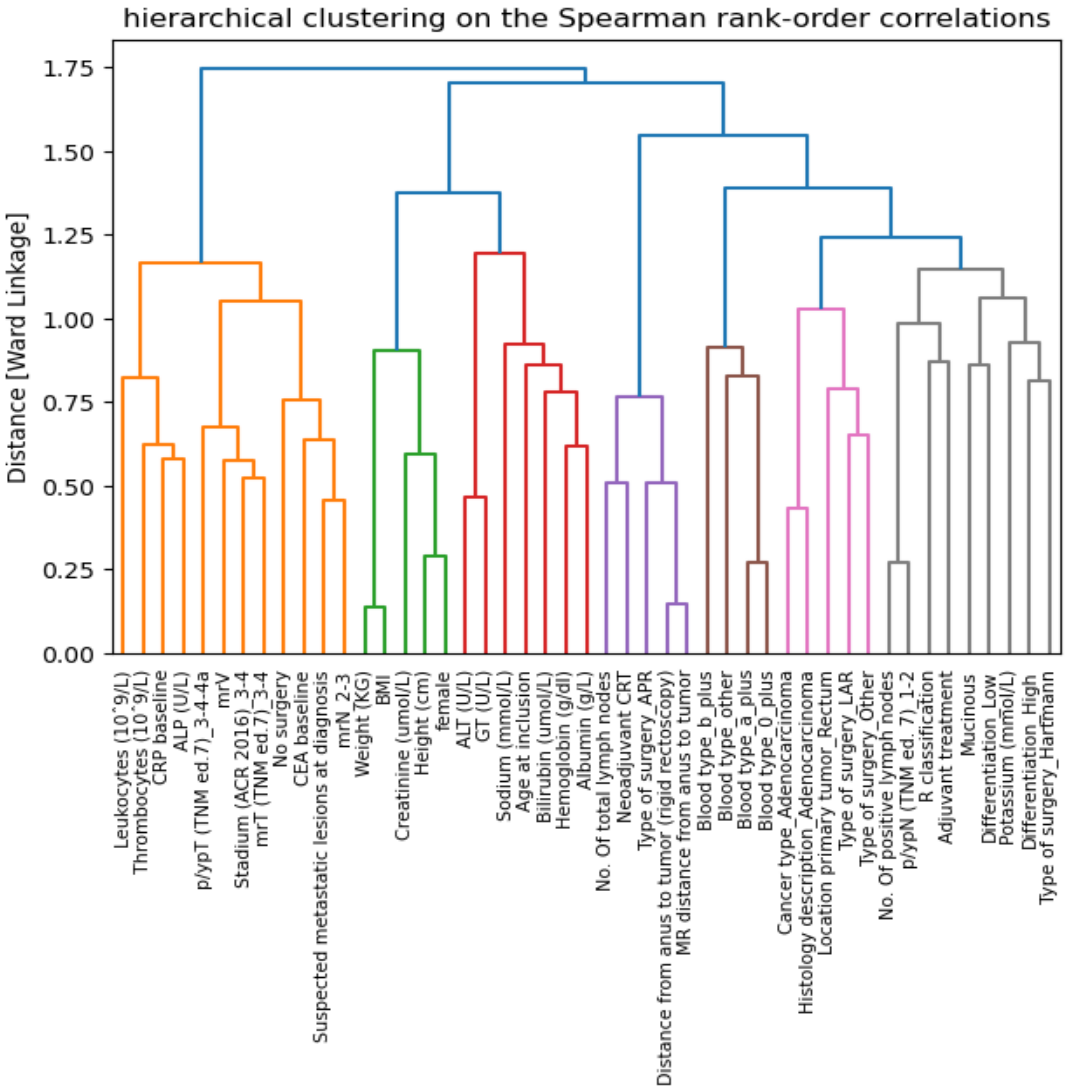

Figure 3.4.4: Dendrogram of the hierarchical clusters based on the ward linkage of the Spearman correlation rank.

3.5 Univariate Analysis

In the univariate analysis, we took a look at the discriminative effect of each feature in terms of the overall survival response variable. We binary encoded the continuous features by separating at the median. We then used the logrank statistical test in order to compare the hazard distributions between the groups separated by each feature.

We also trained a univariate CPH model for each of these features. We used a repeated stratified k-fold method with 4 splits and 25 repeats to validate the results. Harrell's C-index was calculated on the test split for each fold.

We did not use the information learnt by this analysis to make decisions in the modelling process. Because of the experimental setup, we had to be careful and not use information about the unseen test split to train the models. For the OxyTarget data set, we conducted this test on the selected features from the clustering analysis. For the HNC data set, we applied this test to all the predictor variables.

3.6 Model Appropriateness

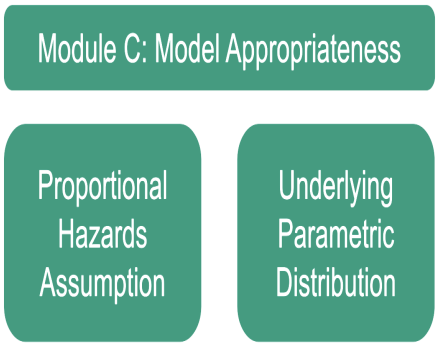

Figure 3.6.1: This figure represents module C in the experimental setup 3.1.1

This section describes module C in the experimental setup 3.1.1. In the theory chapter, we described multiple popular models for survival analysis and some of the assumptions they make. This section describes how we attempt to comply with the proportional hazard assumption. We also attempt to assess whether the chosen models might be a good choice for the task at hand or not.

A weakness in the experimental setup is that we are not able to assess the appropriateness of the models for each fold in the cross-validation. The idea behind this assessment is to verify which of the models is suitable for the problem at hand, not each individual model. In the future, it would be interesting to apply another experimental setup which would allow for this.

For all models and data sets in this thesis we assume non-informative and random censoring.

3.6.1 Proportional Hazards Assumption

As described in section 2.3.2, when the model fails to fulfil the proportional hazards assumptions its coefficients can be less reliable.
The lifelines python library [21] has a built-in method for testing this assumption, with a similar implementation as the function "cox.zph()" from the survival R-library [56]. The tests are calculated by correlating the scaled Schoenfeld residuals and a transformation of time, with the default setting being based on the Kaplan-Meier estimator. The lifelines implementation allows us to detect all features that fail to reject the statistical test given the selected statistical significance level.

$$h_0 = \text{Proportional hazard ratio over time}$$

$$h_1 = \text{Not proportional hazard ratio over time}$$

We used the statistical significance level of $\alpha = 0.05$ and conducted a graphical inspection of the scaled Schoenfeld residuals when the h_0 is rejected.
This procedure was performed on both data sets. We used all observations (except for outliers), the full feature subset for HNC, and the feature subset selected by the unsupervised clustering algorithm for OxyTarget.
In this thesis we compare the models' performance using cross-validation. We are not able to perform this procedure for each fold. With this test, we therefore only assess whether this assumption is somewhat reasonable to make.

3.6.2 Assumption of Underlying Distribution

In this thesis, we are interested in comparing AFT models with the semi-parametric model CPH and the non-parametric model AAF.
We wanted to check whether the assumption about the underlying distribution was justified. In order to do this, we compared the AIC score. We also constructed Q-Q plots with the fitted model quantiles plotted against the empirical quantiles.

3.7 Model Evaluation

This section describes the methodology for evaluating the models. The models we are interested in comparing are: Cox proportional hazards model, Aalen'd additive fitter (AAF), Weibull accelerated failure time model (WAFT) and lognormal accelerated failure time model (LNAFT).

The response variable we are modelling is overall survival (OS). We start by describing the performance metrics and how we calculate them. The next step is to transform and scale the features. Finally, we describe the method for cross-validation.
These steps are shown in figure 3.7.1, which represents module D in the experimental setup 3.1.1.

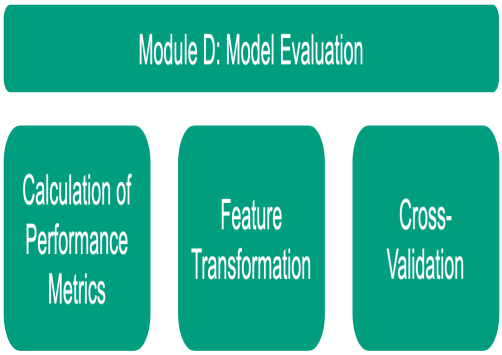

Figure 3.7.1: This figure represents module D in the experimental setup 3.1.1

3.7.1 Calculation of Performance-Metrics

3.7.1.1 Concordance index

We calculate Harrell's concordance index, for both training and test predictions. This was to get an idea of the degree of over-fitting in the model. A model with a high training score and low test score indicates an overfit.
The Uno's concordance index was only calculated on the test data. This metric takes into account the distribution of censored data for both test and train splits. Therefore, we did not find it appropriate to use it on the train split.

3.7.1.2 Brier Score

We calculate the Brier score for the time points between 1 and 5 years. This is 12 to 60 months in the time scale of our data sets. This is because we are interested in evaluating the overall performance of the first 5 years. For CPH and the AFT models, we calculated the Brier score for each month in this period.
The lifelines implementation of the AAF has not implemented a method for calculating survival curves at specific time stamps. For this model, we only calculated the Brier score at the points in time at which the AAF model did calculate the survival probability.
For CPH and the AFT models we also calculated IBS for the specified time period.

3.7.2 Power Transformation and Scaling

We power transform the continuous variables and scale them so that they have a mean of 0 and a standard deviation of 1. This is done using the Yeo-Johnson method [49].

3.7.3 Cross Validation

As described in section 2.6.2, a cross-validation method can give a more robust estimate of the models' performance. Because of the small sample sizes of the data sets used in this thesis, we use this method to get a better estimate of the model's performance.

The method used for cross-validation in this thesis is **repeated stratified k folds** with 4 folds repeated 25 times. This gives a total of 100 folds.
The data is stratified by the survival status in order to give a sufficient amount of censored and uncensored observations on both sides of each split.

In this thesis, cross-validation splits are created and stored for each data set in a separate document. This is done so that all models will train on the same samples. We do our best to keep information about the response variables from leaking between the folds. This makes sure that the information learned from one training split is not used to predict the results for the same data that the information has been learned from. For unsupervised feature selection methods, this does not necessarily apply, because the information about the response variables remains unseen during this process. However, one could argue that the covariates in the test data set are affecting the features selected.
Before fitting the models, we also apply transformations as described in section 3.7.2. Because of the high amount of features in relation to the number of observations, we added an L2 (ridge) penalty of 0.1 to each model.

For each fold in the cross-validation, we calculated the performance metrics as described in section 3.7.1.

3.8 Tools and Software

In this thesis, we used multiple software packages.
For creating and evaluating survival models in Python, we used the Lifelines library [21]. We also used sci-kit-survival [20] for some evaluation metrics. Some methods from Pysurvival [57] were also used.

For pre-processing and cross-validation, sci-kit-learn [36] was integral. Pandas [58] were used for data handling.
The python survival libraries, and also some R packages (R survival [56] and mlr3 [59]) provided useful manuals and instructions.
For general data exploration and manipulation, we used sci-kit-learn [36] and hoggorm [60].
The full list of Python libraries used in this thesis can be found in the GitHub repository A.

CHAPTER
FOUR

RESULTS AND DISCUSSION

4.1 OxyTarget

4.1.1 Univariate Analysis

Table 4.1.1 shows the results of the univariate data analysis of the OxyTarget data set.
The feature with the highest measured concordance index is "Suspected metastatic lesions at diagnosis". Metastasis occurs when the tumour is spreading. This feature achieved an average concordance score of 0.71, with a standard deviation of 0.05. This result tells us that this feature alone is quite descriptive in terms of overall survival. "No. Of positive lymph nodes > 4" also received a high concordance score.
The logrank test failed to reject the h_0 for 5 of the features in this data set.
The features used in this analysis are the ones selected from the clustering algorithm.
The column "Group distributions" shows how many observations there are in each group. "Events per group" show how many OS events occurred for each group.

Table 4.1.1: Univariate analysis of OxyTarget, with logrank test and univariate CPH.

Feature		Logrank Test		Univariate CPH	Group Distributions		Events Per Group	
Name	**Type**	**P-Value**	**Null-Hypothesis**	**Harrell's C-index**	**0**	**1**	**0**	**1**
Suspected metastatic lesions at diagnosis	Binary	<0.001	Reject	0.71 +- 0.05	142	33	25	26
No. Of positive lymph nodes > 0.4	Separated by median	<0.001	Reject	0.68 +- 0.06	96	79	11	40
mrV	Binary	<0.001	Reject	0.63 +- 0.06	98	77	18	33
CRP baseline > 2.6	Separated by median	0.032	Reject	0.59 +- 0.05	89	86	20	31
Weight (KG) > 78.0	Separated by median	0.086	Not rejected	0.56 +- 0.07	90	85	31	20
R classification		<0.001	Reject	0.56 +- 0.04	164	11	43	8
Distance from anus to tumor (rigid rectoscopy) > 8.0	Separated by median	0.165	Not rejected	0.54 +- 0.08	88	87	30	21
Differentiation_High	Binary	0.291	Not rejected	0.53 +- 0.04	151	24	46	5
Location primary tumor_Rectum	Binary	0.198	Not rejected	0.53 +- 0.03	15	160	2	49
Blood type_b_plus	Binary	0.376	Not rejected	0.52 +- 0.02	164	11	49	2
Mucinous	Binary	0.387	Not rejected	0.51 +- 0.03	163	12	49	2
Cancer type_Adenocarcinoma	Binary	0.486	Not rejected	0.51 +- 0.03	18	157	4	47
Potassium (mmol/L) > 4.3	Separated by median	0.418	Not rejected	0.49 +- 0.08	90	85	29	22
Age at inclusion > 65.0	Separated by median	0.587	Not rejected	0.48 +- 0.06	88	87	24	27
Sodium (mmol/L) > 139.0	Separated by median	0.608	Not rejected	0.47 +- 0.07	88	87	27	24
ALT (U/L) > 30.0	Separated by median	0.710	Not rejected	0.47 +- 0.06	95	80	29	22
Blood type_a_plus	Binary	0.792	Not rejected	0.47 +- 0.06	91	84	27	24

4.1.2 Model Appropriateness

4.1.2.1 Proportional Hazards Assumption

The test for proportional hazards did not reject the h_0 for any of the features in the OxyTarget data set.

4.1.2.2 Parametric Distribution Assumption

Table 4.1.2: AIC - scores for OxyTarget distributions.

Weibull	LogNormal	LogLogistic	Exponential
640.4	634.8	638.2	641.1

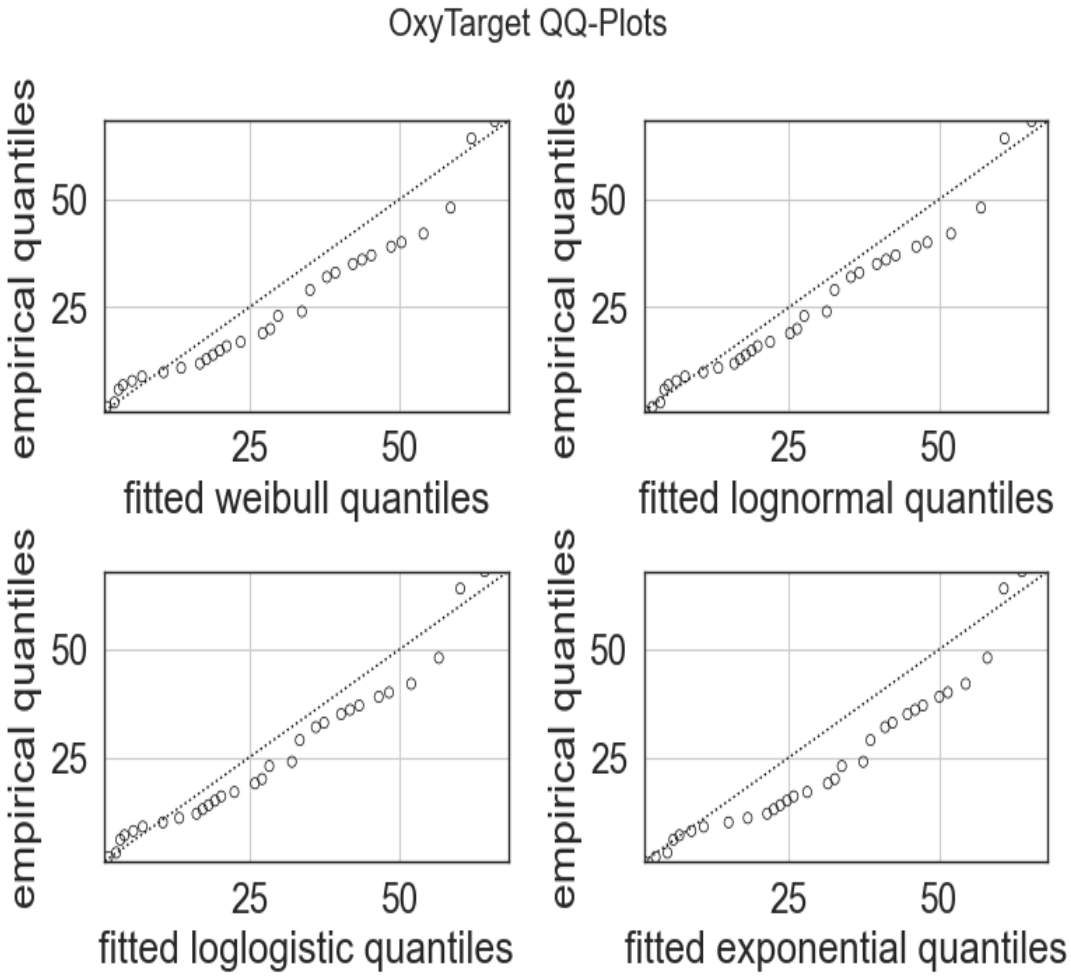

Figure 4.1.1: Q-Q plots comparing parametric distributions for OxyTarget survival distribution.

From figure 4.1.1 it is difficult to tell which distribution is the best fit. The

lognormal and log-logistic models both perform very similarly. To determine the best-suited distribution, we decided to look at the AIC scores.

Table 4.1.2 shows that the lognormal has a slightly lower score which indicates that this is the best-suited distribution to describe the overall survival time in the OxyTarget data set.

4.1.3 Performance Metrics

Table 4.1.3 shows the results for the OxyTarget data set. The results presented are Harrell's C-index, Uno's C-index and integrated Brier score (IBS). All metrics were calculated on the test split, and Harrell's C-index was calculated on the training split as well. The IBS was calculated based on the predictions for each month in the time period of 12 to 60 months.
Aalen's additive fitter (AAF) was not able to predict survival probabilities for the desired time points. Therefore, the IBS score was not calculated for this model.
It is important to note that the LNAFT model was selected after comparing the distributions of the survival times and that the results for this model are to be interpreted as validation scores.

4.1.3.1 C - Index

Cox proportional hazards performed the best in terms of measured concordance index. This is true for all three measurements.
The log-normal AFT model (LNAFT) and the Weibull AFT model (WAFT) performed very similarly. The LNAFT had a higher training score, which suggests that this model had a higher degree of overfit than the WAFT model.
The AAF model performed better than the AFT models on the test data, but worse on the training data. This could be an indication that the AAF model was less overfit than the other models. This also includes CPH which had a difference of 0.07 between the test and train scores.
All models achieved lower scores when the censoring distribution was adjusted for, estimated with Uno's concordance.
These findings suggests that CPH had the highest number of concordant predictions.

4.1.3.2 Integrated Brier Score

An IBS score of 0.25 represents an uninformative model.
CPH, WAFT and LNAFT achieved very similar IBS, well below the worst-case performance threshold. This indicates that the models have a similarly good fit.

Table 4.1.3: Performance metrics from the OxyTarget data set.

Performance Metric	CPH	AAF	WAFT	LNAFT
Harrell's C (train)	0.851 +/- 0.012	0.773 +/- 0.019	0.794 +/- 0.021	0.807 +/- 0.020
Harrell's C (test)	0.780 +/- 0.049	0.735 +/- 0.051	0.724 +/- 0.057	0.726 +/- 0.052
Uno's C (test)	0.767 +/- 0.055	0.726 +/- 0.053	0.715 +/- 0.060	0.714 +/- 0.056
IBS (test)	0.131 +/- 0.019	-	0.132 +/- 0.020	0.131 +/- 0.019

4.1.4 Brier Scores Over Time

Figure 4.1.2 shows the Brier scores. They were calculated for each month in the time period of 12 to 60 months. AAF was not able to predict the survival probabilities for this interval. For this model, we presented the measured Brier scores for the time points the model was able to estimate the survival probability.

The models achieved similar results in terms of the measured Brier score. The AAF model has fewer data points. It is therefore more difficult to interpret the performance. However, the data points available suggest that this model performed slightly worse than the others.
The rising trend in the Brier score suggests that the predictions get less accurate when time increases.
All models had results below the 0.25 threshold for the time interval of measurement. This suggests that the models are informative for this time period.

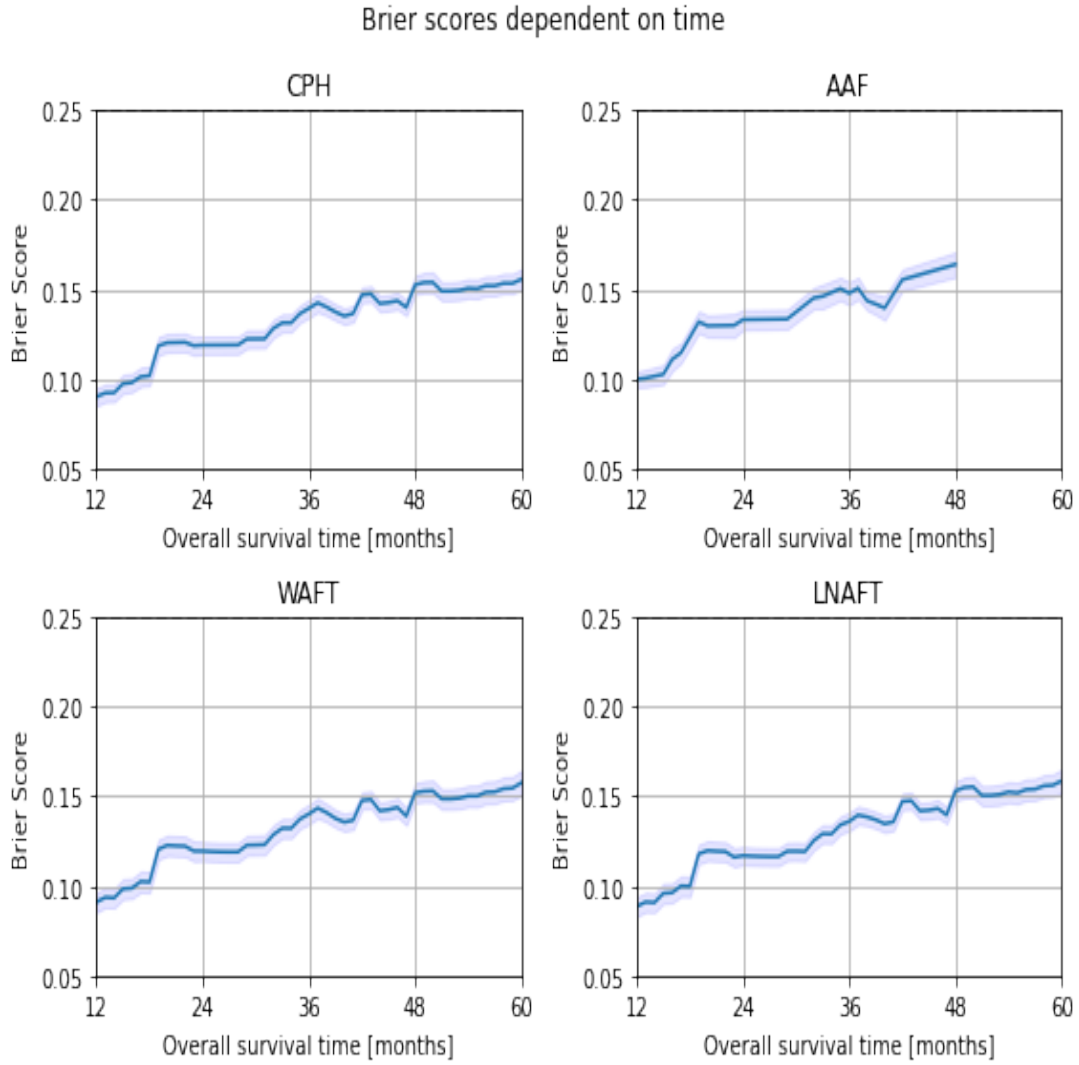

Figure 4.1.2: The figure shows the calculated Brier score with a 95% confidence interval.

4.2 Head and Neck

4.2.1 Univariate Analysis

The results from the univariate analysis are shown in table 4.2.1. Using the logrank test, we were able to reject the h_0 for 11 of the 16 features. This means that there is a significant difference in the hazard distributions between the groups separated by 11 of these features. This told us that there are several features with potential in terms of modelling the overall survival time.
The results from the univariate CPH model support this finding.
The column "Group distributions" shows how many observations there are in each group. "Events per group" shows how many OS events occurred for each group.

Table 4.2.1: Univariate analysis of Head and Neck, with logrank test and univariate CPH.

Feature		Logrank Test		Univariate CPH	Group Distributions		Events Per Group	
Name	**Type**	**P-Value**	**Null-Hypothesis**	**Harrell's C-index**	**0**	**1**	**0**	**1**
uicc8_III-IV	Binary	<0.001	Rejected	0.68 +- 0.05	123	66	28	43
ecog	Binary	<0.001	Rejected	0.67 +- 0.05	126	63	31	40
pack_years > 22.2	Separated by median	<0.001	Rejected	0.66 +- 0.05	95	94	19	52
oropharynx	Binary	<0.001	Rejected	0.64 +- 0.06	49	140	34	37
hpv_related	Binary	<0.001	Rejected	0.63 +- 0.05	112	77	56	15
age > 60.6	Separated by median	0.001	Rejected	0.61 +- 0.05	95	94	25	46
charlson	Binary	0.002	Rejected	0.59 +- 0.05	127	62	39	32
MTV > 6.9	Separated by median	0.072	Not rejected	0.56 +- 0.06	95	94	31	40
TLG > 52.9	Separated by median	0.049	Rejected	0.56 +- 0.06	95	94	30	41
hpva_status_unknown	Binary	0.039	Rejected	0.55 +- 0.05	138	51	56	15
larynx	Binary	0.003	Rejected	0.55 +- 0.04	168	21	56	15
cavum_oris	Binary	<0.001	Rejected	0.55 +- 0.03	177	12	61	10
SUVpeak > 10.0	Separated by median	0.154	Not rejected	0.54 +- 0.06	95	94	31	40
hypopharynx	Binary	0.085	Not rejected	0.53 +- 0.03	173	16	62	9
female	Binary	0.316	Not rejected	0.51 +- 0.05	141	48	56	15
histgrade_high	Binary	0.364	Not rejected	0.50 +- 0.06	57	132	24	47

4.2.2 Model Appropriateness

4.2.2.1 Proportional Hazards Assumption

The test for proportional hazards rejected the h_0 for "hpv_related". Figure 4.2.1 shows the Schoenfeld residuals for this feature. By analysing these residuals, we saw that there is some trend. As described in section 2.3.2, these residuals should not be correlated with time.
Because of this finding, we considered a CPH model stratified by this feature, as well as a model that completely ignores this feature. We compared these two models together with the 'regular' CPH without any intervention. Note that because we used information which will lie within the test splits, we considered the results from these two models as validation results rather than test results.
We keep in mind that stratifying the CPH model is not appropriate for small data sets [61].

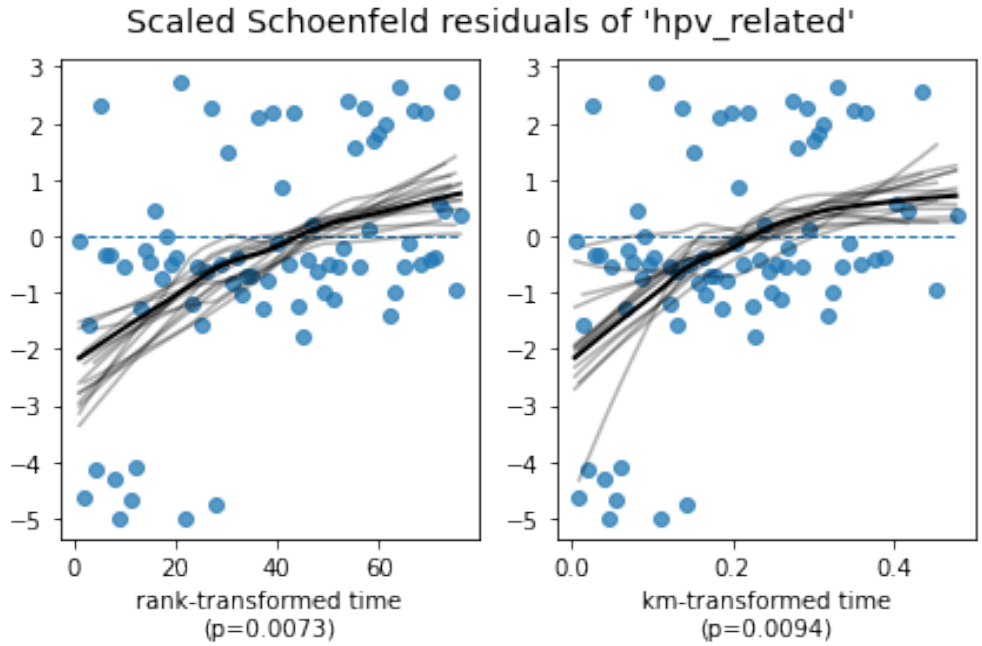

Figure 4.2.1: Schoenfeld Residuals for the feature "hpv_related" in the HNC data set.

4.2.2.2 Parametric Distribution Assumption

Table 4.2.2: AIC scores for HNC distributions.

Weibull	LogNormal	LogLogistic	Exponential
898.1	890.5	895.0	899.8

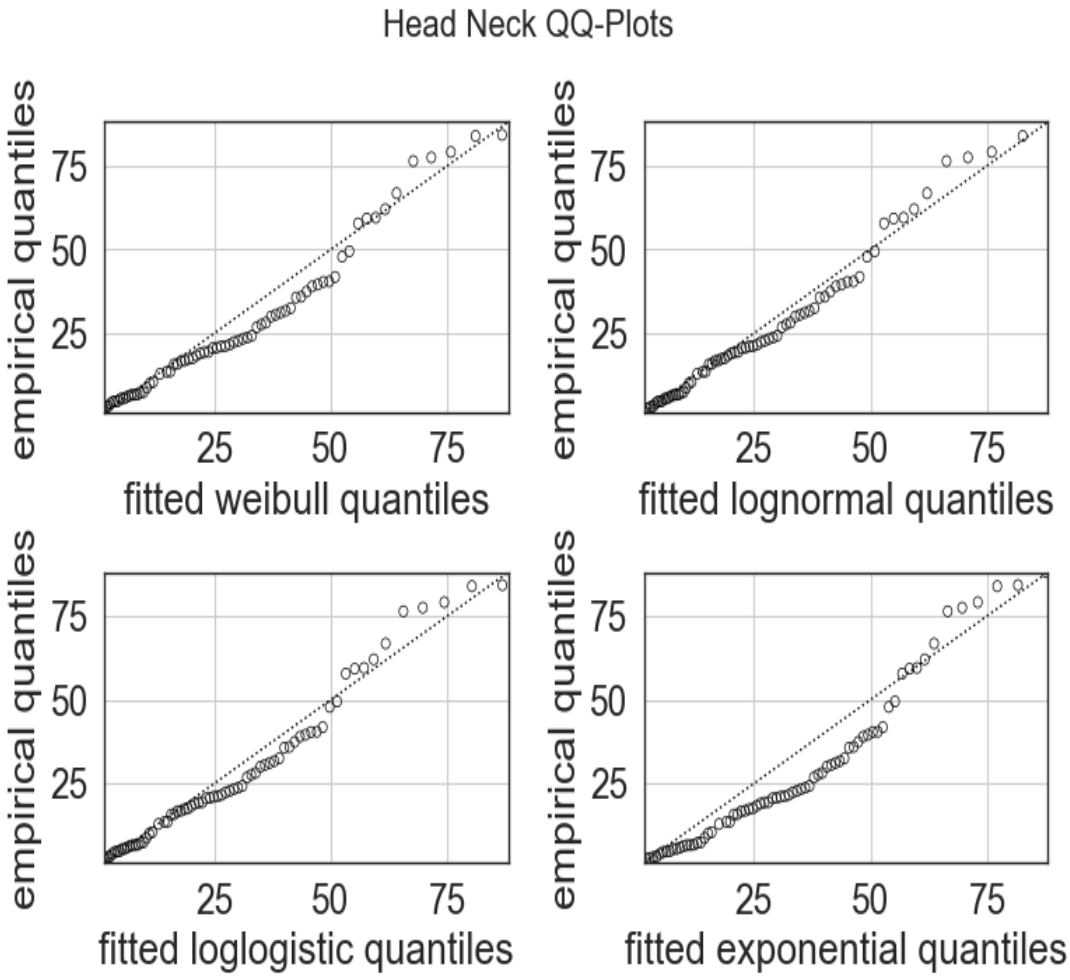

Figure 4.2.2: Q-Q plots comparing parametric distributions for HNC survival distribution.

Figure 4.2.2 shows that the lognormal distribution is the most suitable choice for parametric models for the HNC data set because it aligns slightly more with the 45-degree line.

Table 4.2.2 shows that there is only a slight difference between the suitability of the underlying distributions. Out of these scores, the lognormal is the lowest by a small margin.

Because of these findings, the lognormal distribution is considered to best describe the underlying distribution of overall survival for the HNC data set.

4.2.3 Performance Metrics

Table 4.2.3 shows the calculated performance metrics for the HNC data set. CPH* is the Cox proportional hazards model stratified by the feature 'hpv_related'. CPH** is a Cox proportional hazards model with this feature completely removed. It is important to note that we used information from the entire data set when determining this. These scores should therefore be considered as validation scores. The same is true for the lognormal AFT model, which was selected after analysing the parametric distribution of the survival times. Results for Harrell's C-index are calculated on both training and test data sets, while Uno's C-index and IBS are calculated on the predictions for the test data only. We do not present the IBS score for the AAF model, because this model was not able to predict the survival probabilities for all months in the interval 12 to 60 months.

4.2.3.1 C - index

CPH achieved the highest measured Harrell's C - index. This is true for both the training and test data. This model also had the highest measured Uno's C-index. CPH** had slightly lower scores than this model. CPH* had even lower than both of these two models. WAFT and LNAFT had lower recorded C-index by all measures compared to CPH. AAF achieved the lowest performance measured by the C-scores.
All models had a significant difference between their measured train and test scores, in terms of Harrell's C - index. This suggests that the models might be over-fitting. All models achieved a lower score measured with Uno's C - index compared to Harrell's C-index.
These findings suggests that CPH had the highest number of concordant predictions.

4.2.3.2 IBS

An uninformative model will have an IBS of 0.25. This is almost the case for the CPH* model. The other models performed significantly better, with CPH and LNAFT having a similarly well-performing scores. CPH** and WAFT performed slightly worse. An interesting finding is that the CPH* model achieved to predict

Table 4.2.3: Performance metrics from HNC data set.

Performance Metric	CPH	CPH*	CPH**	AAF	WAFT	LNAFT
Harrell's C (train)	0.818 +/- 0.012	0.799 +/- 0.015	0.814 +/- 0.012	0.781 +/- 0.019	0.797 +/- 0.016	0.804 +/- 0.015
Harrell's C (test)	0.780 +/- 0.041	0.752 +/- 0.050	0.775 +/- 0.040	0.728 +/- 0.058	0.750 +/- 0.051	0.759 +/- 0.045
Uno's C (test)	0.756 +/- 0.058	0.737 +/- 0.058	0.754 +/- 0.054	0.707 +/- 0.066	0.732 +/- 0.059	0.739 +/- 0.057
IBS (test)	0.146 +/- 0.019	0.232 +/- 0.034	0.150 +/- 0.019	-	0.149 +/- 0.020	0.145 +/- 0.020

4.2.4 Brier Scores Over Time

An informative model will have a Brier score of 0.25. This is demonstrated by the dotted line in figure 4.2.3.

All the models had a Brier score starting at approximately 0.10 at 12 months and with a rising trend. This tells us that the predictions get less accurate over time.

CPH* had the steepest trend out of all the models. This model measured considerably worse than the other models. Stratification is not recommended on very small data sets, because it reduces the available training data for each stratum. This can be the reason behind the poor performance. This is interesting because this model achieved relatively well in terms of the C-index score. This suggests that the model's predicted survival probabilities can be poor even though the model is able to discriminate against different risk groups.

As described in the theory chapter, the AAF model is not able to predict the survival probability for all values of t. We can see that the model has fewer data points when t increases. This model performed slightly worse than the best-performing models.

CPH, CPH*, WAFT and LNAFT performed very similarly. These results indicate that these models are informative in terms of the overall survival response variable.

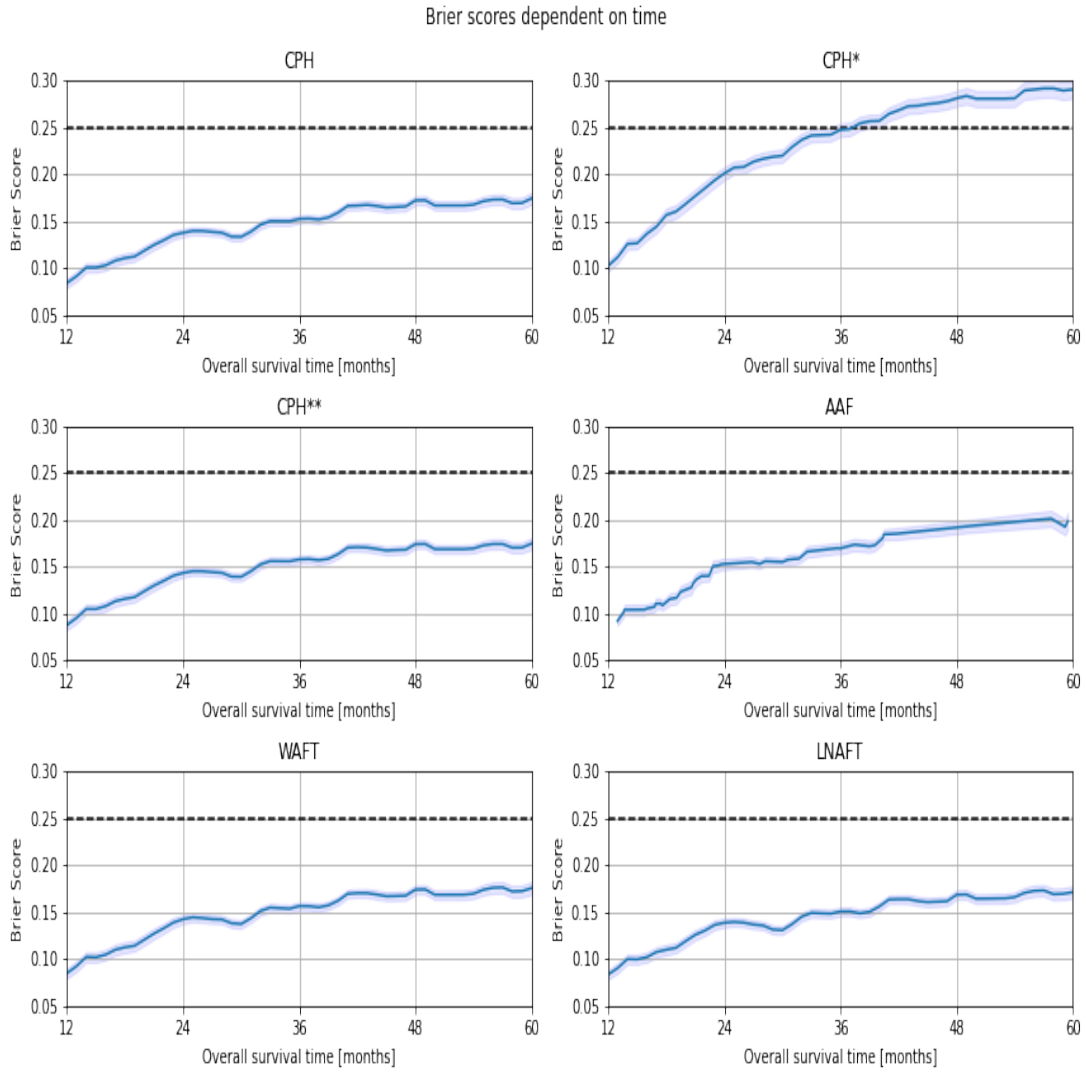

Figure 4.2.3: The figure shows the calculated Brier score with a 95% confidence interval.

4.3 Sources of Error

4.3.1 Data Registration

Different techniques for data registration were used by different doctors. For instance, some data points were unclear whether it was 0 or not registered because of the different registration styles. These data points could include 0, - or simply be left empty.
Other examples of this could be comment-based data entries, where different doctors had different styles of commenting, leaving the resulting data to be difficult to interpret at scale.
For the OxyTarget data set, some of the information was registered after the inclusion date. This can introduce bias and lead to unrealistic estimates of the model performance.

The overall survival was measured from the time of treatment for the HNC data set, and from the time of inclusion for the OxyTarget data set.
For the HNC data set, this allows us to model the overall survival time after the start of treatment.
For the OxyTarget data set, this means that we are modelling the survival time from a random point in time, usually a short time after diagnosis. This allows us to model the survival time given the current state described by the covariates. When information is added after the initial registration, some bias and randomness (due to different amounts of time between inclusion and treatment) are introduced.

4.3.2 Satisfying Model Assumptions

As mentioned in section 3.6, the chosen experimental setup does not allow for assessing the assumptions for each individual model. This is because we had a total of 100 folds. A source of error is therefore the potential violations of these assumptions that are left un-noticed and un-assessed.
In section 2.3.2 we mentioned that some researchers argue whether or not it is necessary to check for the proportional hazard assumption.

4.4 Discussion of Choices

In this section, we describe some of the choices that were made during this thesis, and we discuss some of the reasoning behind these decisions.

4.4.1 Accelerated Failure Time Models

An exponential model would assume a constant hazard. This was not an assumption we were willing to make.
A Weibull model assumes that the hazard changes proportionally with time. We chose the Weibull-based AFT model because it is reasonable to assume that the risk of death is going to increase over time. It is also the most commonly used AFT-based model [2]. Because of these two reasons, we decided to use the Weibull model in this thesis.

We decided to use Q-Q plots and analyse AIC - scores to assess to which degree each distribution fit our response. This would allow us to determine whether the assumption was reasonable or not. We decided to include the best-suited model as well. Because we used information about the response distribution to decide on this model, we have to take this into consideration when interpreting the results. The results for this model are therefore interpreted as validation results, rather than test results.

4.4.2 Method for Evaluation

In this thesis, we worked with some traditional statistical methods and compared them using a data science approach. We worked with very small data sets, which limited the possibilities in terms of the evaluation of the models. Because this thesis's main goal is to measure performance metrics for the different survival models, we had to make some choices in order to find a suitable method for doing so.

We found that many of the available articles focused on the assessment of the feature importance, rather than the model performance. The nature of our experimental setup did not allow us to remove potential confounders, and it made it difficult to inspect each individual model's coefficients. If we were to remove the redundant features, the models would have been exposed to the test data during the training.

4.4.2.1 Cross Validation

In this thesis, we used a cross-validation scheme in order to be able to get a better comparison between these models.
The repeated stratified k-folds method was selected to be able to separate the data into training and test sets with sufficient sizes and also a fairly similar distribution of censored data. This allowed us to be able to extract the aggregate results across all the splits and get some more reliable insights into the performance metrics.

In this thesis, we also considered using other schemes for model evaluation. If the sample size was larger, we would consider separating the data into a train-test split. This scheme would have allowed us for a detailed inspection of the "final" models. A cross-validation scheme could then also allow for hyperparameter tuning and feature selection.

4.4.3 Data Preparation

One of the areas in this thesis where we had to make the most difficult decisions was when preparing the data. There are several things to consider when building a data set for estimation, and it is important to simultaneously both remove redundant and "harmful" information while keeping as much information as possible.

4.4.3.1 Feature Encoding

When encoding the features in the OxyTarget data set we had to make multiple decisions. There were multiple categorical features with more than two categories. Some of the categories had few observations, and some even had no registered OS

events. The most sparse categories were therefore encoded into "others" categories. The HNC data set was already encoded into a suitable format for analysis.
We also had to consider the continuous variables. Some might argue that it can be beneficial to bin some of the continuous variables. Due to the lack of background knowledge, we were not able to distinguish reasonable thresholds and features for this solution. We also had to consider the fact that due to a relatively low amount of observations, it was difficult to get sufficient observations in each bin while also avoiding complete separation in terms of the response variables. A solution could be to separate by percentiles, or by the median. For this thesis, we decided not to do this. Instead, we power transformed and standard scaled the data so that the continuous variables were to be normally distributed with mean 0 and standard deviation 1.

4.4.3.2 Removal of Observations

We considered several methods for the detection of outliers. This is because it was difficult to perform outlier detection without removing too many observations. We were willing to have some observations with potentially "damaging" data points because of this, even though the data sets could be more vulnerable to over-fitting. Therefore we looked at multiple methods and compared the outliers found by all methods. Then we selected a method with a reasonable amount of features that still somewhat cohered with the other methods.

In addition to removing outliers, we had some observations that died from other causes. Due to the definition of overall survival, we decided to keep these observations within the data set.

4.4.3.3 Removal of Features

In this thesis, we did not use any supervised methods for the removal of features. This means that we did not use any methods that consider the response variables. Although it was tempting to hand-select the features which received the highest measured Harrell's C-index in the univariate analysis, we did not do so. This would lead to information bleeding between the validation folds, and we lose the entire purpose of the experimental setup. An exception to this was the choice of including the LNAFT model, which was based on the best-suited parametric distribution. Another exception to this is the models CPH* and CPH**. These models were included in order to analyse whether this would improve the performance.
We considered the removal of redundant features, but with the small sample sizes, we decided to use an experimental setup which did not allow for this.

4.4.3.4 Selection of Transformations

For the OxyTarget data set, we had to consider feature correlation. Some of the survival models were highly sensitive to this, and would not converge before we reduced the feature space sufficiently.
There are multiple methods for reducing the collinearity of a data set, and we decided to try two unsupervised alternatives. These were principal component analysis and hierarchical clustering.

The principal components extracted from the principal component analysis did not give a sufficient amount of explained variance. Therefore we decided to use clustering analysis.
We chose hierarchical clustering based on the Spearman rank correlation. This method allowed for the creation of a dendrogram. Using this dendrogram we found a distance that gave a reasonable amount of features. The first feature in each cluster was selected.
Another master student used RENT, repeated elastic net technique [4] [5] in their thesis on the OxyTarget data set [6]. They considered the binary classification problem for OS and PFS and used medical imaging on the cohort of 81 patients. There is some difference between the features selected by the RENT method and the method used in this thesis. This tells us that some important information might be left out of our analysis.

4.5 Future work

In the future, it would be interesting to use hyper-parameter tuning to find the optimal amount of penalty for each model. An experimental setup which separates the data into train, validation and test splits would be of utmost interest in this scenario. This could allow us to further inspect the results and validity of the final model used on the test set. This hypothetical test setup would prove to be very interesting, and we could even use supervised methods for reducing the feature space. However, this experimental setup might need to be attempted on a data set with a larger sample size.

CHAPTER

FIVE

CONCLUSIONS

In this thesis, we used survival analysis methods to model the overall survival time for patients suffering from rectal and head and neck cancer. Three types of models were considered: Cox proportional hazards, Aalen's additive fitter and accelerated failure time models.
The goal of this thesis was to compare these models using the performance metrics concordance index and Brier score. The Brier score was calculated for each month from 12 to 60 months in order to understand the models' overall performance in the first 5 years. We also calculated the integrated Brier score for this period. For Aalen's additive fitter, we could not calculate the Brier score for the desired points in time. For this model, we calculated the Brier score for the time points available and left out the integrated Brier score.
The performance metrics were estimated using a repeated stratified k-folds cross-validation scheme. We used four splits and 25 repeats for a total of 100 folds. This gave us 100 estimates for the performance of each model per data set.

Cox proportional hazards had the highest achieved performance measured using Harrell's concordance index and Uno's concordance index. This is true for both the rectal cancer data set and the head and neck cancer data set.
The models achieved similar Brier scores on the rectal cancer data set. Aalen's additive fitter performed slightly worse than the other models in terms of this performance metric. All models had a rising trend in the Brier scores, which indicates less accurate predictions when the time increases.

In the future, it would be interesting to use hyperparameter tuning to find the best model parameters. It would also be of interest to apply these methods in combination with supervised feature selection.

REFERENCES

[1] Cancer Registry of Norway. *Cancer in Norway 2021- Cancer incidence, mortality, survival and prevalence in Norway*. Oslo, 2022. URL: https://www.kreftregisteret.no/Generelt/Rapporter/Cancer-in-Norway/cancer-in-norway-2021/ (visited on 04/23/2023).

[2] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. New York, NY: Springer, 2012. ISBN: 978-1-4419-6645-2 978-1-4419-6646-9. DOI: 10.1007/978-1-4419-6646-9. URL: http://link.springer.com/10.1007/978-1-4419-6646-9 (visited on 10/22/2022).

[3] Frank E. Jr Harrell. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Cham: Springer International Publishing, 2015. ISBN: 978-3-319-19424-0 978-3-319-19425-7. DOI: 10.1007/978-3-319-19425-7. URL: http://link.springer.com/10.1007/978-3-319-19425-7 (visited on 10/22/2022).

[4] Anna Jenul et al. "RENT: A python package for repeated elastic net feature selection". In: *Journal of Open Source Software* 6.63 (2021). Publisher: The Open Journal, p. 3323. DOI: 10/gr596n. URL: https://doi.org/10.21105/joss.03323.

[5] Anna Selina Jenul et al. "RENT—Repeated Elastic Net Technique for Feature Selection". In: *152333-152346* (2021). Accepted: 2022-07-14T11:58:54Z. ISSN: 2169-3536. DOI: 10/gr7zrr. URL: https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/3005410 (visited on 05/10/2023).

[6] Lars Jetmund Svartis Engesæth. "Predicting patient outcome using radioclinical features selected with RENT for patients with colorectal cancer". Accepted: 2022-12-06T10:24:15Z. Master thesis. Norwegian University of Life Sciences, Ås, 2022. URL: https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/3036071 (visited on 02/02/2023).

[7] Sofie Fjellvang. "Prediksjon av behandlingsutfall for hode- og halskreftpasienter ved bruk av radiomics og repetert elastisk nett teknikk". Accepted: 2022-08-17T09:11:07Z Journal Abbreviation: Prediction of treatment outcome for head and neck cancer patients using radiomics and Repeated Elastic Net Technique. Master thesis. Norwegian University of Life Sciences, Ås, 2022. URL: https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/3012291 (visited on 05/04/2023).

[8] Farzan Madadizadeh et al. “Applying Additive Hazards Models for Analyzing Survival in Patients with Colorectal Cancer in Fars Province, Southern Iran”. In: *Asian Pacific Journal of Cancer Prevention : APJCP* 18.4 (2017), pp. 1077–1083. ISSN: 1513-7368. DOI: 10/gr596j. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5494219/ (visited on 04/24/2023).

[9] Jesus Orbe, Eva Ferreira, and Vicente Núñez-Antón. “Comparing proportional hazards and accelerated failure time models for survival analysis”. In: *Statistics in Medicine* 21.22 (2002). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim pp. 3493–3510. ISSN: 1097-0258. DOI: 10/c4nxqv. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1251 (visited on 04/05/2023).

[10] National Cancer Institute. *NCI Dictionary of Cancer Terms - NCI*. Archive Location: nciglobal,ncienterprise. Feb. 2, 2011. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/ (visited on 05/06/2023).

[11] Mohammad Sadegh Fazeli and Mohammad Reza Keramati. “Rectal cancer: a review”. In: *Medical Journal of the Islamic Republic of Iran* 29 (Jan. 31, 2015), p. 171. ISSN: 1016-1430. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4431429/ (visited on 05/06/2023).

[12] M. McCourt, J. Armitage, and J. R. T. Monson. “Rectal cancer”. In: *The Surgeon: Journal of the Royal Colleges of Surgeons of Edinburgh and Ireland* 7.3 (June 2009), pp. 162–169. ISSN: 1479-666X. DOI: 10/cwnhg4.

[13] Mahul B. Amin et al. “The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging”. In: *CA: a cancer journal for clinicians* 67.2 (Mar. 2017), pp. 93–99. ISSN: 1542-4863. DOI: 10/f9tkns.

[14] Laura Q.M. Chow. “Head and Neck Cancer”. In: *New England Journal of Medicine* 382.1 (Jan. 2, 2020). Publisher: Massachusetts Medical Society, pp. 60–72. ISSN: 0028-4793. DOI: 10/ggf6bp. URL: https://www.nejm.org/doi/10.1056/NEJMra1715715 (visited on 05/07/2023).

[15] Graunt John. “Natural and political observations upon the bills of mortality”. In: *The Economic Writings of Sir William Petty* 2 (1662).

[16] William Matthew Makeham. “On the Law of Mortality and the Construction of Annuity Tables”. In: *Journal of the Institute of Actuaries* 8.6 (Jan. 1860). Publisher: Cambridge University Press, pp. 301–310. ISSN: 2046-1658. DOI: 10/gr2csm. URL: https://www.cambridge.org/core/journals/journal-of-the-institute-of-actuaries/article/abs/on-the-law-of-mortality-and-construction-of-annuity-tables/3EBB2F12AF8829F453E38C0D77E0E3F8 (visited on 03/27/2023).

[17] D. V. Glass. “Graunt’s life table”. In: *Journal of the Institute of Actuaries* 76.1 (June 1950). Publisher: Cambridge University Press, pp. 60–64. ISSN: 2058-1009, 0020-2681. DOI: 10/gr2csk. URL: https://www.cambridge.org/core/journals/journal-of-the-institute-of-actuaries/article/abs/graunts-life-table/996C728A54CBB11E47954DFB07F56812 (visited on 03/27/2023).

[18] E. L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (June 1, 1958). Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf pp. 457–481. ISSN: 0162-1459. DOI: 10/gdz3gq. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452 (visited on 11/23/2022).

[19] *Regression Models and Life-Tables - Cox - 1972 - Journal of the Royal Statistical Society: Series B (Methodological) - Wiley Online Library*. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x (visited on 03/27/2023).

[20] Sebastian Pölsterl. "scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn". In: *Journal of Machine Learning Research* 21.212 (2020), pp. 1–6.

[21] Cameron Davidson-Pilon. "lifelines: survival analysis in Python". In: *Journal of Open Source Software* 4.40 (2019), p. 1317. DOI: 10/gnccgr. URL: https://doi.org/10.21105/joss.01317 (visited on 03/28/2023).

[22] Nathan Mantel. "Evaluation of survival data and two new rank order statistics arising in its consideration". In: *Cancer Chemother Rep* 50.3 (1966), pp. 163–170.

[23] D. R. Cox. "The Analysis of Exponentially Distributed Life-Times with Two Types of Failure". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 21.2 (1959), pp. 411–421. DOI: 10/gr4mtw. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1959.tb00349.x (visited on 04/06/2023).

[24] MJ Stensrud and MA Hernán. "Why Test for Proportional Hazards?" In: *JAMA Guide to Statistics and Methods* Vol 323 (No. 14 2020), pp. 1401–1402. DOI: 10/ghpbdf. URL: https://jamanetwork.com/journals/jama/article-abstract/2763185 (visited on 04/06/2023).

[25] N. Breslow. "Covariance Analysis of Censored Survival Data". In: *Biometrics* 30.1 (1974). Publisher: [Wiley, International Biometric Society], pp. 89–99. ISSN: 0006-341X. DOI: 10/bxqwbf. URL: https://www.jstor.org/stable/2529620 (visited on 04/24/2023).

[26] Bradley Efron. "The Efficiency of Cox's Likelihood Function for Censored Data". In: *Journal of the American Statistical Association* 72.359 (Sept. 1, 1977). Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080 pp. 557–565. ISSN: 0162-1459. DOI: 10/gdvqq4. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1977.10480613 (visited on 04/24/2023).

[27] Odd O. Aalen. "A linear regression model for the analysis of life times". In: *Statistics in Medicine* 8.8 (1989). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002 pp. 907–925. ISSN: 1097-0258. DOI: 10/c2tgmd. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780080803 (visited on 03/27/2023).

[28] IAN W. MCKEAGUE and PETER D. SASIENI. "A partly parametric additive risk model". In: *Biometrika* 81.3 (Sept. 1, 1994), pp. 501–514. ISSN: 0006-3444. DOI: 10/bcbd2c. URL: https://doi.org/10.1093/biomet/81.3.501 (visited on 03/28/2023).

[29] L. J. Wei. "The accelerated failure time model: A useful alternative to the cox regression model in survival analysis". In: *Statistics in Medicine* 11.14 (1992). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780111409, pp. 1871–1879. ISSN: 1097-0258. DOI: 10/chfpfc. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780111409 (visited on 05/09/2023).

[30] Rinku Saikia and Manash Pratim Barman. "A Review on Accelerated Failure Time Models". In: *International Journal of Statistics and Systems* 12.2 (2017), pp. 311–322. ISSN: ISSN 0973-2675.

[31] Majeed Abdul-Fatawu. "Accelerated Failure Time Models:: An Application in Insurance Attrition". In: *The Journal of Risk Management and Insurance* 24.2 (Dec. 7, 2020). Number: 2, pp. 12–35. ISSN: 2773-9260. URL: https://jrmi.au.edu/index.php/jrmi/article/view/225 (visited on 04/30/2023).

[32] H. Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Proceedings of the 2nd International Symposium on Information Theory* (1973), pp. 267–281.

[33] M. B. Wilk and R. Gnanadesikan. "Probability Plotting Methods for the Analysis of Data". In: *Biometrika* 55.1 (1968). Publisher: [Oxford University Press, Biometrika Trust], pp. 1–17. ISSN: 0006-3444. DOI: 10/djbhpq. URL: https://www.jstor.org/stable/2334448 (visited on 05/09/2023).

[34] Ken Aho, DeWayne Derryberry, and Teri Peterson. "Model selection for ecologists: the worldviews of AIC and BIC". In: *Ecology* 95.3 (2014). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/13-1452.1, pp. 631–636. ISSN: 1939-9170. DOI: 10/gfww5d. URL: https://onlinelibrary.wiley.com/doi/abs/10.1890/13-1452.1 (visited on 05/12/2023).

[35] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Third Edition. Packt publishing ltd, 2019. ISBN: 978-1-78995-575-0.

[36] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[37] Frank E. Harrell Jr et al. "Evaluating the Yield of Medical Tests". In: *JAMA* 247.18 (May 14, 1982), pp. 2543–2546. ISSN: 0098-7484. DOI: 10/d7wchh. URL: https://doi.org/10.1001/jama.1982.03320430047030 (visited on 03/28/2023).

[38] Matthias Schmid, Marvin N. Wright, and Anreas Ziegler. *On the use of Harrell•s C for clinical risk prediction via random survival forests | Elsevier Enhanced Reader*. 2016. DOI: 10.1016/j.eswa.2016.07.018. URL: https://reader.elsevier.com/reader/sd/pii/S0957417416303633?token=9CB809AD9152D60F7E3EB63B4916859B0DB0E2C6682EF7178DAF387BF046633481CE77A0052BFF245originRegion=eu-west-1&originCreation=20230328132744 (visited on 03/28/2023).

[39] Enrico Longato, Martina Vettoretti, and Barbara Di Camillo. "A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models". In: *Journal of Biomedical Informatics* 108 (Aug. 1, 2020), p. 103496. ISSN: 1532-0464. DOI: 10/gr4wk9. URL: https://www.sciencedirect.com/science/article/pii/S1532046420301246 (visited on 04/13/2023).

[40] Hajime Uno et al. "On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data". In: *Statistics in medicine* 30.10 (May 10, 2011), pp. 1105–1117. ISSN: 0277-6715. DOI: 10/cmnt5x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3079915/ (visited on 04/13/2023).

[41] Glenn W Brier et al. "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1 (1950), pp. 1–3. DOI: 10/fp62r6.

[42] Ewout W. Steyerberg et al. "Assessing the performance of prediction models: a framework for some traditional and novel measures". In: *Epidemiology (Cambridge, Mass.)* 21.1 (Jan. 2010), pp. 128–138. ISSN: 1044-3983. DOI: 10/bj7bng. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3575184/ (visited on 05/08/2023).

[43] E. Graf et al. "Assessment and comparison of prognostic classification schemes for survival data". In: *Statistics in Medicine* 18.17 (Sept. 15, 1999), pp. 2529–2545. ISSN: 0277-6715. DOI: 10/bg5khm.

[44] DONALD B. RUBIN. "Inference and missing data". In: *Biometrika* 63.3 (Dec. 1, 1976), pp. 581–592. ISSN: 0006-3444. DOI: 10/fhqxxb. URL: https://doi.org/10.1093/biomet/63.3.581 (visited on 04/27/2023).

[45] Martijn W. Heymans and Jos W. R. Twisk. "Handling missing data in clinical research". In: *Journal of Clinical Epidemiology* 151 (Nov. 1, 2022), pp. 185–188. ISSN: 0895-4356. DOI: 10/gr6pjm. URL: https://www.sciencedirect.com/science/article/pii/S0895435622002189 (visited on 04/27/2023).

[46] Evelyn Fix and J. L. Jr. Hodges. "Nonparametric discrimination: Consistency Properties". In: *USAF School of Aviation Medicine* (1951).

[47] Markus. M Breunig, Hans-Peter Kriegel, and Jörg Sander. "LOF: Identifying Density-Based Local Outliers". In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (2000). URL: https://dl.acm.org/doi/abs/10.1145/342009.335388 (visited on 05/12/2023).

[48] G. E. P. Box and D. R. Cox. "An Analysis of Transformations". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 26.2 (1964). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1964.tb00553.x, pp. 211–243. ISSN: 2517-6161. DOI: 10/gfrhvs. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1964.tb00553.x (visited on 04/05/2023).

[49] In-Kwon Yeo and Richard A. Johnson. "A new family of power transformations to improve normality or symmetry". In: *Biometrika* 87.4 (Dec. 1, 2000), pp. 954–959. ISSN: 0006-3444. DOI: 10/bcc9jv. URL: https://doi.org/10.1093/biomet/87.4.954 (visited on 04/05/2023).

[50] Kine M. Bakke et al. "Sex Differences and Tumor Blood Flow from Dynamic Susceptibility Contrast MRI Are Associated with Treatment Response after Chemoradiation and Long-term Survival in Rectal Cancer". In: *Radiology* 297.2 (Nov. 1, 2020). Publisher: Radiological Society of North America, pp. 352–360. ISSN: 0033-8419. DOI: 10/gr6vb8. URL: https://doi.org/10.1148/radiol.2020200287 (visited on 04/28/2023).

[51] Steven I. Gutman et al. "Background". In: *Progression-Free Survival: What Does It Mean for Psychological Well-Being or Quality of Life? [Internet]*. Agency for Healthcare Research and Quality (US), Apr. 2013. URL: https://www.ncbi.nlm.nih.gov/books/NBK137763/ (visited on 04/29/2023).

[52] Jon Magne Moan et al. "The prognostic role of 18F-fluorodeoxyglucose PET in head and neck cancer depends on HPV status". In: *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 140 (Nov. 2019), pp. 54–61. ISSN: 1879-0887. DOI: 10/gr5wnv.

[53] Yngve Mardal Moe et al. "Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients". In: *European Journal of Nuclear Medicine and Molecular Imaging* 48.9 (Aug. 2021), pp. 2782–2792. ISSN: 1619-7070, 1619-7089. DOI: 10/gn8598. URL: https://link.springer.com/10.1007/s00259-020-05125-x (visited on 04/20/2023).

[54] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *kdd*. Vol. 96. Number: 34. 1996, pp. 226–231.

[55] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest". In: *2008 Eighth IEEE International Conference on Data Mining*. 2008 Eighth IEEE International Conference on Data Mining. ISSN: 2374-8486. Dec. 2008, pp. 413–422. DOI: 10/cndm4g.

[56] Terry M Therneau. *A package for survival analysis in R*. manual. 2023. URL: https://CRAN.R-project.org/package=survival.

[57] Stephane Fotso et al. *PySurvival: Open source package for survival analysis modeling*. 2019. URL: https://www.pysurvival.io/.

[58] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10/ggr6q3.

[59] Michel Lang et al. "mlr3: A modern object-oriented machine learning framework in R". In: *Journal of Open Source Software* (Dec. 2019). DOI: 10/gnmmjn. URL: https://joss.theoj.org/papers/10.21105/joss.01903.

[60] Oliver Tomic, Thomas Graff, and Kristian Hovde Liland. "hoggorm: a python library for explorative multivariate statistics". In: *The Journal of Open Source Software* 4.39 (2019). DOI: 10/gr5g8f. URL: http://joss.theoj.org/papers/10.21105/joss.00980.

[61] Junyong In and Dong Kyu Lee. "Survival analysis: part II – applied clinical data analysis". In: *Korean Journal of Anesthesiology* 72.5 (Oct. 2019), pp. 441–457. ISSN: 2005-6419. DOI: 10/gg8pbp. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6781220/ (visited on 05/11/2023).

APPENDICES

APPENDIX

A

A - GITHUB REPOSITORY

The code utilised during the course of this thesis is located in the following GitHub library:

Github repository link

- `https://github.com/mikkorekstad/M30-DV`

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway